Quality Analysis of Multilingual Neural Machine Translation Systems and Reference Test Translations for the English-Romanian language pair in the Medical Domain

Miguel Rios, Alina Secară, Raluca-Maria Chereji, Dragoș Ciobanu
Centre for Translation Studies, University of Vienna
{miguel.angel.rios.gaona, alina.secara,
raluca-maria.chereji, dragos.ioan.ciobanu}@univie.ac.at

Abstract

Multilingual Neural Machine Translation (MNMT) models allow translation across multiple languages based on a single system. We study the quality of a domain-adapted MNMT model in the medical domain for English-Romanian with automatic metrics and a human error typology annotation based on the Multidimensional Quality Metrics (MQM) framework. We further expand the MQM typology to include terminology-specific error categories. We compare the out-of-domain MNMT with the in-domain adapted MNMT on a standard test dataset of abstracts from medical publications. The in-domain MNMT model outperforms the out-of-domain MNMT in all measured automatic metrics, and produces fewer errors. We also manually annotate the reference test dataset to study the quality of the reference translations, and we identify a high number of omissions, additions, and mistranslations. We therefore question the assumed accuracy of existing datasets. Finally, we compare the correlation between the COMET, BERTScore, and chrF automatic metrics with the MQM annotated translations; COMET shows a better correlation with the MQM scores.

1 Introduction

Neural Machine Translation (NMT) models have achieved competitive performance on low-resource language pairs, particularly for non-specialised domains (Araabi and Monz, 2020). However, in a high-risk and low-resource domain, like the medical domain, the accurate translation of terminology, alongside the absence of hallucinations and mistranslations are crucial for exchanging information across international healthcare providers or users (Skianis et al., 2020). Multilingual NMT (MNMT) models leverage many language pairs and millions of segments (Johnson et al., 2017) within one system. The inclusion of many language pairs helps to improve the translation quality for low-resource languages by transferring knowledge from high-resource languages via similar cross-lingual word representations. Moreover, domain adaptation techniques are used to adapt MNMT models into new domains (Bérard et al., 2020). However, evaluation studies of MNMT models are focused on automatic metrics without providing insights into the quality of the translation of a specialised domain. These automatic metrics require high-quality reference translations which reflect the specialised terminology and style of a given domain, but such translations are difficult to find. Also, given that translation processes and expertise vary among translators and other text producers, the quality of datasets in different language pairs can differ considerably. In addition, justifiable differences between source and target sentence content are caused by legitimate pragmatic translation strategies. Overall, automatic or even semi-automatic translation data gathering processes are not sophisticated enough yet to improve the quality of the source and/or target content, or filter out content mismatches between source and target sentences before aligning them.

In this paper, we study the quality of a pre-trained MNMT model in the medical domain for a low-resource language pair (English-Romanian). Our goal is to compare an out-of-domain MNMT with a fine-tuned in-domain MNMT in terms of automatic

metrics and a manual error typology annotation. We use a pre-trained model based on MBart (Liu et al., 2020) and fine-tune it with a medical in-domain parallel corpus. In addition, we analyse the quality of the reference test dataset, because errors present in the reference translations can bias the findings of automatic metrics.

We test the models on the English-Romanian language pair with a corpus of medical paper abstracts (Neves et al., 2018). We evaluate the quality of both models with automatic metrics and an error typology annotation based on the Multidimensional Quality Metrics (MQM) framework (Lommel et al., 2014), to which we added terminology-based categories from (Haque et al., 2019). The terminology categories provide a fine-grained discrimination of errors. Finally, we analyse the segment-level correlation between automatic metrics (chrF (Popović, 2015), COMET (Rei et al., 2020), and BERTScore (Zhang et al., 2020)) and the MQM error annotations based on the reference translations.

The fine-tuned MBart model outperforms MBart on the automatic metrics. In addition, the error analysis based on the terminology-enhanced-MQM shows that the fine-tuned model also produces fewer errors than the MBart model. The COMET score shows the highest correlation with the MQM scores. However, it is also important to mention that we identified a total of 157 translation errors in 66 of the 75 reference translation segments, as detailed in section 4.1 below; this questions current assumptions regarding the quality of the reference datasets for our chosen language pair and domain.

2 Background and Related Work

MNMT models are based on transferring parameters or information across multiple languages, where low-resource languages benefit from the high-resource languages. The MNMT model shares a common word representation (i.e., word embeddings) across language pairs. During training, the MNMT model clusters word representations with similar contexts from the high- and low-resource segments (Johnson et al., 2017). The low-resource pairs learn meaningful word representations given the access to a large number of similar contexts from the high-resource language pairs. Multiple languages are processed jointly by indicating the target translation direction on each segment of the multilingual corpora in the input training data by using an artificial token (label *<2target>*). For example, an English-Romanian segment pair would be labelled as follows:

<2ro> It is noted that in some cases increase of blood pressure was documented. → Se remarcă faptul că, în unele cazuri, s-a înregistrat creșterea tensiunii arteriale.

MNMT models outperform standard bilingual baselines on translation quality for low-resource languages (Johnson et al., 2017). MBart is an example of a sequence-to-sequence model pre-trained on monolingual data from 25 languages based on a text reconstruction learning objective for MNMT (Liu et al., 2020). MBart incorporates a monolingual training step before the multilingual MT training for a better initialisation of the translation model. In other words, MBart first learns an improved representation of each language with monolingual data. After that, MBart continues with the multilingual translation training based on parallel data. MBart shows a better translation quality compared to previous MNMT models.

However, most MNMT models are general-purpose systems trained with web-crawled corpora (Liu et al., 2020), and as such they struggle with specialised domains (e.g., medical). Domain adaptation aims to improve the translation performance in specialised domains, where fine-tuning is a low-cost and common technique. Fine-tuning consists of resuming the training of an out-of-domain resource-rich MT model with a poor-resourced in-domain corpus (Chu and Wang, 2018). The resulting model is adapted to work with an in-domain language pair, instead of re-training the MNMT model from scratch (Verma et al., 2022).

MT models are usually evaluated with automatic metrics that take into account fluency and adequacy, by comparing the machine translation output against one or more human reference translations (Papineni et al., 2002). Metrics produce a corpus-level score or a segment-level score for a given MT model (Rei et al., 2020). However, automatic metrics are not designed to identify translation errors in MT outputs, such as errors in terminology, for example (Haque et al., 2019).

On the other hand, error typology evaluation frameworks, such as the Multidimensional Quality Metrics (MQM) (Lommel et al., 2014), are based on manually classifying and annotating errors using predefined categories. The MQM error typology covers high-level error categories, such as: *accuracy*, *style*, *terminology*, *linguistic conventions*, *lo-*

cale conventions, *audience appropriateness*, and *design and markup*. Each high-level category can be further expanded into fine-grained categories; for example, *accuracy* can be further sub-categorised into *mis-translation*, *over-translation*, *omission*, etc. Expert evaluators identify an error in the MT output, label it with a category from the typology, and also assign a severity score to it.

Haque et al. (2019) propose a fine-grained error typology with a focus on terminology. They use a legal domain corpus and develop a gold-standard terminology resource of identified terms based on the previous error typology. Given the terminological richness within the medical domain, we found it relevant to supplement MQM (Lommel et al., 2014) with this terminology-specific error typology. Klubička et al. (2017) compare the quality of phrase-based MT, factored phrase-based MT, and NMT with a manual error annotation of 100 segments with MQM for the English-Croatian language pair. The NMT system was the best performing, with fewer errors produced. Freitag et al. (2021) perform a large-scale study based on MQM annotation of systems from the Workshop on Machine Translation, and they use MQM error-based scores for evaluation. Their error annotation shows a preference for human translations over MT outputs, and the automatic metrics correlate positively with the MQM scores.

3 Experiments

Data For fine-tuning, we use the English-Romanian section from the EMEA parallel corpus (ELG, 2020). The EMEA corpus consists of automatically-aligned PDF documents from the European Medicines Agency. We split the corpus into $775,904$ fine-tuning, and $7,837$ validation segments. We evaluate the MNMT models with the test dataset of similarly-automatically-aligned medical publication abstracts from Medline (Neves et al., 2018), which contains 291 segments.

We use BLEU (Papineni et al., 2002; Post, 2018), chrF (Popović, 2015), BERTScore (Zhang et al., 2020), and COMET (Rei et al., 2020) for automatic evaluation. For human evaluation, we use MQM Core extended with the terminology categories from (Haque et al., 2019), which contain eight terminology-related error categories - Partial error, Source term copied, Inflectional error, Reorder error, Disambiguation issue in target, Incorrect lexical selection, Term drop, and Other error -, and three severity levels with corresponding weights - Minor (1), Major (5) and Critical (10).

MNMT Systems We define general MBart (out-of-domain data) and fine-tuned MBart (in-domain medical data) as MNMT models. We perform our experiments with Fairseq (Ott et al., 2019) using an open-source pre-trained model for MBart[1]. We continue training MBart with the EMEA corpus to adapt it into the medical domain, and we perform model selection using BLEU on the validation split. The settings for the fine-tuned MBart are as follows: Adam with learning rate $3e{-}5$, inverse square root scheduler, $2,500$ warm-up updates, $40,000$ updates, dropout 0.3, attention dropout 0.1, label smoothing 0.2, batch size 2048 tokens (256 maximum tokens per batch, and 8 batches for gradient accumulation), and memory efficient fp16 training. We used a 16GB Tesla T4 GPU from the Google Cloud platform for training[2]. The fine-tuning process took 38 hours to complete.

3.1 Results with Automatic Metrics

Table 1 shows the automatic metrics scores for both models. Fine-tuned MBart outperforms the general model on all metrics. The BLEU and chrF scores are statistically significant $p = .001$ based on bootstrap resampling $1,000$ iterations with sacreBLEU[3].

4 Manual Evaluation Analysis

To gain insights into the translation errors produced by the two models, we show a sample of 12 abstracts with a total of 75 segments to three annotators working collaboratively (Esperança-Rodier et al., 2019); the motivation for this joint in-person annotation approach was to increase agreement for identifying terms and errors. The annotators are native Romanian speakers with in-house and freelance translation experience; moreover, one of the annotators also has in-house and freelance translation experience in the medical domain. The annotators had access to the source, the reference, and the output of the two MNMT systems in order to annotate the reference translation, as well as each MT segment, with error categories (Klubička et al., 2017) using the combination of both typologies:

[1] https://dl.fbaipublicfiles.com/fairseq/models/mbart/mbart.cc25.ft.enro.tar.gz

[2] The scripts for our experiments are available at: https://github.com/mriosb08/medical-NMT-HAITrans

[3] https://github.com/mjpost/sacrebleu

	BLEU ↑ (95% CI)	chrF ↑ (95% CI)	COMET ↑	BERTScore ↑
MBart	22.0 [20.0, 24.0]	51.5 [50.07, 52.93]	0.556	0.834
fine-tuned MBart	**25.8** [23.7, 27.9]	**54.9** [53.29, 56.51]	**0.663**	**0.847**

Table 1: Automatic metrics for MBart and fine-tuned MBart.

MQM (Lommel et al., 2014) and the fine-grained terminology typology (Haque et al., 2019).

To perform the annotation, we set up a translation project in Trados Studio 2021[4] and import the source texts, reference translations, and MT output files as bilingual .xlsx files. We install the freely-available Qualitivity [5] plug-in for Studio; this serves as the environment in which the annotators record any identified errors, their severity level and proposed corrections, along with explanatory comments. At the end of the annotation process, we export a report from Qualitivity containing the full annotation data for the reference segments, as well as MBart and fine-tuned MBart outputs.

4.1 Reference Data Quality

We contacted the authors of the publications abstracts to verify how the reference translations were produced for Romanian. Only four authors replied, listing four different approaches to producing the Romanian abstracts: (1) a translation of the English abstract carried out by a colleague of the author; (2) a separate text written in parallel with the English abstract by the author himself; (3) a translation explicitly undertaken by the publisher, but without informing the author of the exact process; and (4) a text which appeared on the publisher's website without having informed the author that it will appear or who will do the translation. These four different approaches in as many replies explain the level of inconsistency in the quality of the reference translations and raise questions about the confidence which should realistically be placed on datasets gathered automatically, without detailed evaluation.

Table 2 shows the result of our manual error annotation for the gold-standard reference translations, and it is interesting to see Addition, Omission, and Mistranslation accounting for 2/3 of the total errors. If this situation is also present in other publicly-available training and/or testing corpora (for this language pair and domain, but perhaps for other language pairs and domains, too), it would be prudent to temper the current hype and expectations regarding machine translation output quality. Professional human translators make informed decisions whether to omit or add information based on the needs of the target audience, the client's brief, and how the current segment fits into the structure of the overall text, so Additions and Omissions are not always errors *per se*. Such pragmatic decisions cannot be expected of current MT models, though. We therefore need to find much better methods for cleaning training and evaluation datasets, and in the meantime trust professional translators a lot more regarding MT output quality.

To ensure consistency, in our experiment the annotators first evaluated the quality of the reference translation for a given source segment before evaluating the quality of the general and fine-tuned hypotheses for that same source segment. However, some of the errors present hindered this approach, such as the - admittedly rare - cases of identifying in a reference translation a word which does not exist in the target language, or identifying wrong numbers used in the reference compared to the source segments. In human translation evaluation practices, such errors would be categorised as Mistranslations (which is where we have included them in our table); in more recent MT evaluation practices, these errors would be categorised as Hallucinations, although the MQM framework did not have such a category at the time of our experiment, so we needed to add it manually to our typology. In any case, seeing how reference translations can contain such inaccuracies, it is less surprising to notice further Hallucinations in MT output. In our experiment, working horizontally on the reference translations and MT hypotheses for each segment, and having three experienced translators collaborate synchronously ensured as much consistency and agreement as could possibly be expected for such a high-effort and time-consuming task.

Given the surprisingly high number of errors identified in the reference translations for the 75 annotated segments, our MT error annotations also

[4] https://www.trados.com/products/trados-studio/

[5] https://community.rws.com/product-groups/trados-portfolio/rws-appstore/w/wiki/2251/qualitivity

Error Type	Reference
Terminology – Partial error	4
Terminology – Source term copied	12
Terminology – Disambiguation issue in target	4
Terminology – Incorrect lexical selection	3
Terminology – Other error	1
Accuracy – Mistranslation	21
Accuracy – Omission	36
Accuracy – Addition	51
Fluency – Mechanical – Grammar	4
Fluency – Content – Stylistics	9
Fluency – Content – Register	1
Fluency – Mechanical – Locale convention	1
Fluency – Content – Inconsistency	1
Fluency – Mechanical – Typography	2
Verity – Completeness	7
Total	157

Table 2: Total number of errors in the Reference Translation for each category.

took into account the corrections which could have been made to the gold-standard published reference translations. Once again, the presence of these errors highlights the importance of not taking for granted the accuracy of existing datasets, as over-reliance on reference sets of an assumed good quality can undermine the result of the evaluation exercise. This can also lead to important discrepancies between the perception regarding the usefulness of individual MT models, and the experience of professional translators using them.

4.2 MNMT Systems Quality

The total number of errors for general-model MBart and fine-tuned MBart are 234 and 140 respectively, demonstrating the improvement brought about by the fine-tuning process with in-domain data. Interestingly, when comparing the gold-standard translations and the fine-tuned MBart system output, we notice 17 fewer errors in the MT output. However, as we have mentioned before, what was labelled as an error for consistency purposes when evaluating the gold-standard was at times justified by the wider translation context. Table 3 shows the number of errors divided by severity for each category present in the abstracts.

The fine-tuned MBart model produces fewer errors than the general model on most categories.

Table A1 shows annotated examples for the *Accuracy*, *Fluency* and *Hallucination* error categories for the fine-tuned MBart. Fewer overall errors were recorded for all the *Accuracy*, *Fluency*, and *Hallucination* categories, with the exception of *Accuracy – Omission* and *Fluency – Mechanical – Grammar* error types. While *Accuracy – Omission* leads to entire sentences being left out, *Fluency – Mechanical – Grammar* displays instances of mismatched feminine and masculine articles (**un pacientă** instead of **o pacientă**), determinate for indeterminate articles (**tratamentul** instead of **tratament**), as well as incorrect prepositions and agreements (**de pacienţi** instead of **ale pacienţilor**). In the *Accuracy – Mistranslations* category, in addition to calques (**descărcat** instead of **externat**; **evolueze** instead of **apară**), we also note mistranslations of some of the English (EN) hedging devices: in some segments, they are eliminated altogether ("We investigated **the extent to which** anthropometric measurements **can be** used to identify"); in other contexts, they are strengthened (in some examples, the EN **could be**, which should be rendered into Romanian (RO) as **ar putea fi**, becomes **poate fi** in RO, which is the equivalent of **can be** in EN). The *Fluency – Content – Stylistics* and *Register* categories contain almost exclusively minor non-idiomatic or informal style choices. *Fluency – Content – Inconsistency* refers to a document-level inconsistency regarding gender: replacing the feminine noun (**pacienta**) with its masculine form (**pacientul**). The *Hallucination* category includes errors which we break into three phenomena: a) direct borrowings from English inflected for RO gender and number (**auriclelui** instead of **auricular**); b) made-up recomposed words (**pre-anaetică** instead of **pre-anestezic**; **raţii** instead of **şobolanii**; **nazofaringinei** instead of **nazofaringelui**; or **adnexectomie** instead of **anexectomie**), and c) changes in numbers (**0,07** instead of **0,17**). All these point to challenges with the setup of the Byte pair encoding (BPE) vocabulary in NMT.

We consider *Terminology* errors central to medical MT evaluation and development. Although an in-depth analysis of such errors is beyond the scope of the current paper, we notice that the fine-tuned model produces fewer terminology-related errors. However, it still performs worse than the general MBart in the following terminology-related categories: *Inflectional error*, *Reorder error*, and *Other*. In Table A2 in the appendix, we show a random selection of source and fine-tuned MBart examples for each *Terminology* error category, and highlight the annotated errors for each category. Within the *Terminology – Other error* category, we identify

Error Type	MBart ↓ minor	major	critical	fine-tuned MBart ↓ minor	major	critical
Terminology – Partial error	5	11	25	**5**	**7**	**11**
Terminology – Source term copied	1	20	1	**0**	**8**	**1**
Terminology – Inflectional error	**0**	**2**	**0**	3	1	0
Terminology – Reorder error	**0**	**0**	**1**	1	1	1
Terminology – Disambiguation issue in target	1	3	10	**0**	**2**	**4**
Terminology – Incorrect lexical selection	1	1	7	**0**	**0**	**6**
Terminology – Other error	**1**	**0**	**8**	0	0	13
Accuracy – Mistranslation	7	10	11	**2**	**9**	**6**
Accuracy – Omission	**0**	**1**	**1**	0	0	3
Accuracy – Addition	1	0	1	**0**	**0**	**1**
Accuracy – Untranslated	0	2	0	**0**	**0**	**0**
Fluency – Mechanical – Grammar	**13**	**4**	**0**	17	4	0
Fluency – Content – Stylistics	11	0	0	**10**	**0**	**0**
Fluency – Content – Register	**1**	3	0	2	**1**	**0**
Fluency – Mechanical – Locale convention	3	0	5	**1**	**0**	**2**
Fluency – Content – Inconsistency	2	0	0	2	0	0
Fluency – Mechanical – Typography	3	0	0	**2**	**0**	**0**
Fluency – Mechanical – Spelling	2	0	0	2	0	0
Fluency – Unintelligible	1	0	1	**0**	**0**	**0**
Hallucination	0	4	49	**0**	**1**	**11**
Total	53	61	120	47	34	59

Table 3: Total errors in MBart and fine-tuned MBart with severity for each category.

two phenomena regarding the treatment of English borrowings and acronyms, as well as evidence of hallucination. The first phenomenon observed is that source terms, including acronyms, are translated, even where a borrowing from English would be the correct translation strategy (**arsură** instead of **burst**; **SSO** instead of **OS**). Secondly, acronyms corresponding to terms with a translation into Romanian are randomly recomposed (**SMO** instead of **MODS**; **RF** instead of **RL**). This points again to challenges with the setup of the Byte pair encoding (BPE) vocabulary in NMT (Araabi et al., 2022; Lignos et al., 2019).

4.3 Automatic Metrics Correlation Analysis

We perform a segment-level correlation analysis between BERTScore, COMET, and chrF with the MQM scores from the manual error annotation. We select metrics with segment-level output, thus not including corpus-level metrics such as BLEU. We use the score and severity weights defined by Unbabel (Freitag et al., 2021) for the MQM typology. The MQM score (↑) is defined as follows:

$$\text{MQM} = 100 \cdot \left(1 - \frac{10 \cdot \text{critical} + 5 \cdot \text{major} + \text{minor}}{\text{tokens}}\right), \quad (1)$$

where **critical**, **major**, and **minor** represent the number of errors annotated, and the number of **tokens** in a segment. Figure 1 shows the Kendall Tau and Spearman correlation with the segment-level MQM scores. COMET, without any medical domain fine-tuning, has the highest correlation with the MQM scores.

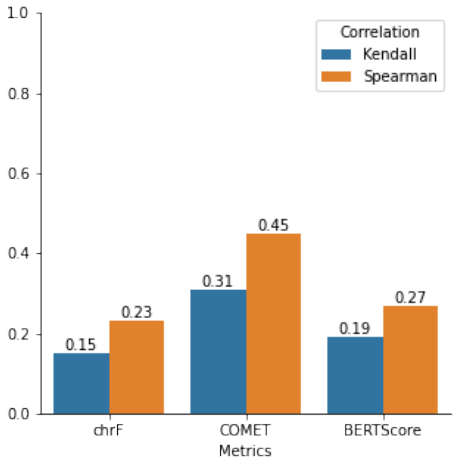

Figure 1: Kendall Tau and Spearman segment-level correlation between automatic metrics chrF, COMET, and BERTScore with the MQM scores.

Further work will investigate these correlations in the case of a *corrected* gold-standard because, given the large number of differences (some erroneous, some justified) between the source and the target segments in the gold standard, we believe it is an unfair task to evaluate translation hypotheses

proposed by MT models against reference translations produced by a variety of methods through a variety of workflows and which, as a result, often do not contain all the information from the source, or which contain additional information unavailable to the MT models, or contain a wide range of translation errors.

5 Conclusions and Future Work

We quantified the impact of domain adaptation on MBart in the medical domain for English-Romanian. The fine-tuned MBart outperforms the general model with automatic metrics and produces fewer errors in the relatively small sample (75 segments belonging to the 12 medical publications abstracts) we annotated.

We show that the gold-standard reference translations provided in the datasest contain a high number of errors. Blindly assuming good quality of the reference translations when performing evaluations can be problematic and the community should be more open about the shortcomings of existing data gathering methods, and incorporate translators' contributions to improving test and training datasets to a greater extent.

While fewer *Terminology* errors were recorded in the *Partial error, Source term copied, Disambiguation issue in target, Incorrect lexical selection*, and *Term drop* categories, in the three remaining ones (*Inflectional error, Reorder error*, and *Other error*), the fine-tuned MBart output actually contained more errors than the general MBart output. Of these three categories, the *Inflectional error* and *Other error* items present in the fine-tuned MBart output are related to the BPE vocabulary. In future work, we plan to extend the BPE vocabulary in MBart (Berard, 2021) to cope with in-domain terminology.

COMET shows a higher correlation with MQM scores compared to other automatic metrics. COMET can be an option for evaluating NMT systems for the medical domain, and in particular for scientific abstracts. At the same time, reference translation datasets need to be prepared much more carefully, keeping in mind shortcomings in the translation output produced by NMT models.

Finally, it is essential to raise awareness among machine translation post-editors, as well as clients, that errors may persist in MT output even after fine-tuning. Errors in NMT output remain difficult to identify due to the apparent fluency of the output, and can thus be overlooked even by subject-matter experts. It is for these reasons that translators should be able to work in post-editing interfaces which stimulate their attention to such errors. It is also why synchronous collaborative translation, revision, and post-editing workflows which use newer, more ergonomic and interactive technologies should be promoted and adopted to a much greater extent.

Acknowledgements

The GPU used for this research was sponsored by the Google Cloud Research Credits Program.

References

Araabi, Ali and Christof Monz. 2020. Optimizing Transformer for Low-Resource Neural Machine Translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3429–3435, Barcelona, Spain (Online), December. International Committee on Computational Linguistics.

Araabi, Ali, Christof Monz, and Vlad Niculae. 2022. How Effective is Byte Pair Encoding for Out-Of-Vocabulary Words in Neural Machine Translation? In *Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 117–130, Orlando, USA, September. Association for Machine Translation in the Americas.

Berard, Alexandre. 2021. Continual Learning in Multilingual NMT via Language-Specific Embeddings. In *Proceedings of the Sixth Conference on Machine Translation*, pages 542–565, Online, November. Association for Computational Linguistics.

Bérard, Alexandre, Zae Myung Kim, Vassilina Nikoulina, Eunjeong Lucy Park, and Matthias Gallé. 2020. A Multilingual Neural Machine Translation Model for Biomedical Data. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online, December. Association for Computational Linguistics.

Chu, Chenhui and Rui Wang. 2018. A Survey of Domain Adaptation for Neural Machine Translation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1304–1319, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

ELG, ELG. 2020. ELG - Bilingual corpus made out of PDF documents from the European Medicines Agency, (EMEA), https://www.ema.europa.eu, (February 2020) (EN-RO).

Esperança-Rodier, Emmanuelle, Francis Brunet-Manquat, and Sophia Eady. 2019. ACCOLÉ: A

Collaborative Platform of Error Annotation for Aligned Corpora. In *Translating and the computer 41*, Londres, United Kingdom, November.

Freitag, Markus, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. 2021. Results of the WMT21 Metrics Shared Task: Evaluating Metrics with Expert-based Human Evaluations on TED and News Domain. In *Proceedings of the Sixth Conference on Machine Translation*, pages 733–774, Online, November. Association for Computational Linguistics.

Haque, Rejwanul, Md Hasanuzzaman, and Andy Way. 2019. Investigating Terminology Translation in Statistical and Neural Machine Translation: A Case Study on English-to-Hindi and Hindi-to-English. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 437–446, Varna, Bulgaria, September. INCOMA Ltd.

Johnson, Melvin, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Klubička, Filip, Antonio Toral, and Víctor M. Sánchez-Cartagena. 2017. Fine-grained human evaluation of neural versus phrase-based machine translation. *The Prague Bulletin of Mathematical Linguistics*, 108(1):121–132, June.

Lignos, Constantine, Daniel Cohen, Yen-Chieh Lien, Pratik Mehta, W. Bruce Croft, and Scott Miller. 2019. The Challenges of Optimizing Machine Translation for Low Resource Cross-Language Information Retrieval. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3497–3502, Hong Kong, China, November. Association for Computational Linguistics.

Liu, Yinhan, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual Denoising Pre-training for Neural Machine Translation. *Transactions of the Association for Computational Linguistics*, 8:726–742, December.

Lommel, Arle, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional Quality Metrics (MQM): A Framework for Declaring and Describing Translation Quality Metrics. *Revista Tradumàtica: tecnologies de la traducció*, (12):455–463. Publisher: Universitat Autònoma de Barcelona.

Neves, Mariana, Antonio Jimeno Yepes, Aurélie Névéol, Cristian Grozea, Amy Siu, Madeleine Kittner, and Karin Verspoor. 2018. Findings of the WMT 2018 Biomedical Translation Shared Task: Evaluation on Medline test sets. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 324–339, Belgium, Brussels, October. Association for Computational Linguistics.

Ott, Myle, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A Fast, Extensible Toolkit for Sequence Modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota, June. Association for Computational Linguistics.

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 of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, 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.

Post, Matt. 2018. A Call for Clarity in Reporting BLEU Scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium, October. 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.

Skianis, Konstantinos, Yann Briand, and Florent Desgrippes. 2020. Evaluation of Machine Translation Methods applied to Medical Terminologies. In *Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis*, pages 59–69, Online, November. Association for Computational Linguistics.

Verma, Neha, Kenton Murray, and Kevin Duh. 2022. Strategies for Adapting Multilingual Pre-training for Domain-Specific Machine Translation. In *Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 31–44, Orlando, USA, September. Association for Machine Translation in the Americas.

Zhang, Tianyi, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*.

A Annotated Examples

Category	Severity	Source	fine-tuned MBart
Accuracy–Mistranslation	major	After the gastrointestinal decontamination, including gastric lavage, activated charcoal and cathartics, the outcome was favourable and 48 hours after admission the patient was discharged.	După decontaminarea gastrointestinală, incluzând lavaj gastric, cărbune activat și catartice, rezultatul a fost favorabil și la 48 de ore după admitere pacientul a fost **descărcat**. [instead of **externat**]
Accuracy–Omission	major	A hole was drilled in the skull over the frontal cortex and electrodes were inserted in order to record the local field potentials.	[**OMISSION**]
Accuracy–Addition	major	Using patients files we recorded the following data: 30 day mortality, development of acute respiratory distress syndrome (ARDS) and MODS, local infectious complications (LIC), intensive care unit length of stay (ICU LOS), days of mechanical ventilation (MV), units of red blood cells units/48 h (RBC).	Utilizând dosarele pacienţilor, am înregistrat următoarele date: mortalitate cu durata de 30 zile, apariţia sindromului de detresă respiratorie acută (SRA) şi MODS, complicaţii infecţioase locale (LIC), durata de şedere la unitatea de terapie intensivă (ICU LOS), zile de ventilaţie mecanică (MV), unităţi de celule roşii **în sânge** /48 ore (RBC).
Fluency–Mechanical–Grammar	minor	Several theories have been proposed in terms of what causative factors are associated with poor outcome in polytrauma patients.	Au fost propuse mai multe teorii în ceea ce privește factorii cauzatori asociați cu rezultate slabe la pacienții cu Au fost propuse mai multe teorii în ceea ce privește factorii cauzatori asociați cu rezultate slabe la pacienții cu **politrauma**. [instead of **politraumă**]
Fluency–Content–Stylistics	minor	The last 20 years have been dedicated to extensive research regarding the pathophysiology of trauma and the consequences of interventions that follow.	Ultimii 20 de ani au fost dedicaţi unei cercetări extinse privind profilaxia traumatismului şi consecinţele intervenţiilor **care urmează**. [instead of **ulterioare**]
Fluency–Content–Register	minor	However, there are significant barriers to more generalized use, but evidence continues to evolve that might one day make this practice a standard of care in the perioperative period.	Cu toate acestea, există bariere semnificative în calea utilizării mai generalizate, dar dovezile continuă să evolueze care ar putea transforma **într-o zi** [instead of **în viitor**] această practică într-un standard de îngrijire în perioada perioperatorie.
Fluency–Mechanical–Locale convention	major	We observed significantly higher mortality in the DCO shock group (25%) compared with the other two groups (ETC - 9.4%; DCO without shock - 6.7%; p = 0.042/0.015). Similar results for: ICU LOS (16.29 ± 6.7 versus 9.92 ± 4.7 and 10 ± 3.9; p = 0.001/0.002), days of MV (10.29 ± 5.7 versus 6.83 ± 4.7 and 6.8 ± 3.4; p = 0.007/0.04), units of RBC/48 h (15.04 ± 4.3 versus 8.08 ± 4.3 and 7.33 ± 1.5; p = 0.007/0.04).	Rezultate similare pentru: LOS ICU (16,29 ± 6,7 faţă de 9,92 ± 4,7 şi 10 ± 3,9; p = 0,001/0,002), zile de MV (10,29 ± 5,7 faţă de 6,83 ± 4,7 şi 6,8 ± 3,4; p = 0,007/0,04), unităţi de RBC/48 ore (15,04 ± **4.3** [instead of **4,3**] faţă de 8,08 ± 4.3 şi 7,33 ± 1,5; p = 0,007/0,04).
Fluency–Content–Inconsistency	minor	The patient underwent coronary catheterization which confirmed a coronary fistula connecting CX with a superior vena cava-right atrium junction, with a hemodynamic significant left- to-right shunt.	**Pacientul** [instead of **pacienta**] a fost supus cateterizării coronariene care a confirmat o fistulă coronariană care leagă CX de o joncțiune atrială venă superioară cava-dreapta, cu un shunt hemodinamic semnificativ de la stânga la dreapta.
Fluency–Content–Typography	minor	This is a retrospective study of severe polytrauma patients with femoral shaft fractures admitted to the intensive care unit of the Emergency clinical Hospital of Bucharest and treated from an orthopaedic point of view by either Damage Control Orthopaedics (DCO) or Early Total Care (ETC) principles.	Acesta este un studiu retrospectiv la pacienţi cu politraum sever, cu fracturi ale căilor femurale, internaţi în unitatea de terapie intensivă a **Spitalului clinic de urgenţă** [instead of **Spitalului Clinic de Urgenţă**] din Bucureşti şi trataţi din punct de vedere ortopedic, fie conform principiilor de control al deteriorării (DCO), fie conform principiilor de îngrijire totală precoce (ETC).
Fluency–Content–Spelling	minor	Decreased plasma concentrations of antioxidants, correlated with a disturbance of the redox balance are responsible for the installation of the phenomenon called oxidative stress (OS).	Scăderea concentraţiilor plasmatice de **antioxidanti** [instead of **antioxidanţi**], corelată cu o tulburare a echilibrului redox, este responsabilă de instalarea fenomenului numit stres oxidativ (SSO).
Hallucination	major	Rats were maintained in deep level anaesthesia (burst-suppression).	**Ratii** [instead of **şobolanii**] s-au menţinut în anestezie profundă (supresie pulmonară).

Table A1: Fine-tuned MBart annotated examples for each Accuracy, Fluency and Hallucination error category. The additional errors present in these examples have not been highlighted in this table.

Category	Severity	Source	Target (fine-tuned MBart)
Partial error	critical	The DX-OSA score may be useful for identifying obese patients with significant OSA who require CPAP (continuous positive airway pressure) treatment, and CPAP could be commenced without the need for polysomnography, therefore, without delaying surgery.	Scorul DX-OSA poate fi util pentru identificarea pacienţilor obezi cu OSA semnificativă care necesită tratament cu CPAP (**tensiune arterială continuă pozitivă** [instead of **presiune pozitivă continuă în căile aeriene**]), iar CPAP poate fi început fără a fi necesară polisomnografie, prin urmare, fără a întârzia intervenţia chirurgicală.
Source term copied	major	The objectives of this study were to reveal possible relations between antioxidant therapy and a number of serum biochemical variables (ALT, AST, APPT, LDH, urea, leukocytes, platelets), the length of mechanical ventilation, the time spent in the ICU, and the mortality rate in major trauma patients.	Obiectivul acestui studiu a fost să evidenţieze posibilele relaţii dintre tratamentul cu antioxidanti şi o serie de variabile biochimice serice (ALT, AST, **APPT** [instead of **APTT**], LDH, uree, leucocite, trombocite), durata ventilaţiei mecanice, timpul petrecut în ICU şi rata mortalităţii la pacienţii cu traumatisme majore.
Inflectional error	minor	Two of these drugs, duloxetine and venlafaxine, are used also in chronic pain management.	Două dintre aceste medicamente, **duloxetină** şi **venlafaxină** [instead of **duloxetina** şi **venlafaxina**], sunt utilizate şi în tratamentul durerii cronice.
Reorder error	major	Although not statistically significant, MODS and ARDS incidences were higher in the DCO shock group: MODS (41.7% versus 22.6% and 20%; p = 0.08/0.17), ARDS (29.2% versus 17% and 20%; p = 0.22/0.53).	Deşi nu au fost semnificative statistic, incidenţele MODS şi ARDS au fost mai mari în **grupul cu şoc DCO** [instead of **grupul DCO cu şoc**]: MODS (41,7% faţă de 22,6% şi 20%; p = 0,08/0,07), ARDS (29,2% faţă de 17% şi 20%; p = 0,22/0,53).
Disambiguation issue in target	major	The drug's efficacy results from its modulating effect on the descending inhibitory pain pathways and the inhibition of the nociceptive input.	Eficacitatea medicamentului rezultă din efectul său de modulare asupra **căilor de durere inhibatoare descendente** [instead of **căilor descendente inhibitorii ale durerii**] şi inhibarea contribuţiei nociceptive.
Incorrect lexical selection	critical	These results correlate with a higher trauma score in these patients, more serious lesions requiring several damage control procedures.	Aceste rezultate sunt corelate cu un **scor traumatic** [instead of **gravitatea traumatismelor**] mai mare la aceşti pacienţi, leziunile mai grave necesitând mai multe proceduri de control al leziunilor.
Other	critical	The global cortical connectivity increased during the burst periods.	Conectivitatea corticală globală a crescut în timpul perioadelor de **arsură**.[instead of **burst**]
Other	critical	Decreased plasma concentrations of antioxidants, correlated with a disturbance of the redox balance are responsible for the installation of the phenomenon called oxidative stress (OS).	Scăderea concentraţiilor plasmatice de antioxidanti, corelată cu o tulburare a echilibrului redox, este responsabilă de instalarea fenomenului numit stres oxidativ (**SSO**) [instead of **OS**].
Other	critical	Once the "two event model" was accepted, it became clear that patients although initially resuscitated, but in a vulnerable condition, have a high risk that a secondary aggression (for example, surgical interventions) would precipitate a state of hyperinflammation and early multiple organ dysfunction syndrome (MODS).	Odată ce „modelul celor două evenimente" a fost acceptat, a devenit clar că pacienţii, deşi iniţial resuscitaţi, dar aflaţi într-o stare vulnerabilă, prezintă un risc crescut ca o agresivitate secundară (de exemplu intervenţii chirurgicale) să precipite o stare de hiper inflamaţie şi sindrom de disfuncţie multiplă precoce (**SMO**) [instead of **MODS**].
Other	critical	The biochemical processes of bioproduction of free radicals (FR) are significantly increasing in polytrauma patients.	Procesele biochimice de bioproducţie a radicalilor liberi (**RF**) [instead of **RL**] cresc semnificativ la pacienţii cu politrauma.

Table A2: Fine-tuned MBart annotated examples for each *Terminology* error category. The additional errors present in these examples have not been highlighted in this table.