What Do Post-Editors Correct? A Fine-Grained Analysis of SMT and NMT Errors

Sergi Álvarez-Vidal
Antoni Oliver
Toni Badia

revista.tradumàtica
tecnologies de la traducció

Sergi Alvarez-Vidal
Universitat Pompeu Fabra
salvarezvid@uoc.edu;
ORCID:
0000-0002-6355-4559

Antoni Oliver
Universitat Oberta de Catalunya
aoliverg@uoc.edu;
ORCID:
0000-0001-8399-3770

Toni Badia
Universitat Pompeu Fabra
toni.badia@upf.edu;
ORCID:
0000-0002-9429-5940

Abstract

The recent improvements in neural MT (NMT) have driven a shift from statistical MT (SMT) to NMT. However, to assess the usefulness of MT models for post-editing (PE) and have a detailed insight of the output they produce, we need to analyse the most frequent errors and how they affect the task. We present a pilot study of a fine-grained analysis of MT errors based on post-editors corrections for an English to Spanish medical text translated with SMT and NMT. We use the MQM taxonomy to compare the two MT models and have a categorized classification of the errors produced. Even though results show a great variation among post-editors' corrections, for this language combination fewer errors are corrected by post-editors in the NMT output. NMT also produces fewer accuracy errors and errors that are less critical.

Keywords: machine translation; MT; NMT; post-editing; neural machine translation; error taxonomy

Resum

Les millores recents en la TA neuronal (TAN) han impulsat un canvi de la TA estadística (TAE) a la TAN. Tanmateix, per avaluar la utilitat dels models de TA per a la postedició (PE), és fonamental analitzar els errors més freqüents i com afecten la tasca. Presentem un estudi pilot d'una anàlisi detallada dels errors de la TA basat en correccions de postedició d'un text mèdic traduït de l'anglès al castellà amb TAE i TAN. Hem utilitzat la taxonomia MQM per comparar els dos models de TA i hem classificat els errors produïts. La nostra anàlisi també inclou una avaluació de la variació entre els posteditors, que se centra en els passatges amb una major variació en la postedició.

Paraules clau: traducció automàtica; TA; TAN; postedició; traducció automàtica neuronal; taxonomia d'errors

Resumen

Los avances recientes en TA neuronal (TAN) han producido un giro desde la TA estadística (TAE) hacia la TAN. Sin embargo, para evaluar la utilidad de los modelos de TA para la posedición, es imprescindible analizar los errores

Rebuda: 30 de gener de 2021 / Acceptació: 2 de juliol de 2021 / Publicació avançada: 15 de desembre de 2021

más frecuentes y cómo afectan a esta tarea. Presentamos el estudio piloto de un análisis pormenorizado de errores en TA basado en las correcciones realizadas por los poseditores en la traducción de un texto médico realizada del inglés al castellano mediante TAE y TAN. Utilizamos la taxonomía MQM para comparar los dos modelos de TA y obtener una clasificación categorizada de los errores resultantes. Nuestro análisis incluye también una evaluación de las diferencias entre poseditores, centrada en los pasajes en los que la posedición presentaba mayor disparidad.

Palabras clave: traducción automática; TA; TAN; posedición; traducción automática neuronal; taxonomía de errores

1. Introduction

Post-editing (PE) of machine translation (MT) has become a very common practice in the translation industry (Lommel and Depalma, 2016) in the last decade. PE increases productivity when we compare it with human translations (Aranberri, 2014) without having a negative impact on quality (Plitt and Masselot, 2010) and it also involves a reduction of costs, as post-editing is usually paid less per word than translating from scratch (Guerberof, 2009). Post-editors "edit, modify and/or correct pre-translated text that has been processed by an MT system from a source language into (a) target language(s)" (Allen 2003: 293).

Statistical MT (SMT) has been well established as the dominant approach in MT for many years. However, recent evaluation campaigns (Bojar et al., 2018; Barrault et al., 2019) have shown neural MT (NMT) outperforms previous systems in terms of quality. These results have driven a technological shift from SMT to NMT in most translation industry scenarios. When assessing the usefulness of MT models for PE, it is essential to analyse the most frequent errors and how they affect the task. Although recent studies suggest that NMT reduces errors and produces more fluent outputs (Bentivogli et al., 2016; Castillo et al., 2017; Toral and Sánchez-Cartagena, 2017), each error type affects the PE effort differently (Daems et al., 2017).

Error annotation has been used to study the quality of the MT products (Vilar et al., 2006; Costa et al., 2015, Popovic 2018) and to investigate whether an MT output is fit for post-editing (Denkowski and Lavie, 2012). However, it is usually conducted as a separate task from post-editing, even though these two tasks are highly related. In fact, post-editing can be understood as an implicit error annotation, as the edits post-editors enter are intended as corrections of translation errors (Popovic and Arcan, 2016). Even though translators' edits may reflect preferential changes or style and do not always correspond to errors (Koponen, 2013; Koponen and Salmi, 2017; Koponen et al., 2019), we have annotated the actual modifications introduced into the raw MT output, as many correct translations for the same source text are possible. Moreover, analysing corrections from different translators working on the same text can give valuable insight to better

understand the variability patterns among translators. That is, how different professionals modify the raw MT output.

We present a pilot study of a fine-grained analysis of MT errors based on post-editors' corrections for an English to Spanish medical text translated with SMT and NMT. Our goal is to study the type of errors post-edited for these two MT models for this text type and language combination, and analyse the differences among translators post-editing the same MT output. In Section 2, we present previous work analysing the differences between these two MT models and the errors they produce. Then, we present the methodology we used, both the MT systems trained in our study and the PE set-up. In the following section we detail the customized MQM taxonomy we used for the error annotation process and then we present the results of the annotation for each post-edited version. Finally, we include a discussion of the results and detail our future research.

2. SMT versus NMT

Automatic metrics such as BLEU (Papineni et al., 2002) are currently used to assess MT quality and have been used to show that in many cases NMT models outperform SMT systems. For example, Junczys-Dowmunt et al. (2016) studied 30 different translation directions from the United Nations Parallel Corpus and Wu et al. (2016) assessed the quality of NMT and SMT outputs for Wikipedia entries translated with these two MT models. However, metrics like BLEU exploit mainly surface-matching characteristics that are largely insensitive to more subtle nuances and have been shown to underestimate NMT quality compared to the assessment conducted with rankings obtained by human reviewers (Shterionov et al., 2018). Moreover, more recent evaluation campaigns have confirmed NMT architectures produce better results with different automatic metrics and human evaluations (Bojar et al., 2018; Barrault et al., 2019). These campaigns, however, do not directly address errors and a more nuanced analysis of the errors produced is needed.

One of the first papers analysing the impact of SMT and NMT in post-editing was Bentivogli et al. (2016). They carried out a study on post-editing NMT and SMT outputs of English to German translated TED talks. They concluded that in general NMT decreased the post-editing effort, but SMT yielded better results for longer segments. Toral and Sánchez-Cartagena (2017) broadened the scope of the former paper adding different language combinations and metrics, and they concluded that although NMT yielded better quality results in general, it was negatively affected by sentence length, and the improvement of the results was not always perceivable in all language pairs. Bentivogli et al. (2018) extended the scope of their previous paper increasing the number of systems analysed, adding an extra language pair and conducting a three-category error analysis on the results. They confirmed the increase in quality for NMT systems and concluded most errors produced in NMT outputs were lexical, especially proper nouns.

Castilho et al. (2017) reported on a comparative study of phrase-based SMT (PBSMT) and NMT, with four language pairs and different automatic metrics and human evaluation

methods. It highlighted some strengths and weaknesses of NMT, which in general yielded better results. It focused especially on post-editing and used PET (Aziz et al., 2012), a computer-assisted translation tool which enables recording time and keystrokes, to compare educational domain outputs from both systems using different metrics. They concluded that NMT reduced word order errors and improved fluency for certain language pairs, so fewer segments required post-editing, especially because there was a reduction in the number of morphological errors. However, they did not detect a decrease in PE effort nor a clear improvement in omission and mistranslation errors.

Koponen et al. (2019) presented a comparison of PE changes performed on NMT, rule-based MT (RBMT) and SMT output for the English-Finnish language combination. A total of 33 translation students edited in this English-to-Finnish PE experiment. It outlined the strategies participants adopt to post-edit the different outputs, which contributed to the understanding of NMT, RBMT and SMT approaches. It also concluded that PE effort was lower for NMT than SMT.

Klubička et al. (2017, 2018) compared the errors produced by an English-Croatian pure phrase-based, factored phrase-based and NMT system performing a manual evaluation via error annotation of the systems' outputs. Two annotators used a metric compliant with MQM (multidimensional quality metrics) and results showed that NMT reduced the number of errors considerably. Ye and Toral (2020) also conducted a fine-grained human evaluation to compare the transformer and recurrent approaches to neural MT for the English-Chinese combination. They followed a tailored MQM taxonomy and observed the transformer produced an overall better translation reducing the number of errors related to accuracy, fluency and comprehensibility.

Even though a product-based analysis of the errors produced in the MT output can help to understand the MT quality, it is not enough to measure the actual effort involved in PE, which Krings (2001) defined as the sum of three aspects: temporal, technical and cognitive effort. Some errors may be easy to identify but require a lot of editing, while others can be easily corrected but may be difficult to spot or solve. For example, lack of coherence, shifts in meaning and structural issues have proved to be good indicators of post-editing effort (Daems et al., 2017).

The analysis and classification of MT errors has been always used as a valuable tool to improve MT systems. Some automatic or semiautomatic tools have been developed to conduct this task. Addicter (Zeman et al., 2011) is a tool for the automatic detection and display of common translation errors which uses a first-order Markov model for aligning reference words with hypothesis words. Hjerson (Popovic, 2011) uses WER alignments and compares the sets of words identified as erroneous due to a mismatch with the reference. However, error classification is usually conducted manually because currently available tools are still not able to distinguish detailed error classes, and are prone to confusions between mistranslations, omissions and additions. This task is usually performed by annotators who identify the errors of the MT output with or without a reference translation. However, with the widespread use of post-editing in the translation workflow, the analysis of post-editing corrections is receiving more and more attention (Popovic, 2018), and can also be understood as an implicit error annotation, as the

edits post-editors introduce are intended as corrections of translation errors (Popovic and Arcan, 2016).

Vilar (2006) suggested one of the first error classifications for MT analysis focused on identifying the main MT problems and grouping errors into five major categories: missing words, word order, incorrect words, unknown words and punctuation. Farrús et al. (2010) designed an error taxonomy with five broad categories for SMT outputs from the Catalan-Spanish language combination. They correlated the different categories with human evaluations and noticed that semantic errors influenced the most in the perception of quality. Federico et al. (2014) used a similar taxonomy focused on detecting MT errors for translations from English into Arabic, Chinese and Russian. Costa et al. (2015) reported an error taxonomy tailored for Romance languages. In their study, highly ranked sentences clearly showed low number of grammatical errors, and a high inter-annotator agreement between two annotators was reported.

The translation industry has also developed error taxonomies which have been included in quality metrics. For the purpose of evaluation, many companies use error-based models that seek to "identify errors, classify them, allocate them to a severity level and apply penalty points with a view to deciding whether or not the translation meets a specific pass mark." (O'Brien, 2011a, p. 58)

The LISA QA metric[1] was initially designed to promote the best translation and localization methods for the software and hardware industries. Although it is no longer in use, its methods are still used in translation quality evaluation. This metric includes three severity levels, but there is no weighting. It consists of a set of 20, 25 or 123 error categories, depending on how they are counted. The SAE J2450 metric originated in the automotive industry and includes seven primary error categories which cover such areas as terminology, meaning, structure, spelling, punctuation, completeness, etc. and two severity levels. In contrast to LISA, it focuses on linguistic quality and includes no formatting or style issues. It also includes two meta-rules to help evaluators make a decision in case of ambiguity.

The TAUS Dynamic Quality Framework (DQF)[2] uses different tools, which include an error taxonomy, for the evaluation of translation quality. It was recently harmonized with the Multidimensional Quality Metrics (MQM) (Lommel et al., 2014), which will be explained in detail in Section 4, to offer translation professionals and researchers a unified model.

3. Methodology

3.1. MT Systems

First of all, we trained a SMT and a NMT system with the same medical-domain corpora to produce the MT output. For this purpose, we used ModernMT (Germann et al., 2016) version 2.4. This version allows training both statistical and neural MT systems. We used

1 http://producthelp.sdl.com/SDL TMS 2011/en/Creating and Maintaining Organizations/Managing QA Models/LISA QA Model.htm

2 https://www.taus.net/data-for-ai/dqf

the default options for this version. One of the salient characteristics of ModernMT is that it can take into account the context of the sentence to be translated for both approaches. We used the previous and the next segment (except for the first segment and the last segment, where we have taken into account the next segment and the previous segment only, respectively) as context. Short contexts are usually enough to calculate the context vector used by ModernMT.

To train the systems we compiled all, to our knowledge, publicly available corpora in the English-Spanish pair. We also created several corpora from websites with medical content:

- The EMEA[3] (European Medicines Agency) corpus.
- The IBECS[4] (Spanish Bibliographical Index in Health Sciences) corpus.
- Medline Plus[5]: we have compiled our own corpus from the web and combined this with the corpus compiled in MeSpEn.
- MSDManuals[6] English-Spanish corpus, compiled for this project under permission of the copyright holders.
- Portal Clínic[7] English-Spanish corpus, compiled by us for this project.
- The PubMed[8] corpus.
- The UFAL Medical Corpus[9] v1.0.

We also included in our training data glossaries and glossary-like databases containing terms and expressions frequently used in the medical domain. Namely, we used the English-Spanish glossary from MeSpEn, the 10th revision of the International Statistical Classification of ICD and SnowMedCT. With all the corpora and glossaries, we created an in-domain training corpus of 2,836,580 segments and entries. We split the corpus into two parts: 99% of the segments for training, and the remaining 1% for testing.

Finally, we added other general corpora for training the MT systems, namely the Scielo corpus, the Europarl corpus[10] (Koehn 2005), Global Voices corpus[11] and News Commentary. The IBECS, Scielo, Pubmed and a part of the MedlinePlus corpus were obtained from the MeSpEn corpus[12] (Villegas et al., 2018).

We evaluated the SMT and NMT systems using MTEval[13]. This software allows you to calculate BLEU, NIST, RIBES and WER using only one reference. We have used all the test sets of the corpus. We also compared our systems with Apertium[14] (Forcada et al.,

3 http://opus.nlpl.eu/EMEA.php

4 http://ibecs.isciii.es

5 https://medlineplus.gov/

6 https://www.msdmanuals.com/

7 https://portal.hospitalclinic.org/

8 https://www.ncbi.nlm.nih.gov/pubmed/

9 https://ufal.mff.cuni.cz/ufal_medical_corpus

10 http://www.statmt.org/europarl/

11 https://globalvoices.org/

12 http://temu.bsc.es/mespen/

13 https://github.com/odashi/mteval

14 http://www.apertium.org

2011) and Google Translate[15] to obtain an overall baseline of general-purpose MT engines to compare our systems against. As shown in the following table, the systems trained in the experiment obtain better results in all metrics than the reference systems used, except for the Google Translate system, which obtained a slightly better NIST result than the MMT Phrase-Based system without context and a better WER result than the two MMT Phrase-Based systems. The MMT Neural system performed consistently better than the MMT Phrase-Based system. In the MMT Neural system we did not see any significant difference between the results obtained when trained with or without context. Taking into account the results from the automatic evaluation, we used both SMT and NMT systems with context to produce the raw MT output.

MT systems	BLEU	NIST	RIBES	WER
Apertium	0.1926	6.4425	0.7131	0.7027
Google T.	0.4025	9.6323	0.8095	0.5300
MMT P.B. no context	0.4242	9.5362	0.8144	0.6378
MMT P.B. context	0.4448	9.8015	0.8193	0.6210
MMT Neural no context	0.5039	11.1062	0.8369	0.4855
MMT Neural context	0.5058	11.1413	0.8363	0.4810

Table 1: Results of the automatic evaluation using mteval.

3.2 PE set-up

We used the two previously trained MT systems to translate a 791-word fragment from a 2018 medical paper detailing a new oncological treatment. Four professional translators post-edited the 41 segments to produce a publishable-quality version. They all had between 5 and 10 years of professional experience and had worked between 3 and 6 years as post-editors in the medical domain. Two of them post-edited the SMT output and the other two the NMT output.

For the task, they used PET (Aziz et al., 2012), a computer-assisted translation tool that supports post-editing. It was used with its default settings. We used this tool because it also logs both temporal and technical post-editing effort. Post-editing effort results were included in a previous paper (Alvarez-Vidal et al., 2021).

15 https://translate.google.es/

Once they had finished, one of the authors of the paper with previous experience in marking MT errors manually annotated the four post-edited versions considering all modified elements as errors. We used the MQM (Lommel et al., 2014) taxonomy because it is a popular framework both in research and the translation industry. It mainly groups errors into fluency and accuracy: fluency relates to the quality of the target text and accuracy evaluates how the target text renders the meaning of the source text. Research has shown NMT produces more fluent translations than SMT, although accuracy may be sometimes compromised (Castillo et al., 2017).

We also included a different weight for every error according to its severity in line with MQM instructions. And following Klubička et al. (2018), we counted the number of words corresponding to each error. We also compared the two post-edited versions which had been annotated for each MT model to study the variation patterns between post-editors.

4. Error annotation

For our analysis, we used the Multidimensional Quality Metrics (MQM) (Lommel et al., 2014) system, which was developed as part of the QTLaunchPad project (funded by the European Union) to address the shortcomings of previous quality evaluation systems. This framework offers a flexible system for annotating errors and provides a list of error types that can be correlated to specific errors present in the MT output. It contains a total of 114 issue types, and it represents a generalized superset of the issues that can be found in current metrics and tools. Furthermore, it has become a popular framework both for the translation industry and research and, in fact, it was conceived as an update of the LISA QA Model, which was widely used in the localization industry.

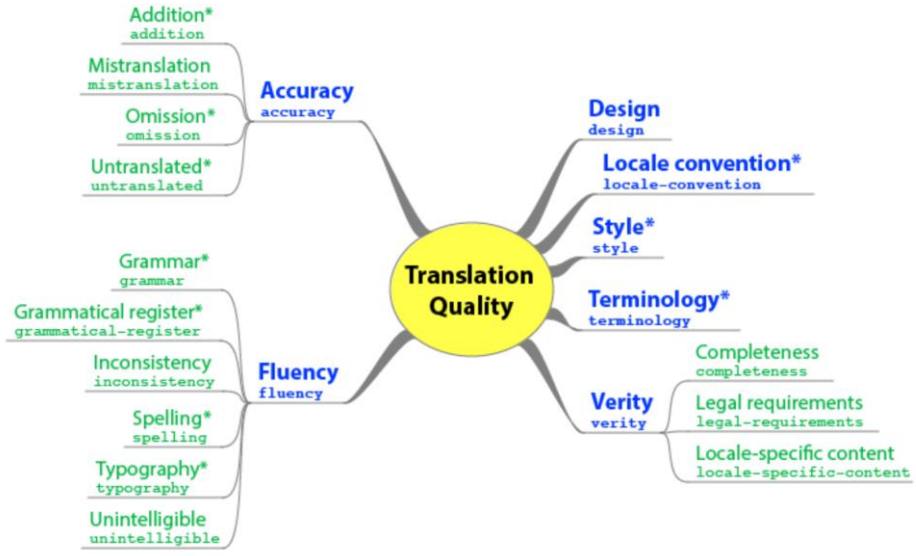

Figure 1. Graphical representation of the MQM core error categories (Lommel et al., 2015)

This framework offers the possibility of describing and defining custom translation quality metrics. Its goal is to provide a flexible vocabulary of quality issue types and a way to use these elements to generate quality scores. Instead of imposing a unique

metric for all situations, it provides a detailed catalogue of different quality issue types, including standardised names and definitions, that can be used to describe particular metrics for specific tasks. The hierarchical structure groups errors into different major issues (such as Fluency and Accuracy) which can be further specified into detailed error types. This enables different levels of granularity, from a coarse analysis to a fine-grained metric, and also facilitates the customisation of the framework for different language combinations. For example, if the analysis focuses on grammar errors, this category can be further specified to include a detailed error description for all the MT output issues encountered. It also includes a guide[16] for the annotators using the MQM framework, and a decision tree designed to standardize the categorization process.

For our analysis, we used four main categories: Terminology, Style, Accuracy, and Fluency. Terminology includes the specific terms related to the domain of the specialized text analysed, which in this case is medicine. Even though in some cases it can coincide with a mistranslation or an omission (which would be part of the Accuracy category), in this category we only included errors which were clearly related to terminological problems from the medical domain. Style groups all modifications introduced by the post-editor which can be considered unnecessary or stylistic. It includes all preferential choices of the different translators when post-editing. Some corrections cannot be considered errors but improve the fluency of the post-edited text and help to better understand the text. Accuracy groups errors which entail adding or removing some part of the source text information. These errors are usually the ones with the biggest impact on the MT output as they usually create critical problems in meaning.

Source: Sixty-nine **patients** had local recurrence and 17 **patients** showed [...].
MT output: Sesenta y nueve **pacientes** presentaron recaída local y 17 pacientes presentaron [...]
PE version: Sesenta y nueve **pacientes** (80%) presentaron recaída local y 17 presentaron [...]

Example 1. Sentence which includes the redundancy category

Finally, Fluency includes errors which have an impact on the quality of the target text, for example, all grammar mistakes produced by the MT system. We have further detailed this category to specify the corresponding type of errors. Apart from punctuation, capital letters and spelling, we have grouped errors mainly taking into account the grammatical category of the error detected. Furthermore, we have included word order (which also includes the modification of the syntactical order of the sentence) and what we have called redundancy. This category usually refers to references within the same sentence or the previous sentence which the MT system has repeated, but that should have been omitted or mentioned with another sort of reference. That is, taking into account the

16 http://www.qt21.eu/downloads/annotatorsGuidelines-2014-06-11.pdf

context, there is a redundant translation, which constitutes a grammatical problem in the Spanish output. For instance, in the following segment "pacientes" was removed the second time it appears, as in Spanish lexical repetitions should be avoided within the same sentence if possible.

5. Results

All post-edited versions were manually annotated using the customized MQM taxonomy by one of the authors. He had extensive experience as a translator for this language combination and had previous experience annotating MT errors for research and industrial purposes. Once the annotation process was completed, we calculated the number of corrections per each category and the mean for each MT system. As it can be seen in Table 2, there is a great divergence between the translators who post-edited the SMT output. In fact, PE1 introduced very few modifications. The results of the translators who post-edited the NMT version are more alike, although PE4 detected many more terminology errors. The mean of all results shows that fewer errors were corrected in the NMT output, although the difference is not statistically significant according to a two-way ANOVA analysis. The most relevant divergence in errors corresponds to accuracy errors, where NMT presented no untranslated elements from the source text and reduced in more than half the mistranslation. In the following sentences we can see examples of the untranslated elements in the SMT version compared with the NMT output:

Source: [...] and the overall **long-term** survival rate is
SMT: [...] y la supervivencia global es [...].
NMT: [...] y la supervivencia global **a largo plazo** es [...].

Example 2. The SMT suggestions present a higher number of omissions

Source: Study Population
SMT: Población.
NMT: Población del **estudio.**

Example 3. Another sentence which presents an omission in the SMT version

As it was a medical text, a considerable number of errors were produced by the use of the wrong terminology. This is in line with previous research, which has shown that in-domain MT outputs usually present a high number of terminology-related errors (Hawakaya and Arase, 2020). However, even though the two MT models were trained with the same data, translators corrected more terminology issues in the NMT version. If we remove from the total results the errors attributed to style, which in most cases correspond to an elective correction introduced into the MT output, results also show NMT output produced less errors (128 errors for SMT versus 119.5 for NMT).

When taking a closer look at the different versions produced by each translator, the total number of corrections in each category tends to be correlated. For the translators who introduce many style corrections, there is also an increase in accuracy and fluency errors. In the fluency category, the main divergence can be found in prepositions and word order. These corrections can be partly associated with the multiple correct possibilities offered by Spanish to translate a given source sentence.

Error Type	SMT PE1	SMT PE2	SMT MEAN	NMT PE 3	NMT PE4	NMT MEAN
Accuracy	**46**	**72**	**59**	**30**	**41**	**35.5**
Mistranslation	24	34	29	18	19	18.5
Omission	6	7	6.5	13	8	10.5
Addition	6	11	8.5	2	14	8
Untranslated	10	20	15			
Fluency	**34**	**74**	**54**	**52**	**55**	**53.5**
Punctuation	5	5	5	6	7	6.5
Verb	4	8	6	4	8	6
Word order	5	6	5.5	12	7	9.5
Prepositions	5	19	12	6	14	10
Capital letters	1	0	0.5	3	0	1.5
Concordance	6	10	8	5	2	3.5
Possessive	0	1	0.5	0	0	0
Articles	5	15	10	7	15	10.5
Spelling	0	0	0	1	1	1
Connectors	2	2	2	1	0	0.5
Pronouns	0	3	1.5	2	1	1.5
Redundancy	1	5	3	5	0	2.5
Style	**3**	**39**	**21**	**24**	**19**	**21.5**
Terminology	**11**	**19**	**15**	**41**	**20**	**30.5**
TOTALS	**94**	**204**	**149**	**147**	**135**	**141**

Table 2. Number of errors post-edited by each translator.

The highest divergence between the versions can be found in the terminology category. Corrections among the different translators are not consistent with the global number of errors. This could be related to the repetition of certain key terms. That is, if a translator decides to modify one term which is repeated throughout the document, all instances should be changed and the total number of corrections increases considerably.

We also included a measure for each of the errors annotated according to the severity of the error: neutral, minor, major, critical. We used the four categories included in MQM and the definitions suggested by O'Brien (2011):

- **Neutral**: Corresponds to stylistic corrections which do not really imply an error and it also includes corrections of issues, features and expressions that do not have a negative impact on the MT output.
- **Minor**: Noticeable errors that do not have a negative impact on meaning and are not confusing or misleading.
- **Major**: Errors that are considered to have a negative impact on meaning.
- **Critical**: Errors which have major effects on the overall meaning, and can compromise product usability, and consumer safety and health.

MT system and post-editor	Neutral	Minor	Major	Critical
SMT PE1	10	42	31	11
SMT PE2	34	105	51	13
NMT PE3	22	87	33	5
NMT PE4	19	70	40	6

Table 3. Severity of the annotated errors post-edited by each translator.

As we can see in Table 3, critical errors were clearly reduced in the two NMT post-edited versions, which seems to indicate that NMT was able to convey better the meaning of the source text. These results can be directly linked to the accuracy errors detected in both systems, in which NMT showed a better performance in reproducing the whole meaning of the source segment into the target.

Finally, we counted the number of words corresponding to each error corrected to calculate the error ratio (Klubička et al., 2018). For each version we divided the number of words that contain an error by the total number of words included in the final post-edited version:

Error ratio = Words with errors / Total number of words

As we can see in Table 4, the percentage of errors is consistent with the global number of errors annotated in each post-edited version. Even though there is a big variability among the SMT versions, the mean of the corrections introduced by the two post-editors (25.6%) is slightly higher than the mean corresponding to the translators who post-edited the NMT output (23.1%).

MT system and post-editor	Error ratio
SMT PE1	16.9%
SMT PE2	34.3%
NMT PE3	25.8%
NMT PE4	20.4%

Table 4. Error ratio calculated for each translator.

6. Discussion and future work

PE is a practice that will increase in the near future, and it is necessary to assess the MT output and understand translators' corrections in order to ensure a satisfying post-editing process and also a final translation of good quality. Error analysis will be a useful tool to achieve it. It can help detect the most frequent errors of each MT system and help prevent the repetitive errors which can be more tedious for post-editors. In our analysis for an English to Spanish medical text, the NMT slightly reduced the number of errors, especially the ones related to omissions or mistranslations from the source text. This fact is reflected in the greater number of critical errors for the SMT version. Even though NMT is usually found to be more fluent than SMT, for this language combination and domain the mean of fluency errors was more or less the same, as was the number of style corrections. NMT conveys the source meaning better but still has problems producing publishable-quality documents for the medical domain.

There was also a great variability among translators. Even though we had only two post-edited versions for each MT output, post-edited versions with higher number of corrections tend to increase in the accuracy, fluency and style categories alike. The terminology category seems to increase independently from the other three. If we focus on the fluency category, the highest divergences can be found in word order and prepositions.

Our future experiments will include increasing the pool of post-editors for a certain text to study with more detail variability among translators and correlate specific error categories with an increased PE effort. We will also broaden the domains and language

combinations for error annotation in order to obtain a larger corpus of post-edited documents.

References

Allen, J. H. (2003). Post-editing. In: Sommer, H. (ed.). *Computers and Translation: A translator's guide*. Amsterdam: John Benjamin. (Benjamins translation library; 35), pp. 297-317.

Alvarez-Vidal, S.; Oliver, A.; Badia, T. (2021). Comparing NMT and PBSMT for Post-editing In-domain Formal Texts: A Case Study. In: Tra&Co Group (ed.). *Translation, interpreting, cognition: The way out of the box*. Berlin: Language Science, pp. 33-47. <https://doi.org/10.5281/zenodo.4544686>. [Accessed: 20211117].

Aranberri, N. (2014). Postediting, productivity and quality. *Tradumàtica, Tecnologies de la traducció,* n. 2012, pp. 471-477. <https://doi.org/10.5565/rev/tradumatica.62>. [Accessed: 20211117].

Aziz, W.; Castilho, S.; Specia, L. (2012). "PET: A Tool for Post-editing and Assessing Machine Translation." In: *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12).* European Language Resources Association (ELRA), pp. 3982–3987. <http://www.lrec-conf.org/proceedings/lrec2012/pdf/985_Paper.pdf>. [Accessed: 20211117].

Barrault, L.; Bojar, O.; Costa-Jussà, M.; Federmann, C.; Fishel, M.; Graham, Y.; Haddow, B.; Huck, M.; Koehn, P.; Malmasi, S.; Monz, C.; Müller, M.; Pal, S.; Post, M.; Zampieri, M. (2019). Findings of the 2019 Conference on Machine Translation (WMT19). In: *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*. Association for Computational Linguistics, pp. 1-61. <https://doi.org/10.18653/v1/W19-5301>. [Accessed: 20211117].

Bentivogli, L.; Bisazza, A.; Cettolo, M.; Federico, M. (2016). Neural versus Phrase-Based Machine Translation Quality: a Case Study. In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pp. 257-267. <https://doi.org/10.18653/v1/D16-1025>. [Accessed: 20211117].

Bojar, O.; Federmann, C.; Fishel, M.; Graham, Y.; Haddow, B.; Huck, M.; Koehn, P.; Monz, C. (2018). *Findings of the 2018 Conference on Machine Translation (WMT18)*. In: *Proceedings of the Third Conference on Machine Translation Shared Task Papers.* Association for Computational Linguistics, pp. 272-303. <https://doi.org/10.18653/v1/W18-6401>. [Accessed: 20211117].

Castilho, S.; Moorkens, J.; Gaspari, F.; Sennrich, R.; Sosoni, V.; Georgakopoulou, Y.; Lohar, P.; Way, A.; Miceli Barone, A.; Gialama, M. (2017). A comparative quality evaluation of PBSMT and NMT using professional translators. In: *Machine Translation Summit XVI: Proceedings of MT Summit XVI, vol.1: Research Track.* Kyoto: Sadao Kurohashi; Hong Kong: Pascale Fung, pp. 116–131. <http://aamt.info/app-def/S-102/mtsummit/2017/wp-content/uploads/sites/2/2017/09/MTSummitXVI_ResearchTrack.pdf>. [Accessed: 20211117].

Costa, A.; Ling, W.; Luis, T.; Correia, R.; Coheur, L. (2015) A linguistically motivated taxonomy for Machine Translation error analysis. *Machine Translation,* v. 29, n. 2, pp.127–161. <https://doi.org/10.1007/s10590-015-9169-0>. [Accessed: 20211117].

Daems, J.; Vandepitte, S.; Hartsuiker, R. J.; Macken, L. (2017). Identifying the Machine Translation Error Types with the Greatest Impact on Post-editing Effort. *Frontiers in Psychology*, n. 8. <https://doi.org/10.3389/fpsyg.2017.01282>. [Accessed: 20211117].

De Almeida, G. (2013). *Translating the post-editor: An investigation of post-editing changes and correlations with professional experience* [PhD Thesis]. Dublin City University, Dublin. <http://doras.dcu.ie/17732/>. [Accessed: 20211117].

Denkowski, M.; Lavie, L. (2012). Challenges in predicting machine translation utility for human post-editors. In: *Proceedings of AMTA 2012*. <https://doi.org/10.1184/r1/6473105>. [Accessed: 20211117].

Farrús, M.; Costa-Jussà, M. R.; Mariño, J. B.; Fonollosa, J. A. R. (2010). Linguistic-based Evaluation Criteria to Identify Statistical Machine Translation Errors. In: *Proceeding of the 14th Annual Conference of the European Association for Machine Translation (EAMT 2010)*, Saint-Raphal, France, pp. 167–173. <https://repositori.upf.edu/bitstream/handle/10230/34496/Farrus_EAMT2010_ling.pdf?sequence=1&isAllowed=y>. [Accessed: 20211117].

Federico, M.; Negri, M.; Bentivogli, L.; Turchi, M. (2014). Assessing the Impact of Translation Errors on Machine Translation Quality with Mixed-effects Models. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (EMNLP 2014), Doha, Qatar, pp. 1643–1653. <https://doi.org/10.3115/v1/D14-1172>. [Accessed: 20211117].

Forcada, M. L.; Ginestí-Rosell, M.; Nordfalk, J.; O'Regan, J.; Ortiz-Rojas, S.; Pérez-Ortiz, J. A.; Sánchez-Martínez, F.; Ramírez-Sánchez, G.; Tyers, F. M. (2011). Apertium: a free/open-source platform for rule-based machine translation. *Machine translation,* v. 25, n. 2, pp. 127–144. <https://doi.org/10.1007/s10590-011-9090-0>. [Accessed: 20211117].

Germann, U.; Barbu, E.; Bentivogli, L.; Bertoldi, N.; Bogoychev, N.; Buck, C.; Caroselli, D.; Carvalho, L.; Cattelan, A.; Cattoni, R.; *et al.* (2016). Modern MT: A New Open-source Machine Translation Platform for the Translation Industry. *Baltic Journal of Modern Computing*, vol. 4, no. 2. <http://www.bjmc.lu.lv/fileadmin/user_upload/lu_portal/projekti/bjmc/Contents/4_2_28_Products.pdf>. [Accessed: 20211117].

Guerberof, A. (2009). Productivity and quality in MT post-editing. In: *Proceedings of MT Summit XII*, pp. 8-13. <https://www.researchgate.net/publication/320467106_Productivity_and_quality_in_MT_post-editing>. [Accessed: 20211117].

Hawakaya, T.; Arase, Y. (2020). Fine-Grained Error Analysis on English-to-Japanese Machine Translation in the Medical Domain. In: *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, Lisboa, Portugal. European Association for Machine Translation, pp. 155-164. <https://www.aclweb.org/anthology/2020.eamt-1.17.pdf>. [Accessed: 20211117].

Junczys-Dowmunt, M.; Dwojak, T.; Hoang, H. (2016). Is Neural Machine Translation Ready for Deployment? A Case Study on 30 Translation Directions. In: *Proceedings of the Ninth International Workshop on Spoken Language Translation (IWSLT).* <https://arxiv.org/abs/1610.01108>. [Accessed: 20211117].

Klubička, F.; Toral, A.; Sánchez-Cartagena, V. M. (2017). Fine-Grained Human Evaluation of Neural Versus Phrase-Based Machine Translation. *The Prague Bulletin of*

Mathematical Linguistics, n. 108, pp. 121–132. <https://doi.org/10.1515/pralin-2017-0014>. [Accessed: 20211117].

Klubička, F.; Toral, A.; Sánchez-Cartagena, V. M. (2018). Quantitative fine-grained human evaluation of machine translation systems: a case study on English to Croatian. *Machine Translation,* n. 32, pp. 195–215. <https://doi.org/10.1007/s10590-018-9214-x>. [Accessed: 20211117].

Koehn, P. (2005). Europarl: a parallel corpus for statistical machine translation. In: *Proceedings of the MT Summit*, vol. 5, pp. 79–86. <https://www.statmt.org/europarl/>. [Accessed: 20211117].

Koponen, M. (2013). This translation is not too bad: An analysis of post-editor choices in a machine translation post-editing task. In: O'Brien, S.; Simard, M.; Specia, L. (eds.). *Workshop Proceeding: Workshop on Post-editing Technology and Practice (WPTP-2).* Allschwil: The European Association for Machine Translation, pp. 1-9. <https://www.researchgate.net/publication/299347281_This_translation_is_not_too_bad_An_analysis_of_post-editor_choices_in_a_machine_translation_post-editing_task> [Accessed: 20211117].

Koponen, M.; Leena, S. (2017). Post-editing quality: analyzing the correctness and necessity of post-editor corrections. *Linguist Antverp, New Series Themes in Translation Studies,* v. 16, pp. 137–148. <https://lans-tts.uantwerpen.be/index.php/LANS-TTS/article/view/439>. [Accessed: 20211117].

Koponen, M.; Leena, S.; Nikulin, M. (2019). A product and process analysis of post-editor corrections on neural, statistical and rule-based machine translation output. *Machine Translation,* v. 33, pp. 61–90. <https://doi.org/10.1007/s10590-019-09228-7>. [Accessed: 20211117].

Krings, H. P. (2001). *Repairing texts: Empirical investigations of machine translation post-editing process*. Kent, OH: The Kent State University Press. (Translation studies; 5).

Lommel, A.; Burchardt, A.; Görög, A.; Uszkoreit, H.; Melby, A. K. (2015). *Multidimensional Quality Metrics (MQM) Issue Types.* <http://www.qt21.eu/mqm-definition/issues-list-2015-12-30.html>. [Accessed: 20211117].

Lommel, A.; Uszkoreit, H.; Burchardt, A. (2014). Multidimensional Quality Metrics (MQM): A Framework for Declaring and Describing Translation Quality Metrics. *Tradumàtica, Tecnologies de la Traducció*, n. 12, pp. 455-463. <https://doi.org/10.5565/rev/tradumatica.77>. [Accessed: 20211117].

Lommel, A. R.; DePalma, D. A. (2016). Europe's Leading Role in Machine Translation: How Europe Is Driving the Shift to MT: Technical report. <http://cracker-project.eu/csa-mt-report/>. [Accessed: 20211117].

O'Brien, S. (2012). Towards a Dynamic Quality Evaluation Model for Translation. *JosTrans, The Journal of Specialised Translation,* n. 17, pp. 55–77. <https://jostrans.org/archive.php?display=17>. [Accessed: 20211117].

Papineni, K.; Roukos, S.; Ward, T.; Zhu, W. J. (2002). BLEU: A Method for Automatic Evaluation of Machine Translation. In: *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, pp. 311–318. <https://doi.org/10.3115/1073083.1073135>. [Accessed: 20211117].

Popovic, M. (2011). Hjerson: An Open Source Tool for Automatic Error Classification of Machine Translation Output. *The Prague Bulletin of Mathematical Linguistics,* v. 96 (october), pp. 59-68. <https://ufal.mff.cuni.cz/pbml/96>. [Accessed: 20211117].

Popović, M.; Lommel, A.; Burchardt, A., Avramidis, E.; Uszkoreit, H. (2014). Relations between different types of post-editing operations, cognitive effort and temporal effort. In: Proceedings of the 17th Annual Conference of the European Association for Machine Translation. Allschwil: The European Association for Machine Translation, pp. 191-198. <https://www.aclweb.org/anthology/2014.eamt-1.41.pdf> [Accessed: 20211117].

Popovic, M.; Arcan, M. (2016). PE2rr Corpus: Manual Error Annotation of Automatically Pre-annotated MT Post-edits. *LREC,* pp. 27-32. <https://www.aclweb.org/anthology/L16-1005.pdf>. [Accessed: 20211117].

Popović, M. (2018). Error Classification and Analysis for Machine Translation Quality Assessment. In: Moorkens J.; Castilho S.; Gaspari F.; Doherty S. (eds.). *Translation Quality Assessment. Machine Translation: Technologies and Applications*, vol 1. Cham: Springer. <https://doi.org/10.1007/978-3-319-91241-7_7>. [Accessed: 20211117].

Shterionov, D.; Superbo, R.; Nagle, P.; Casanellas, L.; O'Dowd, T.; Way, A. (2018). Human versus Automatic Quality Evaluation of NMT and PBSMT. *Machine Translation*, v. 32, n. 3, pp. 217–235. <https://doi.org/10.1007/s10590-018-9220-z>. [Accessed: 20211117].

Toral, A.; Sánchez-Cartagena, V. M. (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: Vol. 1, Long Papers* (Valencia), pp. 1063–1073. <https://www.aclweb.org/anthology/E17-1100.pdf>. [Accessed: 20211117].

Vilar, D.; Xu, J.; D'Haro, L. F.; Ney, H. (2006). Error Analysis of Machine Translation Output. In: *Proceedings of the Fifth International Conference on Language Resources and Evaluation*, Genoa, Italy 2006, pp. 697–702. <http://www.lrec-conf.org/proceedings/lrec2006/pdf/413_pdf.pdf>. [Accessed: 20211117].

Villegas, M.; Intxaurrondo, A.; Gonzalez-Agirre, A.; Marimon, M.; Krallinger, M. (2018). The MeSpEN Resource for English-Spanish Medical Machine Translation and Terminologies: Census of Parallel Corpora, Glossaries and Term Translations. In: *LREC MultilingualBIO: Multilingual Biomedical Text Processing*. ELRA. <http://lrec-conf.org/workshops/lrec2018/W3/pdf/8_W3.pdf>. [Accessed: 20211117].

Wu, Y.; Schuster, M.; Chen, Z.; Le, Q. L.; Norouziet, M.; *et al.* (2016). Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. <https://arxiv.org/abs/1609.08144>. [Accessed: 20211117].

Ye, Y.; Toral, A. (2020). Fine-grained Human Evaluation of Transformer and Recurrent Approaches to Neural Machine Translation for English-to-Chinese. In: *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, Lisboa, Portugal. European Association for Machine Translation, pp. 125-134. <https://eamt2020.inesc-id.pt/proceedings-eamt2020.pdf>. [Accessed: 20211117].

Zeman, D.; Fishel, M.; Berka, J.; Bojar, O. (2011). Addicter: What is wrong with my translations? *The Prague Bulletin of Mathematical Linguistics,* v. 96 (october), pp. 79-88. <https://doi.org/10.2478/v10108-011-0013-2>. [Accessed: 20211117].