Adapting Machine Translation Education to the Neural Era: A Case Study of MT Quality Assessment

Lieve Macken, Bram Vanroy and Arda Tezcan
LT[3], Language and Translation Technology Team
Ghent University
Belgium
{firstname.lastname}@ugent.be

Abstract

The use of automatic evaluation metrics to assess Machine Translation (MT) quality is well established in the translation industry. Whereas it is relatively easy to cover the word- and character-based metrics in an MT course, it is less obvious to integrate the newer neural metrics. In this paper we discuss how we introduced the topic of MT quality assessment in a course for translation students. We selected three English source texts, each having a different difficulty level and style, and let the students translate the texts into their L1 and reflect upon translation difficulty. Afterwards, the students were asked to assess MT quality for the same texts using different methods and to critically reflect upon obtained results. The students had access to the MATEO web interface, which contains word- and character-based metrics as well as neural metrics. The students used two different reference translations: their own translations and professional translations of the three texts. We not only synthesise the comments of the students, but also present the results of some cross-lingual analyses on nine different language pairs.

1 Introduction

Machine translation (MT) is increasingly being used in professional translation workflows. "MT literacy and awareness of MT's possibilities and limitations" forms therefore, according to the EMT competence framework (EMT, 2022), an integral part of professional translation competences. At Ghent University, we offer a 5-credit course *Machine Translation and Post-editing*, which is part of the one-year postgraduate programme *Computer-Assisted Language Mediation*[1] and the two-year *European Master in Technology for Translation and Interpreting*[2]. The MT part of the course aims to provide a comprehensive overview and covers topics such as the main linguistic challenges for MT, different approaches to MT (rule-based, statistical and neural MT) and MT evaluation. The students also acquire hands-on experience in building and evaluating their own MT systems using MutNMT[3].

The use of automatic evaluation metrics to assess Machine Translation (MT) quality is well established in the translation industry. The more traditional word- and character-based metrics are relatively easy to run and it is therefore easy to incorporate them in a university course. But, much more technical knowledge is required to get the neural metrics, which are based on large pre-trained language models, up and running. Despite their better performance, they are therefore less popular in translation courses. In this paper, we discuss how we introduced the topic of MT quality assessment in a course for translation students of varying language backgrounds. The students got one dedicated lecture on the subject of MT evaluation, which covers both manual and automatic evaluation methods and had to critically reflect both on the suitability of MT for three different texts and on the usefulness of automatic evalua-

[1] https://www.ugent.be/lw/vtc/nl/opleidingen/postgraduaten/calm/calmbrochure
[2] https://em-tti.eu/
[3] https://multitrainmt.eu/index.php/en/neural-mt-training/mutnmt

Nurminen, Brenner, Koponen, Latomaa, Mikhailov, Schierl, Ranasinghe, Vanmassenhove, Vidal, Aranberri, Nunziatini, Escartín, Forcada, Popovic, Scarton, Moniz (eds.)
Proceedings of the 24th Annual Conference of the European Association for Machine Translation, p. 305–314
Tampere, Finland, June 2023.

tion. They made use of an early version of MATEO (MAchine Translation Evaluation Online) (Vanroy et al., 2023)[4], an easy-to-use web interface for evaluating MT output by means of a variety of both word- and character-based and neural MT evaluation metrics.

2 Related Research

The term *Machine Translation literacy* has been introduced by Bowker and Buitrago Ciro in the context of scholary communication (2019) and has since then been picked up by other scholars. O'Brien and Ehrensberger-Dow used the term in the context of professional translation and described MT literacy as "knowing how MT works, how it can be useful in a particular context, and what the implications are of using MT for specific communicative needs" (2020, p. 146). Its growing importance is of course related to the ever-increasing quality of MT systems.

Several initiatives have been taken to develop and distribute publicly available learning materials tailored to teaching MT to translation students. In the framework of the MultitraiNMT Erasmus+ project (Forcada et al., 2022) an open access course book has been published (Kenny, 2022) targeting both language learners and translators. The project also developed an open source pedagogically-oriented neural MT platform called MutNMT, in which students can go through the various stages of building an MT engine, from uploading parallel corpora, over training and evaluating an MT system and inspecting translations.

Another initiative is DataLit^{MT} (Teaching Data Literacy in the Context of Machine Translation Literacy), in which, among others, a Python notebook has been created to explain translation-oriented Natural Language Processing (NLP) concepts to translation students (Krüger, 2022).

With MATEO, Ghent university adds another didactic tool to the existing set. MATEO differs from the aforementioned tools in the sense that it gives non-technical users access to both more traditional (word- and character-based) metrics and state-of-the-art neural automatic evaluation metrics via a web interface.

3 Data collection

The students' project consisted of two parts. In the first part students were asked to manually translate three English texts into their L1. They were told that their translations would be used to assess MT quality. In addition, they had to reflect upon translation difficulty of the three texts. The second part dealt with MT evaluation. Students were asked to evaluate the MT output of three different MT engines for the three texts using manual and automatic evaluation methods. For the manual evaluation, students ranked the three MT suggestions (from best to worst; equal rankings allowed) and provided accuracy[5] and fluency scores on a 5-point scale for each MT sentence. Accuracy scores relate to the amount of content and meaning of the source sentence that is retained in the MT output. Fluency scores relate to the degree to which a sentence meets the standards and conventions of the target language.

For the automatic evaluation the lecturers provided the students also with professional translations for the three texts. The students made use of an early version of the MATEO web interface to obtain automatic scores for 6 different metrics (BLEU, TER, ChrF, BERTScore, BLEURT and COMET, see section 3.1 for more details) using two different reference translations: their own translation and the professional translation. MATEO contains easily accessible descriptions of the different metrics and students could look up more details by clicking on the links to the original research papers. The students were asked to compare perceived translation difficulty with obtained MT translation quality and critically reflected on different aspects of the MT evaluation task.

The English source texts were taken from the LeConTra data set (Vanroy and Macken, 2022), which contains Dutch student translations of English source texts enriched with translation process data in the form of keystroke logging. We selected three texts (see Table 1) of different difficulty level based on two parameters derived from the process data: average translation duration (per token) and average number of revisions per segment (1 indicating no revision). The length of the selected source texts varies from 188 to 231 words (or 214 to 260 tokens). The three texts also differed in terms of lexical richness, calculated as mean segmental type-token ratio (with a window of 100). The first text (T1) deals with lovesickness and is the most informal text containing figurative

[4] https://lt3.ugent.be/mateo/

[5] Adequacy is often used as a synonym for accuracy in the context of MT evaluation

and connotative content. The second text (T2) discusses the consequences of globalisation and can be considered more objective than the first text. The third text describes the discovery of the hidden laboratory used by Leonardo da Vinci and is predominantly denotative (T3). According to our selection criteria, T3 was the easiest text and T2 the most difficult text to translate from English into Dutch.

id	T1	T2	T3
lecontra_id	T23	T07	T20
avg. revisions	1.47	2.26	1.33
avg. tok. transl. dur. (s)	3.9	5.4	3.3
sents	10	9	9
tokens	214	216	260
avg. sent. len.	21.4	24.0	28.9
lex. richness (MSTTR)	0.73	0.78	0.69

Table 1: Source text statistics of the three texts

The Dutch professional translations were retrieved from the Dutch Parallel Corpus (Macken et al., 2011). The professional translations for the other languages were obtained via a translation agency; the cost was 15 cents per word supplemented by 21% VAT.

Students were asked to compare the MT quality of three different MT systems. They were given the output of Facebook's multilingual translation model M2M100 1.2B (Fan et al., 2021) and had to create the MT output with Google Translate and a third MT system of their choice. In total, the students worked with 9 different target languages (Dutch, French, Brazilian Portuguese, Romanian, Turkish, Farsi, Kazakh, Ukrainian and Russian). All students had either received formal translation training in a prior training programme or had gained work experience as a translator. For the remainder of this paper we only retained the submission of the most experienced student per target language. Informed consent was obtained for all student submissions included in this study.

Google Translate was chosen as a state-of-the-art system that covers many languages, including low resource ones. M2M100 also supports many of these languages but unlike Google's service, it is an open-source model. Our study therefore also sheds some light on the performance of open models compared to closed ones. It should be noted that we did not use the largest available M2M model (12B parameters) but instead opted for a computationally more feasible variant (1.2B parameters). Students were free to choose a third system themselves based on their own preference and the target language that they worked on. Most students used DeepL as third MT system. Bing was to used for Farsi and LingvaNex was used for Kazakh.

In section 4 we not only synthesize the comments of the students, but also present the results of some cross-lingual analyses as the obtained data set allows us to examine the automatic evaluation metrics across typologically different languages.

3.1 Metrics

At the time of writing, the MATEO interface supports the following automatic evaluation metrics.

BLEU (Papineni et al., 2002, BiLingual Evaluation Understudy) calculates a precision score of n-gram matches (consecutive words or tokens) between a machine translation and one or more reference translations. BLEU has become a widely used evaluation metric due to its simplicity and computational efficiency, despite calls to "retir[e] BLEU as the de facto standard metric" (Mathur et al., 2020, p. 4992) because of its low correlation with human judgements.

ChrF (Popović, 2015; Popović, 2017, Character F-score) is based on the comparison of character n-grams (rather than token n-grams like BLEU) between a machine translation and one or more reference translations, and it calculates an F-score based on the precision and recall of the n-gram sequences. Because of its emphasis on characters, ChrF is language-independent and tokenization-independent which makes it straightforward to use. It also correlates better with human judgements compared to BLEU (Freitag et al., 2022).

TER (Snover et al., 2006, Translation Edit Rate) is based on edit distance and measures the number of edits (token insertions, deletions, substitutions and shifts) required to transform a machine translation into a reference translation. The total TER score is calculated by the number of aforementioned edits, divided by the number of words in the reference translation. (For readability's sake, we also multiplied them by 100.) While TER is an intuitive metric to show the differences between an MT candidate and reference translations, it has

received the same criticisms as BLEU (Mathur et al., 2020) due to its low correlation with human judgements especially when TER scores are used to compare two MT systems.

BERTScore builds on the success of pre-trained, multilingual language models (Zhang et al., 2020). Rather than relying on string-based token or character matching statistics, it embeds given candidate and reference tokens in a multidimensional vector space and then calculates the similarities between the two, and aggregating scores into Precision, Recall and an F-score (in this paper we use the BERTScore F-score). As such BERTScore is not restricted to the surface form and is capable of covering paraphrasing. BERTScore uses existing pre-trained models under the hood to retrieve the token embeddings without retraining the model. We use the default models associated with each language, that is multilingual BERT (Devlin et al., 2019) in all our cases, except for a Turkish BERT for Turkish.[6]

BLEURT (Sellam et al., 2020, Bilingual Evaluation Understudy with Representations from Transformers) is another neural metric based on pre-trained models. Unlike BERTScore, however, it is a learnt metric. The metric uses existing BERT (Devlin et al., 2019) or RemBERT (Chung et al., 2023) models as a starting point and trains them in a three-stage fashion. First, regular BERT pre-training. Secondly, the model is pre-trained on synthetic data related to translation evaluation to learn signals from, among others, BLEU, ROUGE, BERTScore as well as factors such as back-translation likelihood. Finally, the model is fine-tuned on task-specific MT quality ratings from the WMT Metrics Shared Tasks (Freitag et al., 2022). Overall, BLEURT was shown to correlate much better with human ratings than metrics such as BLEU and TER and also outperforming the non-learnt neural metric BERTScore. We use the recommended BLEURT-20 checkpoint.

COMET (Rei et al., 2020, Crosslingual Optimized Metric for Evaluation of Translation) is a learnt metric like BLEURT above. It relies on a pre-trained multilingual model XLM-R (Conneau et al., 2020) which is then fine-tuned on human judgement scores, including data from the WMT Metrics shared tasks (Freitag et al., 2022), the QT21 corpus (Specia, 2017), and a proprietary MQM annotated corpus. Unlike BLEURT, COMET also uses the source sentence as input to calculate a final evaluation score. The authors show that COMET outperforms metrics such as BLEU and ChrF as well as BERTScore. We use the recommended Estimator wmt22-comet-da checkpoint.

[6] https://huggingface.co/dbmdz/bert-base-turkish-cased

4 Results

4.1 Students' findings

In what follows we first synthesise the comments of the students on the manual and automatic evaluation task and perceived translation difficulty.

4.1.1 Overall MT Quality

As mentioned in section 3, all students used Google Translate and Facebook's multilingual translation model M2M100 1.2B. For most languages the third system was DeepL, except for Kazakh and Farsi for which respectively LingvaNex and Bing was used. According to the students' scores and comments Google Translate and DeepL delivered better translations than M2M. The differences in quality between DeepL and Google Translate were small and varied across language pairs and across the three texts. For Kazakh LingvaNex was considered to be worse than Google Translate but better than M2M. For Farsi Google Translate was the best system, and Bing and M2M were on par.

4.1.2 Perceived translation difficulty

With regards to perceived difficulty of the human translation task, agreement among students was moderate. Most students found the first text the easiest to translate and the third text the hardest, which is not what we expected based on the English-Dutch process data. Individual factors such as interest in the topic, background knowledge and translation experience in specific genres (e.g. literary translation) were frequently mentioned as factors determining translation difficulty apart from text-specific characteristics. Text-specific difficulties that were commented on by the students were situated at the lexical level (domain-specific terminology, idiomatic expressions, proper names) and at the structural level (noun stacking, word order differences and complex sentences). Some students also referred to stylistic elements such as Text 1 and Text 2 having a rich vocabulary (evidenced by the use of

non-frequent words and synonyms) and various instances of figurative language. Students also struggled to disentangle long complex sentences. For example in the third text a student mentioned that he had to read the sentence *The Tuscan-born scientist, painter, philosopher and poet was aged 51 when he returned to Florence in 1503 after many years in Milan, where he already had established his reputation, and a period of extended travel*, a couple of times, to understand what the phrase *and a period of extended travel* refers to. The linguistic distance between source and target language was also mentioned several times. At the lexical level several students mentioned that there is no straightforward translation equivalent for the word *lovesickness* in their target language. At the structural level, most difficulties relate to differences in word order (e.g. Turkish and Farsi).

Obtained MT quality did not always align with perceived difficulty. For most target languages, MT achieved the best scores for the third text despite the abundance of proper names and long complex sentences. Only for Farsi proper names were not always rendered correctly (Bing left some proper names untranslated and M2M did not write proper names in the Persian alphabet). Students suggested various reasons for the third text having the best MT quality. One student referred to the more denotative and objective nature of the text, which makes it more suitable for machine translation. The texts contains mainly factual information. Moreover, students suggested that the data the NMT systems were trained on most probably covered all proper names and titles. The first text was considered the most informal, with more figurative and connotative content and thus allows for more creativity in human translation, but proved therefore to be more challenging for machines.

4.1.3 Manual versus automatic evaluation

Students reflected upon the (dis)agreement of their manual assessments with the obtained automatic scores. According to one student *"the best automatic scores are the ones that correlate most with the human assessment, and do not have massive scoring disparities when using the professional translation as a reference and when using the student translation as a reference"*. Taking text-averages into account to compare top-middle-bottom rankings, most students found that both the more traditional and the neural automatic evaluation metrics fit this criterion at text level, but not at sentence level. Several students pointed out that ChrF worked better for their target language than the word-based metrics BLEU or TER as it can capture differences on character level which makes it more suitable for highly inflected languages such as Russian, Kazakh and Turkish.

Neural metrics (BERTScore, BLEURT and COMET) were perceived to be more comparable to the human evaluation for most language pairs. A notable exception is Kazakh for which the student suspects that there was not enough data to train the neural metrics properly. Students attributed the better agreement to the ability of the neural metrics to capture semantic similarities between the MT output and the reference translation (e.g., synonyms or paraphrases), making the neural metrics less reliant on exact matches in word choice. However, critical remarks were made that text-level assessment may not provide a comprehensive assessment as more extreme values get levelled out. One specific problem mentioned at sentence level was that COMET sometimes produced 0s even though the MT output was not completely incorrect and still preserved some meaning of the source sentence. Also the opposite was true and some sentences got a 100 COMET score even when the MT output was flawed. Most students did not express a preference for a particular neural metric. The only exception was the Turkish student who preferred BERTScore.

4.1.4 Impact of the reference translation

For most language pairs, the professional translations deviated more from the source text than the student translations. The only exception was the translation delivered by the Ukrainian professional translator, which in hindsight, was a post-edited version of DeepL as the average TER score was exceptionally low and 5 out of 28 sentences even received a TER-score of 0, which means that the professional translated sentence was identical to the DeepL version. The professional translations exhibited more occurrences of paraphrasing, reordering, and structural changes, whereas the student translations followed the structure of the source sentences more closely. This finding seems to be in line with translation process research where expertise is taken into account. Inexperienced translators have been shown to treat translation as a more lexical task, whereas professional translators pay more attention to higher order concerns such as coherence and style (Séguinot, 1991).

These characteristics of the reference translations have an impact on the obtained scores, and students suggested that this impact is higher for the word/character-based scores than for the neural ones. One can expect that the more 'literal' the human translation is, the higher the automatic scores are. Overall, the student translations resembled more the machine translations, which also stay quite close to the source text. One student noted that the professional translations sound nicer in terms of style, but that this makes it it harder to accurately judge the quality of the MT systems.

4.2 Cross-lingual analyses

In this study we have collected translations and manual student assessments of MT quality for nine different languages. These data sets enable us to compare between metrics and languages, taking into account the origin of the reference translation (professional vs. student).

4.2.1 Correlation between human ranking and automatic metrics

Human ranking is an evaluation technique to compare different MT systems against each other. Students were asked to rank, for each sentence, the MT systems from best to worst. Similarly, we can use automatic metric scores for each system to rank the MT systems from best to worst, per metric. In this section we investigate how well the ranks from a given metric correlate with the human ranking with Spearman correlations. We make the distinction between the cases where the professional translation was taken as a reference when calculating the automatic metrics (PROF) and when the student translation served as a reference (STUD). Note that the negative correlation for TER is to be expected because a higher TER score is "worse" (indicating more edits needed) but for other metrics a higher score is "better".

In Table 2 we see that the ranks of MT systems as assigned by individual metric scores correlate moderately with the ranks of those MT systems assigned by human evaluators. Generally speaking, neural metrics correlate better with human ranks than word-based metrics. ChrF, a character-based metric, correlates relatively well with manual ranks, on-par or exceeding the correlations of BERTScore and COMET. BLEURT rankings correlate best with human rankings, both in the students and professional setting.

ref_type	metric	spearman ρ
PROF	BLEU	0.37
	ChrF	0.44
	TER	-0.26
	BERTScore	0.41
	BLEURT	**0.52**
	COMET	0.47
STUD	BLEU	0.39
	ChrF	0.43
	TER	-0.29
	BERTScore	0.43
	BLEURT	**0.50**
	COMET	0.42

Table 2: Correlations between the ranks assigned to MT engines by automatic metrics and the manual ranks assigned by students. $p < .001$ for all correlations. Best correlations are highlighted in bold.

We find that the absolute correlation for word-based metrics (BLEU, TER) are higher when using student translations as references instead of professional translations, whereas the other metrics correlate less in the student setting. An explanation may be found in what was mentioned in the previous section: student translations followed the structure of the source sentences more closely, whereas professional translations deviated more from the source text. When this behaviour is combined with MT systems' tendency to opt for more common words and to stay close to the source text, we can expect student translations to be more similar to the MT output and score better on lexical matching metrics.

4.2.2 Correlation between accuracy and fluency scores and automatic metrics

In addition to ranking the different MT engines for each translated sentence, students were also asked to rate the accuracy and fluency on a scale of 1 to 5 (5 being the best score). This allows us to correlate automatic metric scores for each sentence with manually annotated accuracy and fluency scores for those sentences using the data of all MT systems.

Table 3 indicates four things. First, neural metrics, in general, correlate better with accuracy and fluency than word-based metrics. Note, however, that ChrF correlates well, especially when using the student translation as reference.

Second, accuracy is overall better correlated with automatic metrics than fluency. However, this is not or barely the case for the word-based metrics BLEU and TER. This seems to imply that the other metrics cover accuracy more than fluency, relatively speaking.

ref_type	metric	spearman ρ (accuracy)	spearman ρ (fluency)
PROF	BLEU	0.34	0.36
	ChrF	0.39	0.35
	TER	-0.30	-0.32
	BERTScore	0.43	**0.40**
	BLEURT	**0.45**	0.37
	COMET	0.41	0.36
STUD	BLEU	0.41	0.40
	ChrF	0.46	0.38
	TER	-0.37	-0.37
	BERTScore	0.48	0.45
	BLEURT	**0.53**	**0.46**
	COMET	0.46	0.40

Table 3: Correlations between the automatic metric scores and the manual accuracy and fluency ratings. $p < .001$ for all correlations. Best correlations are highlighted in bold.

STUD	ChrF			BLEURT		
	M2M	**GT**	**MT3**	**M2M**	**GT**	**MT3**
FA	43.95	**53.97***	45.24	60.77	**65.87***	55.92
KZ	19.05	**62.19***	57.64	22.18	**83.58***	75.58
FR	65.00	**73.50**	69.68	67.49	**77.52**	76.73
NL	66.61	**72.33***	69.47	75.09	79.72	**79.82**
PT	75.10	**88.66**	76.70	76.25	**85.72**	77.29
RO	63.09	65.54	**74.33***	76.67	79.94	**83.19***
RU	53.86	64.12	**67.89**	65.70	**80.59**	79.94
TR	46.94	53.40	**54.81**	67.44	71.19	**74.61**
UA	46.96	**52.18**	51.22	66.23	**74.18**	74.04

Table 5: MT system performance with respect to ChrF and BLEURT when student translations are used as reference (STUD). The highest scores are highlighted in bold. Statistically significant improvements achieved by the best-performing system in comparison to the second-best system are indicated with a star symbol (*) for $p < 0.05$.

Third, using student translations as reference translations when calculating the automatic metrics again yields higher correlations in all settings. As mentioned in the previous section, this can likely be explained by the more 'literal' translations of student translators yielding higher metric scores when using student references.

Finally, the correlations are stronger than in the previous ranking correlation, especially in the student reference scenario and more so in terms of accuracy. The higher correlation compared to the previous section may be explained by the effect of reducing MT metric scores to ranks. It is possible that reducing the MT scores to a 3-point ranking scale in the previous section and correlating it with another 3-point ranking "smooths away" some tendencies. For instance, in the scenario that M2M has a score of 67 for a given metric, DeepL 93, and Google Translate 97, then the ranks were reduced to 3, 2, 1 respectively. But from those ranks it is not clear that DeepL is relatively much closer to Google Translate. In this section we use the full range of the metrics and correlate them with a five-point scale without any rescaling or ranking. That means that correlations can be drawn more easily, because in the example above the low score 67 in M2M can be reflected in lower accuracy/fluency scores (e.g. 2) compared to higher ones for DeepL and Google Translate (e.g. 4 and 5).

4.2.3 MT system performance

With access to many different languages and three different MT systems, we can make a number of observations about the average quality that is achieved for each language and MT system. In Tables 4 and 5, we analyse the translation performance for M2M, Google Translate (GT) and the third MT engine (MT3), using two of the metrics that correlated well with human judgements: a character-based metric ChrF, and a neural, learnt metric BLEURT. For most languages, the third MT engine (MT3) is DeepL, except for Farsi (FA) and Kazakh (KZ), where Bing and Lingvanex were used respectively. The metric scores in Table 4 use the professional translations as reference, whereas the scores in Table 5 are based on the student translations as reference. For both scenarios, we used paired t-test to measure the statistical significance of the differences between the means of the metric scores obtained for the best and the second best-performing systems, per metric, per language.

PROF	ChrF			BLEURT		
	M2M	**GT**	**MT3**	**M2M**	**GT**	**MT3**
FA	45.35	**61.40***	50.00	62.20	**72.41***	60.59
KZ	19.22	**49.86***	46.72	22.55	**83.23***	75.67
FR	54.54	61.99	**66.75***	59.79	66.97	**71.74***
NL	53.06	54.88	**56.23**	67.51	71.71	**72.27**
PT	54.61	**57.48**	57.39	66.28	**70.05**	67.50
RO	59.71	60.19	**65.33***	73.86	75.74	**79.14***
RU	48.55	52.45	**53.13**	66.74	**75.84**	74.72
TR	45.36	49.08	**51.64***	66.38	70.23	**72.91***
UA	62.94	65.08	**84.87***	75.4	81.57	**87.56***

Table 4: MT system performance with respect to ChrF and BLEURT when professional translations are used as reference (PROF). The highest scores are highlighted in bold. Statistically significant improvements achieved by the best-performing system in comparison to the second-best system are indicated with a star symbol (*) for $p < 0.05$.

Looking at Tables 4 and 5, we observe that ChrF and BLEURT tend to agree on the best system, with the exceptions of Dutch (NL-STUD) and Russian (RU-PROF and RU-STUD). However, for these languages, the differences in evaluation scores between the best system and the second-best one are not statistically significant.

Looking at the performance of the MT engines for both PROF and STUD, notably, we observe that M2M performs worst in general, with the exception of the BLEURT scores for Farsi, where M2M performs slightly better than Bing. Furthermore although Kazakh is an officially supported language for this engine, the M2M output resulted in very low scores with respect to both metrics. It is possible that the low-resource nature of this language pair is one of the main causes of the low performance. For FA and KZ, we observe that Google Translate not only outperforms M2M but also Bing (FA) and Lingvanex (KZ).

When professional translations are used as reference (PROF), for the remaining languages, DeepL (MT3) seems to be the better MT engine in general, as it outperforms Google Translate for French (FR), Dutch (NL), Romanian (RO), Turkish (TR) and Ukrainian (UA) with respect to both metrics. Moreover, the improvements in all these languages, except NL, are statistically significant. While DeepL performs worse than Google Translate for Portuguese (PT) with respect to both metrics, and Russian (RU) with respect to BLEURT, the differences in estimated translation quality for these languages are not statistically significant.

When we look at the results obtained for STUD, in Table 5, we see similar trends for FA and KZ. For both languages, Google Translate outperforms Bing and Lingvanex with statistically significant improvements. For the remaining languages, we see more balanced results for the best-performing system. For FR, PT and UA, Google Translate achieves higher scores with respect to both metrics than DeepL. However, none of these improvements is statistically significant. Similar to the case of PROF, for RO and TR, DeepL outperforms Google Translate with respect to both metrics and with statistically significant differences for RO. For NL and RU, the two metrics do not agree on the best-performing system (Google Translate vs. DeepL) and only for NL the differences in ChrF scores are statistically significant.

Again, the metric scores seem to be higher in general when student translations are used as reference (STUD). To illustrate this difference more clearly, in Table 6 we analyse the differences between the average estimated translation quality when the student (STUD) and professional (PROF) translations are used separately. To this end, we provide the difference between the two cases by subtracting the average metric scores obtained on professional translations from the ones obtained on student translations, per language, per MT engine. Similar to the previous analyses, we use paired t-test to measure the statistical significance of the differences between the means of the metric scores (PROF vs. STUD in this case).

	ChrF			BLEURT		
	M2M	**GT**	**MT3**	**M2M**	**GT**	**MT3**
FA	-1.39	-7.43*	-4.76	-1.42	-6.53*	-4.67
KZ	-0.17	12.34*	10.92*	-0.38	0.35	-0.09
FR	10.46*	11.51*	2.93	7.70*	10.54*	4.99
NL	13.55*	17.45*	13.24*	7.58*	8.01*	7.54*
PT	20.50*	31.18*	19.31*	9.97*	15.67*	9.79*
RO	3.38	5.35	9.00	2.81	4.20*	4.05
RU	5.31	11.67*	14.76*	-1.04	4.76*	5.22*
TR	1.58	4.32	3.17	1.06	0.96	3.51
UA	-15.98*	-12.90*	-33.64*	-9.17*	-7.39*	-13.52*

Table 6: Difference between the average metric scores when the student and professional translations are used as reference (student minus (-) professional). Statistically significant differences are indicated with the star symbol (*) for $p < 0.05$.

In Table 6, positive values indicate that the score was higher when student translations were used as reference (STUD), while negative values indicate that the MT output yielded a higher score when professional translations (PROF) were used as reference. By looking at these results, we can see a general tendency that using student translations as reference leads to higher evaluation scores with respect to both metrics and for all MT engines, for the majority of the languages, with the exception of Farsi and Ukrainian. Especially for Google Translate and MT3, the results illustrate that both a neural-based (BLEURT) and a character-based (ChrF) evaluation metric estimate the performance of the MT engines to be higher when student translations are used as references, in comparison to using professional translations instead. These differences are also measured to be statistically significant in most cases.

There are potential explanations for the discrepancy between the results observed for FA and UA, for which the two metrics result in higher average scores when professional translations are used as references. For UA, one explanation is, as stated earlier, that the professional translator post-edited the DeepL (MT3) output to achieve correct translations. A plausible explanation for FA is that given the linguistic distance between English and Farsi, it is not possible to stay close to the source structure and that especially for longer sentences restructuring is needed, which the student apparently

did to a greater extent than the professional translator.

5 Conclusion

Machine translation is taking an increasingly prominent place in professional workflows and so is the assessment of MT quality. MT evaluation methods, both human as well as automatic, thus deserve sufficient attention in MT courses targeting translation students. Whereas research demonstrates that the newer neural automatic evaluation metrics correlate better with human judgements than the more traditional word- and character-based metrics, the neural metrics are not often used in translation courses as quite some technical skills are required to get them up and running

This paper focused on MT evaluation and how it can be taught to translation students. Via the MATEO web interface students had access to six different automatic metrics: two word-based, one character-based and three neural metrics. Students translated three English source texts from scratch into their L1 and assessed MT quality afterwards using manual methods and automatic evaluation metrics. They were asked to critically reflect upon obtained results. Perceived difficulty and MT quality did not always align, which seems to suggest that translation students and machine translation systems face different problems during translation.

Many of the comments that the students made were afterwards confirmed in the cross-lingual analyses. According to the students' comments Google Translate and DeepL delivered better translations than Facebook's M2M100 1.2B, with differences in quality between Google Translate and DeepL varying across language pairs. Within the word/character-based metrics, ChrF was found to be the better metric, especially for highly inflected languages. Overall, the neural metrics were perceived to be more comparable to human evaluation, a statement that was partially confirmed in the cross-lingual analyses, in which BLEURT came out as best metric, but in which ChrF also correlated well.

Automatic metrics were considered to be useful for MT quality assessment, but only for text-level evaluations. It is important to note that all metrics work on different scales, and that scores are therefore not comparable across languages, which makes it difficult to compare results.

Different analyses showed that, in general, the obtained automatic scores were higher when the student translations were used as reference translations. Students tended to stay closer to the source text, whereas professional translators deviated more from the source text. As we worked with students' data, our data set was limited and is too small to make firm conclusions, but it seems worthwhile to further explore the impact of the origin of the reference translations on translation quality assessment.

The data set (source texts, reference translations and MT output) is freely available on GitHub[7].

References

Bowker, Lynne and Ciro Jairo Buitrago. 2019. *Machine Translation and global research: towards improved machine translation literacy in the scholary community*. Esmerald Publishing, Bingley.

Chung, Hyung Won, Thibault Fevry, Henry Tsai, Melvin Johnson, and Sebastian Ruder. 2023. Rethinking Embedding Coupling in Pre-trained Language Models. January.

Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July. Association for Computational Linguistics.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

EMT. 2022. European Master's in Translation competence framework 2022.

Fan, Angela, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. *The Journal of Machine Learning Research*, 22(1):107:4839–107:4886, January.

[7] https://github.com/ardate/MT-quality-assessment_MATEO

Forcada, Mikel L., Pilar Sánchez-Gijón, Dorothy Kenny, Felipe Sánchez-Martínez, Juan Antonio Pérez Ortiz, Riccardo Superbo, Gema Ramírez Sánchez, Olga Torres-Hostench, and Caroline Rossi. 2022. MultitraiNMT erasmus+ project: Machine translation training for multilingual citizens (multitrainmt.eu). In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 291–292, Ghent, Belgium, June. European Association for Machine Translation.

Freitag, Markus, Ricardo Rei, Nitika Mathur, Chikiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 Metrics Shared Task: Stop Using BLEU – Neural Metrics Are Better and More Robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid), December. Association for Computational Linguistics.

Kenny, Dorothy. 2022. *Machine translation for everyone: Empowering users in the age of artificial intelligence. (Translation and Multilingual Natural Language Processing 18).* Language Science Press, Berlin.

Krüger, Ralph. 2022. Using Jupyter notebooks as didactic instruments in translation technology teaching. *The Interpreter and Translator Trainer*, 16(4):503–523.

Macken, Lieve, Orphée De Clercq, and Hans Paulussen. 2011. Dutch parallel corpus: a balanced copyright-cleared parallel corpus. *META*, 56(2):374–390.

Mathur, Nitika, Timothy Baldwin, and Trevor Cohn. 2020. Tangled up in BLEU: Reevaluating the Evaluation of Automatic Machine Translation Evaluation Metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997, Online, July. Association for Computational Linguistics.

O'Brien, Sharon and Maureen Ehrensberger-Dow. 2020. MT Literacy - A cognitive view. *Translation, Cognition & Behavior*, 3(2):145–164.

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, USA, July. Association for Computational Linguistics.

Popović, Maja. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal, September. Association for Computational Linguistics.

Popović, Maja. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Rei, Ricardo, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A Neural Framework for MT Evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online, November. Association for Computational Linguistics.

Sellam, Thibault, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning Robust Metrics for Text Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online, July. Association for Computational Linguistics.

Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231.

Specia, Lucia. 2017. QT21 data. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Séguinot, Candace. 1991. A study of student translation strategies. In Tirkkonen-Condit, Sonja, editor, *Empirical Research in Translation and Intercultural Studies*, pages 79–88. Gunter Narr, Tübingen.

Vanroy, Bram and Lieve Macken. 2022. LeConTra: A learner corpus of English-to-Dutch news translation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1807–1816, Marseille, France, June. European Language Resources Association.

Vanroy, Bram, Arda Tezcan, and Lieve Macken. 2023. MATEO: MAchine Translation Evaluation Online. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, Tampere, Finland, June. European Association for Machine Translation.

Zhang, Tianyi, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *Proceedings of ICLR 2020*.