Francesca D'Angelo*

Assessing translation quality: a survey of research into human translation, post-editing and machine translation

Abstract
L'articolo analizza il rapporto tra traduzione umana, post-editing e traduzione automatica affrontando, in particolare, la discussa questione della qualità traduttiva. Vengono presentati diversi approcci e metodologie, identificando una serie di problematiche riguardanti la traduzione in termini di processo, prodotto e sistemi di valutazione. È inclusa, inoltre, una riflessione sull'osservata tendenza all'oggettivazione degli standard di traduzione e all'utilizzo di metriche automatiche, con particolare attenzione alle possibili implicazioni che ciò potrebbe avere per la comunità scientifica e di traduzione.

Parole chiave: Traduzione automatica; Post-Editing; Traduzione umana; Qualità traduttiva

The article investigates the relationship between human translation, post-editing and machine translation addressing, in particular, the controversial issue of translation quality. An overview of different methodologies and approaches is provided, identifying a number of perceived issues concerning translation in terms of process, product, and evaluation systems. A discussion on the observed tendency to objectify translation standards, and to automatic metrics, is included with a focus on the implications it may have for the research and translation community.

Keywords: Automatic Translation; Post-Editing; Human Translation; Translation Quality

1 Introduction

The considerable growth of interest in the translation field in recent years, due to the pressure to increase the productivity of translation in terms of both amount of text and processing time, has led researchers from multiple fields of study including linguistics, computer sciences and artificial intelligence to investigate the role of machine translation (MT). On one hand, automat-

* Francesca D'Angelo, Alma Mater Studiorum Università di Bologna, francesca.dangelo16@unibo.it.

ic translation is seen as a challenging opportunity to foster and support the translation process, in the form of post-editing, increasing the speed and productivity of translators. On the other, the increasing spread of MT software, to respond to the significant growth of linguistic content to translate, contributes to questioning the translation output in terms of expected quality level. Starting with a technical overview of the most influential approaches to MT, from Statistical Machine Translation (SMT) to the advent of Neural Machine Translation (NMT) systems, the paper aims to provide a historical background of automatic translation and post-editing. In particular, the change of perspective towards post-editing, presented from a diachronic perspective from its introduction in the late 50s to the latest implementations in different fields, shows how the changing demand has brought an overall amelioration of the process. Besides, the article sheds light on the relationship between human translation and MT addressing the controversial issue of translation quality, human parity and superhuman performance, highlighting the complexity of the concept, its different categorisations as well as the different approaches available, from human to automatic quality evaluation systems. Finally, a debate on the increasing tendency to objectify translation quality, through the development of indicators and standards, is included with a focus on its impact in terms of needs and expectations concerning the quality assessment of MT and post-edited texts. In particular, the paper addresses the most recent developments, since 2013, regarding the role of post-editing effort in assessing the quality of neural machine-translated texts.

2 Approaches to Machine Translation: a Critical Overview

MT is a recently developed subfield of computational linguistics that inquires how computer software can develop efficient systems that can translate between human languages (Oladosu et al. 2016). Before starting the discussion on the productivity and quality of MT output, a technical overview of the most important approaches to MT is necessary to better understand the effectiveness of automatic translation and how it can support human translation. A first distinction can be made in MT between single and hybrid approaches to investigate their mechanism, affordances and constraints. Afterwards, NMT is presented as a new approach in the field of automatic machine translation. It has

gained great popularity among researchers in this field because of the promising translation results achieved, in addition to the simplicity of its structure.

Single approaches to MT encompass rule-based, direct-based, corpus-based, and knowledge-based methods. Rule-Based MT employs morphological, syntactic, and semantic rules to address word order issues. Direct-Based MT relies on bilingual dictionaries for word substitution, followed by syntax rearrangement, suitable for unidirectional translation. Corpus-based MT includes example-based and statistical methods, with the latter utilising machine learning on parallel corpora for translating new sentences. Finally, knowledge-based MT relies on extensive semantic and pragmatic knowledge for translation choices.

In contrast, the hybrid approach combines multiple MT techniques, typically integrating statistical and rule-based approaches. This hybrid approach includes word-based, phrase-based, syntax-based, and forest-based models. Word-based models focus on lexical word dependencies but struggle with long-distance reordering. Phrase-based models, on the other hand, use phrases as translation units, allowing local reordering and handling idiomatic expressions. Finally, syntax-based models analyse hierarchical sentence structures, offering string-based and tree-based parsing options. However, it should be noted that tree-based systems may suffer parsing errors due to relying on a single best parse tree. To overcome these issues, forest-based translation was introduced as a hybrid approach, blending string and tree-based methods to enhance MT efficiency and reduce parsing errors. It is a combination that facilitates faster decoding, making it a valuable addition to the MT landscape. In summary, the choice between single and hybrid MT approaches depends on the specific requirements of a translation task, with hybrid approaches offering advantages in terms of flexibility and quality improvement through the integration of multiple techniques.

These translation systems, however, still present significant flaws that more recent translation machine translation systems have tried to overcome. Indeed, past MT systems were mainly rule-based systems with the aim of creating grammatical rules for the source and target language. Indeed, MT acted as a translation between languages based on this set of rules. The problem was mainly the addition of new content and new language pairs

since maintaining and extending such a set of rules was too time-consuming and costly. Hence, SMT was created to respond to these translation problems (Koehn 2010). SMT systems create statistical models by analysing an aligned set of source and target language sentences (i.e. training set). On one hand, the advantage of SMT concerns its automatic learning process and the relatively easy adaptation. The disadvantage, on the other hand, refers to the training itself as it is necessary to create a usable tool and a large database of source and target language segments. Another critical aspect of SMT arises when dealing with grammatically more complicated languages.

In detail, the NMT has recently started to be promoted to solve these technical issues associated with SMT systems. To give some indication of the speed of change, in 2015, only one neural machine translation system was submitted at the shared task for machine translation organized by the Conference on Machine Translation (WMT). In 2017, almost all submitted machine translation systems were neural. The system looks at the sentence as a whole and can form associations between phrases even at greater distances in the sentence. In particular, at the IWSLT 2015 evaluation expedition, NMT was able to overcome state-of-the-art phrase-based machine translation systems on English-German, the language pair famous for its difficulty due to morphology and grammatical differences (Bentivogli et al. 2018).

The SMT system consists of several components tuned separately, whereas the NMT model is a large end-to-end single network that consists of two sub-recurrent neural networks: the encoder and the decoder. Moreover, if the SMT system needs many features that are accurately defined to do the translation, the NMT model depends on a training corpus to learn the translation task, with less or no feature engineering effort by linguists. Another major advancement that is worth discussing concerns the ability of NMT to seize potential long-distance dependencies and complicated word alignment information. In addition, the NMT model does not require a large memory space, such as those used by the SMT to store a translation model, a reordering model and a language model.

Nonetheless, it is important to notice that although convolutional and sequence-to-sequence provided good translation accuracy, the latter was reduced as the length of the input sentence increased. These models have

adopted an encoder–decoder approach that compresses all the necessary information of a source sentence into a fixed-length vector, which made it difficult for the models to handle long sentences, especially those longer than the training sentences. This problem has been solved with the introduction of the attention mechanism, which has achieved great popularity and has been used in various fields. In MT, the three architectures used are Stacked RNN with Attention, Self-attentional Transformer, and Fully Convolutional Models (ConvSeq2Seq). In 2015, to deal with the problem of fixed-length vectors, Bahdanau et al. (2014) proposed a model that extends the encoder-decoder approach by allowing automatic search for portions of a source sentence, which have relevance to the prediction of a target word, without explicitly forming these portions as a hard segment. Instead of encoding a whole input sentence into a vector of fixed length, the model converts it into a sequence of vectors. Each time during the decoding process, the decoder searches the input sentence for the words that have the most relevant information to generate the target word. The target word is predicted based on a context vector of all relevant words, and all previously predicted target words.

3 Background of Machine Translation and Post-Editing Research

The changing landscape in the translation industry, due to the recorded technological advances, has raised important questions on the relationship between human translators and MT. In particular, a change of perspective towards MT output has been acknowledged in the last decades. From its advent, an overall scepticism could be observed towards machine translated works and their uses in industries were peripheral and limited. Nowadays, thanks to the availability of free online MT systems, e.g. *DeepL, Bing Microsoft Translation, Reverso, Smartling* etc., translation output has started to be used as a rough version to be post-edited by human translators. More precisely, "post-editing is the correction of raw machine translated output by a human translator according to specific guidelines and quality criteria" (O'Brien 2001: 1). The advances recorded in the MT field, have brought the translation industry to increase the demand for post-editing instead of translating from scratch or revising human translations. If on one hand, the interest in the field is relatively recent, post-editing is one of the earliest uses envisioned for MT systems.

The history of post-editing can be dated back to the late '50s and early '60s when it was considered a "surprisingly hot topic" (Garcìa 2012: 293). One of the first officially recorded uses of post-editing in MT systems refers to the translation of scientific texts from Russian to English at the RAND Corporation (see Edmundson and Hays 1958). This approach entailed a good command of English on behalf of the post editor but not necessarily knowledge of the source language. Indeed, he/she would work on the MT text supported by a grammar code indicating the morphological information of the case, number, part of speech etc. In the 60s, post-editing was employed by the US Air Force's Foreign Technology Division and Euratom. However, after a negative report by the Automatic Language Processing Committee in 1966 pointing out that it was not worth the effort in terms of quality and productivity compared to human translation, post-editing systems stopped receiving funds (Koponen 2016).

Despite not living up to the expected performance in terms of time, quality and productivity, MT and post-editing continued to be developed and refined. Starting from the 70s, post-editing processes were implemented by important organisations such as the EU and the Pan-American Health Organisation. However, balancing between the advantages of post-editing in terms of speed and productivity of translators on one hand and the translation quality is still a controversial issue in translation research. An influential work by Gaspari et al. (2015) provides an interesting overview of the perception of MT texts and the use and effectiveness of post-editing in the translation and localisation field. The large-scale survey of MT competencies, based on data derived from 438 validated respondents including freelance translators, language service providers, translator trainers and academics, reveals interesting information about the needs and expectations of the community of translators.

The study highlights the increasing use of MT and post-editing, the increased demand for this service compared to the past and infers that this demand is expected to rise in the future. In more detail, the authors found that MT was used by 30% of the respondents, while 21% considered it useful and declared to be likely to use it in the future. Interestingly, 38% of the participants who declared to use MT pointed out that the text was always post-edited, 32% of them never performed post-editing and the remaining 32% of the interviewees used post-editing discontinuously. However, an important

observation to clarify the data referred to the overall level of satisfaction is needed. Most participants who declared to use MT did not use any customisation tool before performing the translation. Since half of the respondents show a low level of satisfaction, the authors analysed the correlation between the two factors finding that those who did not customise MT, were not satisfied with the translation. This explains how fundamental customisation is to meet the translation needs more specifically and to ameliorate the quality of the translation output. As Gaspari et al. (2015) suggest, it is possible to improve customisation by ameliorating the translators' assets and technological competencies such as customised glossaries in linguistic pre-processing of the text. Besides, the respondents who declared to perform post-editing in MT resorted to human evaluation as the most common approach to quality assessment, another controversial issue that is worth discussing in more detail.

An interesting study by Cettolo et al. (2013) provides valuable insights into assessing the quality of neural machine-translated texts, particularly focusing on post-editing efforts. Indeed, in 2013, a novel approach was introduced to evaluate machine translation output. This evaluation took place during the 10th IWSLT evaluation campaign, centred on transcribing and translating lectures using the TED Talks corpus. The assessment encompassed various language pairs, including English-German-French, with optional tracks for 12 languages. Eighteen teams participated, submitting 217 runs, which were evaluated using objective metrics and compared to previous systems. Instead of traditional subjective rankings, the study investigated the post-editing effort required by professional translators to improve machine-generated translations. This led to the adoption of the Human-mediated Translation Edit Rate (HTER) metric, which measures the minimal edit distance between the machine-generated translation and its manually revised version. HTER demonstrated a strong correlation with human evaluations of translation quality. Overall, the post-editing task offered a dual advantage: it highlighted specific translation errors and provided additional reference translations, enhancing the assessment of MT systems. Notably, the study revealed that the most proficient system required minimal post-editing effort, underscoring the potential of machine translation to assist human translators, with an optimal HTER score threshold of 19%.

As already argued, in 2015, a pivotal moment in MT occurred when a NMT system, as detailed by Luong and Manning (2015), surpassed Phrase-Based Machine Translation (PBMT) systems in the IWSLT competition. This marked a substantial improvement in translation quality, especially for complex language pairs like English-German, heralding the onset of the NMT era, following a period when NMT was computationally and resource-intensive compared to PBMT. For the purpose of the current discussion, it is worth recalling a work by Bentivogli et al. (2018) which sheds light on the role of post-editing in MT quality assessment, taking into account the developments of NMT quality. Particularly, the study underscores NMT's substantial advancements, showcasing its superior translation quality and post-editing efficiency across challenging language pairs and diverse sentence lengths. More specifically, the research compared NMT and PBMT outputs by analysing high-quality post-edits performed by professional translators on IWSLT data. This approach, unlike conventional MT evaluation reliant on arbitrary reference translations, enabled a comprehensive evaluation of systems, incorporating post-editing effort and error types. It also holds practical relevance for integrating MT into Computer-Assisted Translation (CAT) tools, where post-editing is common. The key findings highlighting NMT's superiority over PBMT can be summarised in the following points:

- NMT notably reduced overall post-editing effort.
- NMT consistently outperformed PBMT across various sentence lengths.
- NMT generated a higher proportion of low-error MT outputs, crucial for CAT tools.
- NMT exhibited significantly fewer errors, with lower error rates.
- NMT produced fewer lexical, morphology, and word order errors compared to PBMT.

4 Translation Quality: from Human Evaluation to Automatic Metrics

What clearly emerges from the current discussion about post-editing is that the task significantly differs from the traditional process of translation and revision. Considering how common this practice has become, organisations implementing MT are now trying to balance cost and productivity to assess translation. Nonetheless, getting to commonly shared metrics to as-

sess translation quality is not an easy task since the definition of translation quality may significantly vary depending on the factor under investigation.

Indeed, studies addressing the quality of MT texts have been concerned about the supposed inferior level of MT texts compared with the quality level of manually translated texts. A study by Fiederer and O'Brien (2009) compared a set of sentences translated manually or by post-editing to address this question. The different sentences were rated according to three criteria: i.e. clarity (how understandable the sentence was); accuracy (how close the target text meaning was to the source text); and style (naturalness and appropriateness). The findings indicate that post-edited translations were rated higher in terms of clarity and accuracy whereas, in terms of style, the version of manually translated texts was preferred. On the other hand, Carl et al. (2011) asked evaluators to rank in order of preference the manually translated texts and the post-edited versions. A slightly higher, although not significant, preference was recorded for the post-edited translations. However, to better interpret these findings, it would be worth including other variables in the analysis such as the proficiency of the evaluators in both languages, experience in the translation field etc. Other studies assessing the quality of MT texts have focused on the number of errors found in the translation. For instance, Plitt and Masselot (2010), assessed manually and post-edited versions of translated texts according to the criteria employed by the company's quality assurance team. Interestingly, although both versions were considered acceptable for publication, the evaluators considered the manually translated texts as needing more corrections.

On the whole, the literature on MT and human translation evaluation shows that post-editing can lead to quality levels close to manually translated texts. Depending on the quality criteria employed, post-editing texts are, in fact, sometimes even preferred to the manual versions. Nonetheless, all these studies discussed and reviewed so far involve human evaluators to rate in order of preference or choose between manually translated and MT versions. Currently, the trend has moved from human evaluation towards automatic tools since they are less time and cost-consuming. It should be noticed, indeed, that human judgement is mainly based on two main criteria, i.e. adequacy and fluency, and it is rather subjective. Hence, automatic eval-

uation measures are sought. The most common and better developed nowadays are BLEU (Bilingual Evaluation Understudy); WER (Word Order Rate); PER (Position-Independent Word Error Rate); and NIST (National Institute of Standards and Technology). Each automatic evaluator is developed on certain standards and has its affordances and constraints.

Developed by Papineni et al. (2002), BLEU addresses the evaluation problem by comparing the system output with a reference translation of the same text. The validity of BLEU has been proved through correlations with human evaluations. WER is developed by computing the number of substitutions, insertion and deletion operations performed to convert the generated translation into the reference translation. Where several reference translations are provided as source text, the evaluator calculates the minimal distance to this set of references (Nieben et al. 2000). One of the main disadvantages of WER is that it requires perfect word order. To solve this issue, the position-independent word was introduced (i.e. PER). NIST, on the other hand, is based on the BLEU metrics but with some differences. It calculates the-gram precision the same way as BLEU but, at the same time, it also calculates how informative a particular n-gram is. As O'Brien (2011: 3) points out, the limitations of these automatic metrics are well acknowledged by the translation community. Indeed, automatic metrics are not supposed to properly predict the usefulness, adequacy, and reliability of MT technologies. In addition, it can be argued that if on one hand the usefulness of automatic evaluation metrics is deemed, on the other, they believe that too much importance has been given to them "since real translation quality is what we should be concerned with" (O' Brien 2011: 3).

Besides, on the expected level of translation quality, it is important to notice that the discussions surrounding the achievements of 'human parity' and 'super-human performance' in the domain of Neural Machine Translation, as asserted respectively by Hassan et al. (2018) and Bojar et al. (2018), have engendered substantial scholarly scrutiny. These assertions have catalysed an extensive examination of the appropriateness of the evaluation metrics employed in the assessment of NMT systems. Presently, the landscape of neural methodologies in machine translation presents novel challenges, with NMT outputs exhibiting a remarkable fluency. It has become evident that traditional automated metrics may inadequately capture the nuanced quali-

ty of neural systems, which are distinguished by their elevated fluency. This recognition underscores the imperative for the development and adoption of more nuanced and context-sensitive evaluation approaches, as witnessed by different researchers (e.g. Belouadi & Egere 2022; Mathur et al. 2020; Marie et al. 2021), to comprehensively assess the capabilities and limitations of these advanced neural machine translation systems. The aforementioned terminology such as "human parity" (Hassan et al. 2018) and even "super-human performance" (Bojar et al. 2018; Barrault et al. 2019) has been employed, particularly within the context of the Workshop on Machine Translation (WMT) evaluation campaigns. This suggests that MT systems are presumed to have reached a level of quality equal to, or possibly surpassing, the level of human translation, at least in the specific evaluation framework employed. In practical evaluations, machine-generated translations were consistently preferred over those produced by professional human translators. While these achievements appear impressive on the surface, it is imperative to engage in a more extensive and nuanced discussion, placing these claims into a broader context for comprehensive analysis.

A remarkable study by Hassan et al. (2018) investigates the challenge of defining and accurately evaluating human parity in translation. The authors adopt the following definition: "If there is no statistically significant difference between human quality scores for a test set of candidate translations from a machine translation system and the scores for the corresponding human translations then the machine has achieved human parity" (Hassan et al. 2018: 2). The paper provides an overview of Microsoft's machine translation system and assesses the translation quality on the well-established WMT 2017 news translation task, specifically from Chinese to English. The findings indicate that Microsoft's latest neural machine translation system has achieved a new state-of-the-art performance, and its translation quality is on par with that of professional human translations. Moreover, it significantly surpasses the quality of translations produced by non-professional crowd-sourced sources.

The same study has been reassessed by other groups of researchers, such as Toral et al. (2018), who observed specific shortcomings referred to the WMT evaluation. First, the translations to be evaluated were problematic

since the source texts were translations from other languages. That is, texts which were not considered appropriate for evaluation since they may present issues with paraphrasing, idiomaticity etc. Second, the evaluators were not always chosen among professional translators but they included participants of the study or remote crowd-workers, preferring more direct translations. Another critical aspect observed by the authors concerns the perception of translation quality, which presents a margin of variability according to the evaluator (i.e. end-users, MT developers, and professional translators). Similar conclusions were propounded by a work by Läubli et al. (2020). More specifically, the experiment's results pointed to problems dealing with the type of evaluation performed at the 2018 WMT by text segment. Hence, it cannot really take into account the text as a coherent whole. In addition, as argued by the authors, the perceived quality in human evaluation depends on a number of variables such as the choice of the evaluators, the availability of the linguistic contexts and the creation of reference translations.

On the controversial notions of human parity and superhuman performance in MT studies, it is worth recalling Toral's (2020) point of view. In his work, the concept of "super-human" in the context of artificial intelligence is discussed, notably in games such as *Go*. Nonetheless, Toral underscores that the simplicity of *Go*, mainly concerning the number of possible moves at each stage of the game, contrasts sharply with the complexity of human languages, particularly in the domain of translation. Unlike *Go*, language does not have a clear winner, and there is no single definitive solution. Accordingly, labelling AI systems achieving human-level performance in translation as "human parity" is problematic, as machine translation (MT) operates fundamentally differently from human translators. While AI systems may excel at text cohesion to some extent, achieving discourse external coherence remains beyond their reach due to the requirement of worldly knowledge. This limitation extends to any context-dependent decisions that extend beyond a single sentence.

Hence, human parity and superhuman performance represent controversial issues since, for all the factors just discussed, several criticalities arise when dealing with the evaluation process of translations. Indeed, a closer look at current MT systems demonstrates that they are still far from reaching the aforementioned "human parity". NMT generally considers the

sentence level, despite the efforts of some recent systems trying to include larger contexts. If one considers some specific aspects, the advances in the field are evident. For example, NMT systems, thanks to transformers, managed to assemble different fragments of texts overcoming the problem of ill-formed sentences. Nonetheless, MT still remains quite literal since it is based on knowledge inferred from large collections of parallel data. Moreover, another major issue regards the type of text to translate and the languages involved. In particular, the annual WMT evaluators report that human-like performance is only reachable for some specific language pairs. Overall, it has been argued that MT works better with purely informative texts written in a direct and simple style. On the other hand, when dealing with tasks including different types of texts, terminological issues arise.

5 Productivity of Machine Translated Texts

Concerning the validity and viability of automatic and human translations, one of the main issues to consider is productivity. Technically speaking, with "productivity" we indicate "the ratio of the quantity and quality of units produced to the labour required per unit of time" (O'Brien 2011: 2). Although this definition only looks at the economical aspect of productivity, it conveys the importance of speed of translation in an ever-changing, demanding society looking for the best product obtained in the most limited time. Hence, being productivity one of the major concerns of companies and organisations needing translation services, it is fundamental to better analyse this concept as applied to the translation process. In particular, post-editing productivity also involves the cognitive effort required to achieve the result. In particular, analysing effort in translation means observing how much time and cognitive work is involved during the process. In other words, effort and productivity are inversely related.

Specifically, O'Brien (2011) examines two automatic metrics employed to predict the quality of MT output: i.e. General Text Matcher (GTM) (Turian and Melamed 2003) and Translation Edit Rate (TER) (Snover et al. 2006). GTM metrics assess the similarity between the raw MT output and reference sentence having precision, recall, and their harmonic mean as main criteria. Also, GTM metrics can match adjacent words. According to Turian and Melamed (2003),

the main point of strength of this metric is that it correlates well with the human judgement of adequacy and fluency, two factors of crucial importance in the human evaluation of machine-translated texts (Ma and Cieri 2006).

TER measures the number of edits required to change raw MT output into a reference sentence. The developers of TER tried to achieve the highest correlations with human judgements. The main reason why it was selected for the experiment is that, unlike other metrics, TER does not require a large number of reference sentences to correlate with human judgements. Besides, since the focus of O'Brien was to investigate the effort involved in the translation process, it was selected as this metric records the number of edits necessary to convert raw MT output into a reference sentence. Based on the data obtained from the analysis carried out in the study, the author concludes that there is significant evidence to demonstrate that MT automatic metrics (at least the two under investigation) and actual post-editing productivity do correlate. Hence TER and GTM can be considered reliable metrics that convey post-editing productivity. Nonetheless, the limits of this research are also highlighted, suggesting further investigations that also test them in more detail in terms of the accuracy level of the individual segment. Finally, it would be worth exploring the impact of different language pairs, directions and domains.

6 Towards Standardisation and Customisation of Translation Quality

As regards the translation quality assessment methodologies employed in industry, the development of lists of error types to evaluate translation started to spread in the '90s. One of the most influential models worth reporting is the Localisation Industry Standards Association (LISA) which has continued to be employed in its different adaptations even though it ceased in 2011. LISA is based on a model that includes errors categorised according to three main levels: minor, major, and critical in the opinion of the evaluator. The translation output can be accepted or rejected based on the threshold predefined by the evaluator. That is to say, the status depends on how tolerant or demanding they decide to be. Despite the possibility of customising LISA according to the company's specific requirements, one main drawback of the model is that it does not allow us to have intermediate levels of acceptability of the translation since it can only be either accepted or rejected.

The tendency to objectify translation quality according to quantifiable criteria has led to an urge to standardise the process and, thus, to develop an ISO certification parameter, i.e. the ISO/TS 116699: 2012. It is a guideline standard that serves as guidance concerning best practices for all phases of the translation project. The ISO consists of a framework of 21 parameters classified into five main areas: source content, requirements for the target, production tasks, environment, and relationship. The standards conceive translation quality in these terms: "When both requesters and translation service providers agree on project specifications, the quality of a translation – from workflow and final delivery perspective – can be determined by the degree to which the target content adheres to the predetermined specifications" (ISO/TS 11669: 2012).

However, several scholars disagree with this definition of translation quality. For instance, Koby et al. (2014), to contrast this view, oppose a broad and narrow definition of translation quality. The broad view assumes that there cannot be absolute specifications valid for all translation activities and requirements. On the other hand, the narrow definition focuses on the textual aspect of translation. That is, activities and processes such as summarising, paraphrasing etc. are not considered as part of the translation process. Hence, explicit specifications, according to this view, are often unnecessary because requesters cannot have a clear picture of what a translation project requires. An interesting point to stress is that to evaluate translation quality a proper identification of translation nods is fundamental and "any effort to measure translation quality is doomed by confusion without an explicit definition of translation quality" (Koby et al. 2014).

An important attempt to develop indicators to achieve a more effective translation quality assessment comes from the Translation Automation User Society (TAUS). Different stakeholders were employed to achieve this goal. Several variables were considered including communicative function, end-user requirements, context, modes of translation (i.e. HT, raw MT output and post-edited MT), content profiling and quality estimation (Castilho et al. 2018). One of the most significant achievements of TAUS is the adoption of the Dynamic Quality Framework (DQF), where quality issues started to be considered before translating.

In the same research context, the EU-funded QTLaunchPad project developed the Multidimensional Quality Metrics (MQM) framework. It provides a flex-

ible way to create and use appropriate metrics for each translation task that can meet both the requester's and users' expected outcomes. The main affordance of the MQM is that it provides a shared metric that can be used for human and machine translation. Based on the identification of over 100 specific translation issues, the metric can be selected by the users depending on the type of project requirements and priorities to support and improve the translation assessment phase. Moreover, once set the specific metric with the selected parameters, it can be stored in a library to be easily reused across similar projects in the future. More specifically, the metric is defined as completing the following tasks:

- Task 1 "Complete specifications": it defines expectations about the translation and serves as the basis for contractual obligations. The 21 parameters included in this task cover all aspects of the translation product, project, and process.
- Task 2 "Select dimensions": MQM dimensions are high-level aspects including fluency, accuracy, verity, design, and internationalisation.
- Task 3 "methods": aimed at minimising human effort, this section includes a basis for the assessment: i.e. analytic, holistic, task-based.
- Task 4 "select issues": for each dimension, a number of related issues must be chosen to measure it following specifications. For example, when assessing fluency, and how linguistically well formed the target or source text is, the issues to be selected may concern spelling, grammar, register, and style. Nonetheless, it must be noticed that the selected issues vary from the genre and text type. For example, style may not be relevant when dealing with technical texts.
- Task 5 "Set issues weights": weights are used to set the relative importance of different issues. For example, terminology may have a different weight than style in certain types of texts.
- Task 6 "Determine thresholds": they can be set per issue or for dimension and are extremely important as they set the criteria of acceptability of the translation output expressed in percentage values.
- Task 7 "implement a workflow": each MQM metric must be implemented in an appropriate workflow with accompanying assessment tools which may include "sanity checks" as well as objective outcomes and decision response: i.e. approved, perform inspection, send back to the translator etc.

7 Conclusion

The paper has addressed the relationship between automatic and human translation, mainly in terms of productivity and quality, taking into account the most influential studies in the field of translation, artificial intelligence and computational linguistics. To investigate this question, a technical overview of the main approaches to MT was provided, focused on how different MT systems work, the paradigms behind their development and the main linguistic criteria included. As regards the controversial issue of translation quality, human parity and super human performance, it can be argued that translation is a complex process involving multiple domains: cognitive, social, cultural, and technological. Hence, finding a unique definition of translation quality that takes into account the multidisciplinary aspect of the process, and capturing its intrinsic complexity is not an easy task. In addition, the discussion on the use of MT, post-editing and automatic metrics shows, essentially, that the post-editing of MT texts has become a part of the translation workflow, raising new important questions in translation research.

What emerges from the analysis of these automatic metrics is that the rise of MT and, consequently, of MT output has contributed to evaluating translation quality a much debated topic in the translation research community. As Lommel et al. (2014) point out, "human and machine translation [...] quality evaluation methods have been fundamentally different in kind, preventing comparison of the two". If on one hand, it is true that, in the translation industry, quality is mainly related to customer opinion, on the other, it can be noticed a tendency to quantify the assessment process. Indeed, the evaluation models used are mainly error-based, targeted at computing the number of errors detected, classified according to certain standards, and weighted by a reviewer or post-editor. This requires the evaluation system to set the predetermined errors according to logical or hierarchical criteria to be acknowledged and used to provide an effective and objective evaluation of the translation output.

Another crucial point that is worth recalling from the studies reviewed is that post-editing, since its introduction in the translation industry, has obtained a more central position due to the acknowledged benefits in terms

of the productivity of translators. Nonetheless, as human-machine interaction has increased in professional practice, due to the continuing growth in digital content, the aforementioned translation quality has become even more challenging to define, capture, and assess. To operationalise and measure translation quality, different attempts have been made with the aim of achieving evaluation standards based on quantitative criteria (i.e. the ratio of quantity and quality to time). However, the limits of automatic metrics have been discussed pointing to the need to customise the evaluation process, both for human and MT output, according to the specific requirement of the user. It is essential to choose an approach that overcomes dichotomies and is able to join together the two opposite sides of a continuum between the source-oriented concept of fluency and the target-oriented concept of accuracy. Hence, the state of the art suggests that a pragmatic, targeted to end-users method is needed that takes into account the notion of adapting the evaluation system to the purpose of the translation.

To conclude, it can be argued that this overview of methodologies and approaches to translation, post-editing and translation quality assessment, discussing their strengths and weaknesses, is by no means exhaustive. It aims at identifying a number of perceived issues concerning the translation field, in terms of target to achieve, processes, and product evaluation. What arises, particularly for the notion of quality, is that the integration of translation technologies has profoundly changed the relationship between humans and machines, making the boundaries between the two more blurred. Understanding how translation technologies evolve and develop and how the most effective and appropriate evaluation approach can be selected is essential, nowadays, to successfully integrate these technologies in the translation industry. Hence, the ability to adapt to new translation tools and to conceive translation quality with more flexibility and fluidity is crucial considering the impact it may have in terms of effectiveness. Indeed, the translation industry and research "need a way to compare different sorts of translation as objectively as possible, with an emphasis on identifying problems and the metrics adopted to this end should be built on a well-defined foundation including at least clearly stated definitions of translation, quality, and translation quality" (Koby et al. 2014).

References

Abiola, O.B, Adetunmbi, Adebayo, Oguntimilehin, Abiodun, 2015, "Using hybrid approach for English-to-Yoruba text to text machine translation system (proposed)" *International Journal of Computer Science and Mobile Computing* 4(8), 308-313.

Akinwale, O.I., Adebayo, Olusola, Adetunmbi, Olumide Olayinka, Obe and A. T. Adesuyi, 2015, "Web-Based English to Yoruba Machine Translation" *International Journal of Language and Linguistics* 3(3), 154-159.

Bahdanau, Dzmitry, Cho, Kyunghyun, Bengio, Yoshua, 2014, "Neural machine translation by jointly learning to align and translate" *Computation and Language* arXiv preprint arXiv:1409.0473.

Barrault, Loïc, Bojar, Ondřej, Costa-jussà, Marta R., Federmann, Christian, Fishel, Mark, Graham, Yvette, Haddow, Barry, Huck, Matthias, Koehn, Philippe, Malmasi, Shervin, Monz, Christof, Muller, Matthias, Pal, Santanu, Post, Matt, and Zampieri, Marcos, 2019, "Findings of the 2019 conference on machine translation (WMT19). In Proceedings of the Fourth Conference on Machine Translation" *Florence, Italy, August. Association for Computational Linguistics*, 2: 1-61.

Belouadi, Jonas and Steffen Eger, 2022, "UScore: An Effective Approach to Fully Unsupervised Evaluation Metrics for Machine Translation." *ArXiv* abs/2202.10062.

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

Bojar, Ondřej, Federmann, Christian, Fishel, Mark, Graham, Yvette, Haddow, Barry, Koehn, Philipp, and Monz, Christof, 2018, "Findings of the 2018 conference on machine translation (WMT18)" In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers,* pages 272–303, Belgium, Brussels, October. Association for Computational Linguistics.

Carl, Michael, Dragsted, Barbara, Elming, Jakob, Hardt, Daniel and Arnt Lykke Jakobsen, 2011, "The process of post-editing: a pilot study" in Bernadette Sharp et al. (eds) *Proceedings of the 8th International NLPSC workshop. Special theme: Human-machine interaction in translation*, Copenhagen Studies in Language 41. Frederiksberg, Samfundslitteratur, 131–142.

Castilho, Sheila, Doherty, Stephen, Gaspari, Federico, Moorkens, Joss, 2018, "Approaches to Human and Machine Translation Quality Assessment" in: Moorkens, J., Castilho, S., Gaspari, F., Doherty, S. (eds) *Translation Quality Assessment. Machine Translation: Technologies and Applications,* vol 1. Springer, Cham.

Cettolo, Mauro, Niehues, Jan, Stüker, Sebastien, Bentivogli, Luisa and Marcello Federico, 2013, "Report on the 10th IWSLT evaluation campaign" in *Proceedings of the 10th International Workshop on Spoken Language Translation: Evaluation Campaign,* Heidelberg, Germany.

Edmundson, Harold P. and David G. Hays, 1958, "Research methodology for machine translation" *Mechanical Translation* 5(1), 8-15.

Fiederer, Rebecca and Sharon O'Brien, 2009, "Quality and machine translation: A realistic objective?" *The Journal of Specialised Translation* 11, 52–74.

Garcìa, Ignacio, 2012, "A brief history of postediting and of research on post-editing" in Anthony Pym and Alexandra Assis Rosa (eds) *New Directions in Translation Studies. Special Issue of Anglo Saxonica* 3(3), 292–310.

Gaspari, Federico, Almaghout, Hala and Stephen Doherty, 2015, "A survey of machine translation competencies: insights for translation technology educators and practitioners" *Perspectives: Studies in Translatology*, 1–26.

Hassan, Hany, Aue, Anthony, Chen, Chang, Chowdhary, Vishal, Clark, Jonathan, Federmann, Christian, Huang, Xuedong, Junczys-Dowmunt, Marcin, Lewis, William, Li, Mu, Liu, Shujie, Liu, Tie-Yan, Luo, Renqian, Menezes, Arul, Qin, Tao, Seide, Frank, Tan, Xu, Tian, Fei, Wu, L., Wu, Shuangzhi, Xia, Yingce, Zhang, Dongdong, Zhang, Zhirui, and Ming Zhou, 2018, "Achieving Human Parity on Automatic Chinese to English News Translation" *Microsoft AI & Research* arXiv.

International Organization for Standardisation 2012 ISO/TS 11669:2012 technical specification: translation projects – general guidance. International Organization for Standardisation, Geneva. Available via: https://www.iso.org/standard/50687.html.

Koby, Geoffrey S., Daryl R. Hague, Arle Lommel and Alan K. Melby, 2014, "Defining translation quality" *Revista Tradumàtica* (12), 413–420.

Koponen, Maarit, 2016, "Is Machine Translation Post-editing Worth the Effort? A Survey of Research into Post-editing and Effort" *The Journal of Specialised Translation* (25), 131-148.

Lommel, Arle, Uszkoreit, Hans, Burchardt, Aljoscha, 2014, "Multidimensional Quality Metrics (MQM): a framework for declaring and describing translation quality metrics" *Revista Tradumàtica* 12:455–463.

Ma X, Cieri, C, 2006 "Corpus support for machine translation at LDC" in *Proceedings of the Fifth International Conference on Language Resources and Evaluation*, Genoa, Italy, 22-28 May, 859- 864.

Marie, Benjamin, Fujita, Atsushi and Raphael Rubino, 2021, "Scientific credibility of machine translation research: A meta-evaluation of 769 papers" in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing* (Volume 1: Long Papers), pages 7297–7306, Online. Association for Computational Linguistics.

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

O'Brien, Sharon, 2011, "Towards predicting post-editing productivity", *Machine Translation* 25, 197–215.

Oladosu, John Babalola, Esan, Adebimpe, Adeyanju, Ibrahim, Adegoke, Benjamin, Olaniyan, Olatayo & Omodunbi, Bolaji, 2016, "Approaches to Machine Translation: A Review", *Journal of Engineering and Technology* 1(1), 120-126.

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, and Christof Monz.

Papineni, Kishore, Roukos, Salim, Ward, Todd, Zhu, Wei-Jing, 2002, "BLEU: a Method for Automatic Evaluation of Machine Translation" In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics* Stroudsburg PA, USA, 311-318.

Plitt, Mirko and François Masselot, 2010, "A Productivity Test of Statistical Machine Translation Post-Editing in a Typical Localisation Context", *The Prague Bulletin of Mathematical Linguistics* 93, 7–16.

Quirk, Chris. & Corston-Oliver, Simon, 2006, "The impact of parse quality on syntactically-informed statistical machine translation" In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, 56-63.

Snover, Matthew, Dorr, Bonnie, Schwartz, Rich, Micciulla, Linnea, Makhoul, John, 2006, "A study of translation edit rate with targeted human annotation" in *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas,* August 8-12, Cambridge, Massachusetts, USA, 223-231.

TAUS, 2010, Machine Translation Post-editing Guidelines. https://www.taus.net/think-tank/best-practices/postedit-best-practices/machine translation-post-editing-guidelines.

Thang, Luong, Hieu, Pham, and Christopher D. Manning, 2015, "Effective Approaches to Attention-based Neural Machine Translation" in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1412–1421, Lisbon, Portugal. Association for Computational Linguistics.

Toral, Antonio, 2020, "Reassessing claims of human parity and super-human performance in machine translation at WMT 2019" in *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation,* pages 185–194, Lisboa, Portugal, November. European Association for Machine Translation.

Toral, Antonio, Castilho, Sheila, Hu, Ke, and Andy Way, 2018, "Attaining the unattainable? Reassessing claims of human parity in neural machine translation" in *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 113–123, Brussels, Belgium, October. Association for Computational Linguistics.

Turian, Joseph, Shen, Luke, Melamed, I. Dan, 2003, "Evaluation of machine translation and its evaluation" in *Proceedings of the MT Summit IX*, New Orleans, USA, 23-27 September 2003, 386-393.