Translation Quality Management in the AI Age. New technologies to Perform Translation Quality Management Operations

Jennifer Vela-Valido

revista.tradumàtica
tecnologies de la traducció

Vela-Valido, Jennifer
Universidad de Las Palmas de Gran Canaria
jennifer.vela101@alu.ulpgc.es;
ORCID: 0000-0003-4690-4542

Abstract

This article presents a selection of some of the latest technologies to perform translation quality management operations such as quality assurance and translation quality assessment using artificial intelligence and machine learning; it also discusses the impact of these technological advances in the latest academic research trends and the translation industry.

Keywords: translation quality, quality evaluation, quality assessment, quality control, quality management, translation industry, AI, translation technology

Resum

Aquest article presenta una selecció de les tecnologies que permeten dur a terme tasquesde gestió, avaluació i control de la qualiat de la traducció a través dels avenços de la intel·ligència artificial i l'aprenentatge automàtic. També revisa l'impacte d'aquestes tecnologies sobre les tendències investgiadores i sobre la indústria de la traducció.

Paraules clau: qualitat de la traducció, avaluació de la qualitat, estimació de la qualitat, control de qualiat, gestió de la qualitat, indústria de la traducció, IA, tecnologies de la traducció

Resumen

Este artículo presenta una selección de tecnologías para llevar a cabo tareas de gestión, evaluación y control de calidad de la traducción utilizando los últimos avances en inteligencia artificial y aprendizaje automático. También revisa el impacto de estas tecnologías en las tendencias investigadoras y en la industria de la traducción.

Palabras clave: calidad de la traducción, evaluación de la calidad, estimación de la calidad, control de calidad, gestión de la calidad, industria de la traducción, IA, tecnologías de la traducción

Rebuda: 30 de gener de 2021 / Acceptació: 1 de juliol de 2021 / Publicació avançada: 15 de desembre de 2021

Revista Tradumàtica. Número 19 (desembre 2021).
Cita recomanada: VELA-VALIDO, JENNIFER; (2021). «Translation Quality Management in the AI Age: New Technologies to Perform Translation Quality Management Operations». Revista Tradumàtica. Tecnologies de la Traducció, 19, 93-111.
<https://doi.org/10.5565/rev/tradumatica.285>

http://revistes.uab.cat/tradumatica
ISSN 1578-7559

1. Introduction

The translation industry, and especially language service providers, have traditionally concentrated their efforts on making use of the latest technologies to increase the efficiency of the translation process (Doherty, 2016). Proof of this is the level of sophistication of the current translation tools and the application of computer paradigms to create automatic translations, first with statistical models (statistical machine translation) and, more recently, with neural models (neural machine translation).

This uptake in the use of computer-assisted translation tools and machine translation engines has evolved in parallel with different models and tools to support translation quality assurance and translation quality assessment tasks (Doherty et al., 2018), giving way to new challenges and questions. For example: How do these new technologies affect the provision of written language services? What are the advantages and disadvantages of incorporating Artificial Intelligence (AI) and Machine Learning (ML) into translation quality management workflows, such as quality assurance tasks or quality assessment tasks?

This article tries to answer these questions by providing an overview of some of the most recent AI-powered quality assurance and assessment technologies and tools used in the translation industry. It also seeks to open up new perspectives on the use of AI and ML in the quality management workflows of both human and machine translations, a topic that is an increasingly important area in the translation quality management field although, due to its novelty, has yet to gain more significant weight in scientific research.

2 Definitions

One particular challenge when it comes to translation quality management is the terminological inconsistency coming from both professional practice and academic research, which means that the names of the related tasks (such as translation quality evaluation, translation quality control, translation quality assurance, translation quality assessment, and so on) vary depending on the approach (academic vs professional), the author or the procedure (evaluation of the translation process vs evaluation of the translation product). In this article, we will use the terminology that is currently more widely used and accepted by both academia and the industry, including international standards like ISO 9000 (which plays a major role in the translation industry, as it defines the fundamentals and key terminology of quality management), ISO 17100:2015 (which establishes the requirements for translation services), and some of the definitions contained in ASTM F2575 (the standard guide for quality assurance in translation published by ASTM International). We have also included some key definitions provided by Arle Lommel and Alan K. Melby, two of the most prominent scholars specialised in translation quality research, and Lucia Specia, a researcher specialised in machine translation quality evaluation and machine translation estimation systems. The combination of these resources will allow us to make a clear distinction among some concepts that might sometimes appear to overlap or be used inconsistently.

2.1 End to end processes

2.1.1 Quality management (QM)

"The integration and coordination of management activities focused on ensuring the organization fulfils stakeholder requirements predictably, consistently, and reliably" (Lommel & Melby, 2018: 4). The ASTM WK46369 proposal echoes the trend that is already followed by the principal actors in the translation industry and establishes that any translation quality management system should be compatible with the principles and key concepts contained in ISO 9000 (ASTM, 2021). ISO 9000 states that quality management "can include establishing quality policies (3.5.9)[1] and quality objectives (3.7.2), and processes (3.4.1) to achieve these quality objectives through quality planning (3.3.5), quality assurance (3.3.6), quality control (3.3.7), and quality improvement (3.3.8)." (ISO, 2015b).

2.2 Before production phase

2.2.1 Quality planning

"Part of quality management (3.3.4) focused on setting quality objectives (3.7.2) and specifying necessary operational processes (3.4.1), and related resources to achieve the quality objectives" (ISO, 2015b). In the context of translation quality planning, these activities are aimed to design a system of policies, processes, and procedures that need to be followed to be able to produce products (translations) that can meet stakeholder requirements (Lommel & Melby, 2018: 4).

2.2.2 Machine Translation Quality Estimation (MTQE)

In the context of machine translation, quality estimation can be defined as a quality management task aimed at "estimating the quality of a system's output for a given input, without any information about the expected output" (Specia et al., 2010: 40). In other words, quality estimation systems utilise automatic metrics to "predict whether a new source string will result in a good or bad translation" (Way, 2018: 173) before the production phase, rather than assessing the MT segment after production by comparing how similar it is to a reference translation segments (see section 2.4.2 below).

2.3 During production phase

2.3.1 Quality assurance (QA)

"Part of quality management (3.3.4) focused on providing confidence that quality requirements (3.6.5) will be fulfilled" (ISO, 2015b). To provide that confidence to the different stakeholders (management, customers, and even third parties) the assurance

1 Note: The numbers between parenthesis refer to the section numbering of the standard cited.

activities audit the quality processes and procedures put in place (Lommel & Melby, 2018). It is important to note that quality assurance is often used as a synonym for quality assessment in the industry, and this creates quite a lot of confusion. The main goal of the quality assurance workflows is to improve the product to the agreed quality, while the quality assessment activities (as we will see in the definition below) aim to evaluate the quality of the final product. Translation quality assurance activities take place during the production phase, and can include the following tasks:

Revision (also referred to as the first step of the "editing" process in the standard ASTM F2575): "Bilingual examination of target language content (2.3.3) against source language content (2.3.2) for its suitability for the agreed purpose" (ISO, 2015a). The standard ASTM F2575 also mentions that the main goal of the reviser is to check the accuracy of the translation and the correctness of the terminology (ASTM, 2006).

Review (also referred to as the second step of the "editing" process in the standard ASTM F2575): "monolingual examination of target language content [...] for its suitability for the agreed purpose" (ISO, 2015a). According to ASTM F2575, the reviewer focuses only on the target text to check coherence and readability, although they can check the source text if necessary (ASTM, 2006).

Formatting and compilation: This task might vary significantly depending on the characteristics of the project and the specifications, the applications used and even the languages required (ASTM, 2006).

Proofreading and Verification: According to ASTM F2575, this can be a quality assurance step or part of the quality control step. It can also be performed after the editing phase or at the same time. In any case, the proof-reader's mission is to focus on checking the target text for typographical errors, formatting issues, or incorrect spelling (ASTM, 2006).

2.3.2 Quality control (QC)

"Part of quality management (3.3.4) focused on fulfilling quality requirements (3.6.5)" (ISO, 2015b). The standard ASTM F2575 considers that the translation QC step is linear and should consist "of random sampling or a full check of final deliverables or both as the last step in the process" (ASTM, 2006: 10). However, more recent studies from authors such as Lommel and Melby (Lommel & Melby, 2018) consider that translation quality control activities should assess processes and performance in real-time, that is, during the whole production phase, to verify that the quality measures are being fulfilled.

2.4 After production phase

2.4.1 Quality assessment or evaluation (also referred to as "post-project review" or post-mortem in the standard ASTM F2575)

"Performance evaluation procedure conducted at the end of a project to determine how well the project conformed to original specifications" (ASTM, 2006: 3). Ideally, this step takes place before the delivery to the requester, although it can also be carried out by

the requester when accepting the delivered translations to evaluate whether their quality requirements have indeed been fulfilled, and to compare the results against the Key Performance Metrics (KPIs) agreed. To avoid any confusion with the abbreviation used for Quality Assurance (QA), some scholars (Colina, 2008; Görög, 2017; Jiménez-Crespo, 2009; O'Brien, 2012) and recent standards (ASTM, 2021) lean towards the use of Translation Quality Evaluation (TQE) or Translation Quality Assessment (TQA), which can be used interchangeably in this context. Translation quality assessment can be performed to evaluate both human and MT outputs.

2.4.2 Machine Translation Evaluation (MTE)

Evaluation or assessment of MT systems via their output, either with human evaluations or with automatic metrics. "The main purpose of the state-of-the-art automatic evaluation metrics is to compare the output of an MT system, which are assumed to be good, because they are human quality" (Castilho et al., 2018: 25). Some of the most popular automatic metrics used nowadays both in the industry and research projects are the Bilingual Evaluation Understudy, or BLEU (Papineni et al., 2001) and METEOR (Lavie & Agarwal, 2007), although, as we will see later in this article, several new and promising metrics are now able to outperform BLEU and METEOR in terms of correlation with human judgments. These automatic metrics can be used to evaluate how much effort would be required for post-editing, to assess the evolution and efficiency of different iterations of the engine used, to compare efficiency gains before and after an engine has been trained, and to assess how well different engines are suited to the type of text to be translated, the language pairs chosen, or the quality requirements of the translation Project.

Quality Management		
Before production	During Production	After production
Quality planning Quality estimation (QE)	Quality Assurance (QA) Quality Control* (QC)	Quality Assessment (TQA) /Quality Evaluation (TQE)

Figure 1. Quality Management process, as outlined in the definitions of this article. Note: Quality Control is considered in this article a linear step, as defined by the standard ASTM F2575.

3 Translation quality management in the AI Age

Several researchers and technology experts have signalled that our society is now going through a "fourth industrial revolution" (Schwab, 2016) or "Intelligence Revolution" (Marr, 2020) thanks to the recent developments in AI and big data. But what is AI exactly and

how does it affect the translation field? There seems to be some confusion around the terms machine learning, AI, and deep learning, as sometimes these terms seem to be used interchangeably even though they mean different things. Artificial intelligence is the catch-all term that covers machine learning and deep learning, and it refers to the machines being used to enhance decision-making in specific fields of expertise, hence why several experts prefer to use the term "narrow AI" (Dickson, 2019). Machine learning refers to the "training of computers, using algorithms, to parse data, learn from it and make informed decisions based on the accrued learning" (Vandenberg, 2018: 47). Finally, deep learning refers to a subset of machine learning in which multi-layered neural networks learn from vast amounts of data.

To better understand the role that AI plays in the development of different quality management processes and activities, it is important to go back to the first use of AI to enhance machine translation outputs. One of the important breakthroughs in natural language processing and machine translation came from a Google Research group in 2013, the year when the group published a paper called "Efficient Estimation of Word Representations in Vector Space" (Mikolov et al., 2013). This research group took Google News corpus, pushed it into a neural engine, and allocated 300 vector values to each word in the corpus, which allowed them to calculate what the following words would be in relation to the current words and what were the most probable words in the current context given the surrounding words. This process provided a very comprehensive way of associating words to each other, opening the door to further developments in this field. A few years later, Facebook AI research group, headed by Piotr Bojanowski (Grave et al., 2018), took this principle and applied it to the whole data contained in the Internet using crawling technologies. Thanks to this new approach, the team was able to produce a very complete language model of each of the languages in scope (157), opening the way to more sophisticated neural translation technologies. Instead of training machines on a given corpus for a specific purpose and with a specific vocabulary (as happens with statistical machine translation models), the vector space allowed this training to be done on a comprehensive total corpus of data to analyse the relationships between those words. As a result of this breakthrough, many of the companies and research groups that were already working on the applications of AI to translation production using NMT engines started applying the same technology to other human or machine translation-related tasks, such as translation quality assessment, translation quality assurance, machine translation quality estimation, and translation quality metrics and indexes.

3.1 Translation quality assurance and assessment for human translation

Nowadays, there are two main trends in terms of methodology for translation quality assurance and assessment workflows and metrics, depending on whether the translation has been done by a human or by a neural machine translation engine.

In the first case, all the translation quality assurance and translation quality assessment tools that are used in the industry (like Xbench, QA Distiller, or Verifika) are

conceived as support tools for human translation experts. This is mostly due to the fact that, even though these tools have improved (and will continue improving) significantly in the past years, there are still certain categories of mistakes that these tools are not able to detect automatically (for example, meaning, tone of voice or style). These tools can also produce "false-positive" errors, such as a difference in length from source to target (even though different languages have different semantic and morphological structures) or a difference in spacing rules (for example, a number and its unit measures are written without a space in-between in English while in Spanish this space is mandatory). For this reason, it is essential to count with a quality assurance specialist or an experienced translator who can accurately assess the settings of these tools, select the most appropriate ones for each type of language pair and translation requirements, fine-tune them to try to reduce these "false positives" and analyse the reports generated by these tools to detect which errors are "false positives" and mark them accordingly.

Quality risk assessment is another quality task that has traditionally been mostly human-driven and focused on human translation outputs. For this reason, this process has the potential to benefit greatly from new data-driven approaches and AI technologies to make better process and quality management decisions, and there are several tools and applications that are exploring this approach.

In the following section reviews four very promising examples of AI applied to translation quality assurance, quality assessment tasks, and quality risk assessment of human translations with different degrees of automation, and discusses the potential future uses and the limitations of these proposals.

3.1.1 Human translation quality assurance and assessment: Inter-language Vector Space (ILVS)

The research team of XTM International built on the neural language processing framework developed by Google and Facebook AI research group using vector space algorithms by taking each vector space for each language model (source and target) to create what they called "Inter-language Vector Spaces" (or ILVS) (Jaworski, 2020a). This new technology is based on "deep learning, neural networks and algebraic algorithms for supervised learning of vector transformations" (Jaworski, 2020b: 2) and consists of three phases: First, a monolingual text corpus for different languages is fed to a neural network. This network is able to compute 300-dimensional vector representations (called word embeddings) for all the words in the corpus. After that, there is a training phase in which several transformation matrixes convert these vectors between the languages. To do this, the network is fed with the text corpus mentioned before, which contains a great number of contexts for each word. Once the training of the neural network has been completed, the network is able to predict the context of a word using the skip-gram model. Finally, the converted vectors are stored in an index in multiple languages (Jaworski, 2020b).

In terms of its potential application to translation quality management, the developers of this technology affirm that it can be incorporated into a CAT tool to perform automatic

translation quality tasks that would be performed during or after the translation cycle, in particular, to highlight probable translation errors in real-time in categories that have traditionally proved to be quite difficult (if not impossible) to be checked automatically, such as meaning, since the contextual similarity captured by the ILVS allows this technology to perform a semantic analysis of the source and target translation.

Since this technology is still being tested by the research team, this could be a great opportunity for some academic or industry research teams to design some testing scenarios with real cases to explore the efficiencies and capabilities of this new technology in different translation environments and conduct an empirical analysis comparing this approach with other methods used to perform translation quality assessment and evaluation.

3.1.2 Automatic text quality assessment: NwQM, a neural quality assessment framework for Wikipedia

In a paper published in October 2020 by researchers from IIT Kharagpur and Adobe Research, the authors presented NwQM, a deep machine learning model that takes different key information sources such as article text, metadata, and images to assess the overall quality of the articles published in Wikipedia (Reddy et al., 2020). Although the main goal of the model is to perform an automatic quality assessment of the articles written in English and does not focus on translated texts, the approach taken to evaluate the quality of the texts analysed could potentially be adapted and used as an additional tool to evaluate the quality of a human translated text, just as the strategies developed initially to enhance natural language processing tasks such as natural language understanding and natural-language generation helped the development of the machine translation technology.

One of the key reasons why it would be very interesting to study the potential applications of this neural quality assessment framework in a translation quality management model and compare it to other current approaches to document-level MT assessment is that, as will be seen further on, NwQM is able to detect mistakes across a whole text, such as long-range semantic errors, and coherence inconsistencies (for example, wrong personal pronouns in subsequent sentences). This means that revisers and evaluators could have a more holistic view of the quality of a complete target text, instead of just the quality of each separate sentence of a given translation.

But how does this model work exactly? First, the authors relied on the use of a contextual representation to encode the text contained in the article by segmenting the articles organically based on sections. After this first step, they applied BERT (Devlin et al., 2019), a language pre-training model used to get contextual word representation to find explicit signals about the article quality. These signals included the presence of bias, coherent wording, and grammar correctness. Finally, the researchers applied the end-to-end fine-tuning strategy proposed to enhance the accuracy of the model. According to the results published in the paper, the use of this new machine learning model based on the BERT model has shown that "fine-tuned BERT performs extremely well in subject-verb agreement task [...] hence it is able to capture long-range semantic dependency.

Besides, since it is pre-trained with next sentence prediction objective, it remembers context across multiple sentences" (Reddy et al., 2020: 4).

However, it is important to note that this proposal is specifically designed to optimise and automatise the text quality assessment of large amounts of texts, such as the corpora hosted in Wikipedia. For this reason, this model has three potential limitations that should be taken into account before considering testing its capabilities in the assessment of human translations. First, the accuracy of the results of the neural quality assessment model is highly dependent on the volume of information fed into it, which means it would be less reliable when used with smaller corpora. Second, this model is completely automatised and has not been tested yet within a translation quality framework, therefore, it would be essential to test and compare its accuracy using human evaluators. And finally, the experiment only used texts in English, so it is not clear how the performance of NwQM would be in other languages.

3.1.3 Translation metrics and quality indexes using business analytics: the holistic TQI proposed by Wordbee

One of the key points of an effective translation quality assessment system is that it needs to have an analytic approach to quality, and it has to be measurable, repeatable, and objective. Only this way will it be possible to manage the quality of the translations produced. The result of these measurements (or metrics) is called the TQI (Translation Quality Index). This index may vary depending on the purpose of the quality assessment process, the context, or even the scope (external, to standardise the outsourced translation tasks; or internal, to measure the quality of a given special translation task).

There are several standard TQIs in the translation industry, such as the one proposed by the LISA QA Model in 1995 (updated until 2007, when version 3.1 was launched) (Mateo, 2014), the SAE J2450 (SAE International, 2005), a standard presented by a working group made of SAE and GM representatives in 2005 (Petrova, 2019) or, more recently, MQM-DQF, proposed by TAUS and DFKI in 2015 (DFKI, 2015). All these TQIs follow the analytic approach and therefore focus on assessing the product quality.

However, there is also a complementary approach that focuses on assessing process quality by combining business analytics tools and specific KPIs (key performance indicators) used in the common business practice to come up with a holistic vision of the translation quality provided by an external party (Muzii, 2014). One of the industry pioneers of this new approach to TQI is Wordbee, a translation management system that tracks, compiles, and analyses all the data generated in its environment by its users, and transforms it into what is called "business analytics". These AI-powered analytics are able to provide an analysis of the typology and recurrence of past errors, measure the current performance, and predict future outcomes and performances. Consequently, their integrated business analytics tool can correlate the quality translation data gathered with the KPIs to offer a very innovative and comprehensive TQI.

KPI	Definition	Data input
Capacity Utilization Ratio (CUR)	Output produced in a given time frame (Capacity)	Daily worked volume Daily assigned volume
DIFOT (Delivery In-Full, On-Time) rate	Ability to fulfil orders and meet customer expectations (Timeliness)	Job acceptance time Timeliness rating
FPY (First Pass Yield) rate	Percentage of units coming out of a process (Effectiveness)	Adherence to instructions Reliability
Order Fulfilment Cycle Time (OFCT)	Average time to deliver a service from order to customer receipt (end-to-end delivery capacity)	Jobs completed on time Timeliness rating
Rework Level	Percentage of items inspected requiring rework (Quality)	Work Quality rating Segments requiring correction

Figure 2: Correlation between quality KPIs and quality data input proposed by Wordbee

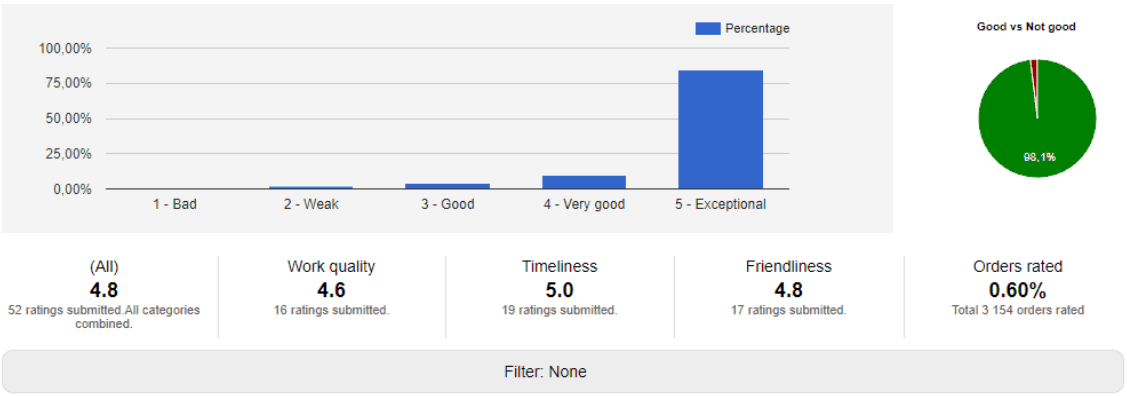

Figure 3: Client Rating section within the Business Analytics tool offered by Wordbee (Wordbee, 2018)

3.1.4 Quality risk assessment for human translation using big data and machine learning

Thanks to the latest developments in machine learning, some tech-driven translation service providers (especially developers of TMS systems) have focused their efforts on developing new tools and functionalities with the goal of automatising quality risk assessments and providing a prediction of human translation quality results (Esselink, 2019). One of the most advanced products in the market with such capability is the Smartling TMS, recognised by CSA Research as the Leader for Language-Oriented Translation Management Systems (Sargent et al., 2019). This translation platform tracks more than 100 parameters, such as the translation process followed, the time spent, the visual context available, the grammar and spell checks performed, or the translation expert rating; and determines whether each factor had an impact on quality outcomes using a machine-learning algorithm with around 75 factors. The weight of all the factors considered produces a 1-100 score called the "Quality Confidence Score" (or QCS)

(Pielmeier, 2017), which represents the likelihood that a translation is of "professional quality" if it were to be manually evaluated by a human translation expert. Another remarkable aspect about the algorithm developed by Smartling is that it aggregates data gathered from seven billion translated words processed by Smartling's translation management platform, and thus combines data and machine learning to predict translation quality.

Figure 4: CQS parameters (Pielmeier, 2017)

3.2 Translation quality assurance, assessment, and estimation for machine translation

As mentioned in the previous section of this article, machine translation quality management workflows can differ quite a lot from those used in human translations. While it is true that it is possible (and quite frequent) to use the same translation quality assurance tools in the quality assurance phase (after the post-editing step, or even at the same time), the quality assessment or evaluation methodologies present very significant differences in comparison with those used in human translation.

In order to advance in the evaluation of new and coming MT systems that have been developed in the past decade, organizations such as the Association for Computational Linguistics (ACL) organise the Conference on Machine Translation (or WMT), dedicated to the evaluation of the performance of machine translation models, both from an academic and a professional standpoint (Németh, 2019). One of the peculiarities of this conference is that its organisers release a collection of shared tasks related to machine translation every year, and the participant researchers compare their techniques and results against others in the field. These tasks include a translation task, in which participants translate a common set with their MT systems; and an evaluation task, which

focuses on automatic metrics and quality estimation. When all the tasks are completed, the WMT publishes an official ranking of the MT systems in each translation task. It is important to mention that one significant component of the WMT has always been the manual evaluation and that the official rankings are the result of the evaluation performed by human annotators, as the WMT organisers consider that human evaluation should be the standard of MT evaluation. Furthermore, this ranking "has enabled the development of automatic metrics by providing a gold standard against which metrics can be compared" (Bojar et al., 2016: 27).

The following sections introduce some of the most recent efforts to evaluate and estimate the quality of MT quality with different levels of human intervention.

3.2.1 Machine translation evaluation without human intervention: the KoBE model proposed by Google Research

Although there is a strong consensus among the industry and the scholars that human evaluation is the gold standard for Machine Translation evaluation (Birch et al., 2016), it is also clear that this type of manual evaluation is, unfortunately, time-consuming and labour-intensive (Bojar et al., 2016), especially for massive-scale experiments designed to evaluate the performance of different MT engines across different languages, scopes, and domains. This lack of scalability has already been addressed using different automatic approximations of human judgment, such as BLEU (Papineni et al., 2001) or METEOR (Lavie & Agarwal, 2007). However, both methods still have some important limitations in terms of scope, quality of results, or bias (Gekhman et al., 2020). To overcome this, the Google Research team presented the KoBE model in 2020, a new evaluation method based on a large-scale multilingual knowledge base that is publicly available in the Google Knowledge Graph Search API. To test this new approach, the team used the benchmark for evaluation without references from the Fourth Conference on Machine Translation shared task on quality estimation (WMT19). KoBE obtained the best results in 9 out of 18 language pairs analysed and produced scores closer to human judgments on an absolute scale, which seems to suggest that this knowledge-based evaluation method can be a very powerful tool to evaluate the performance of MT engines in big-scale experiments without human intervention or reference translations. To foster further research on knowledge-based evaluations using their proposed model, the research team also released a data set with 1.8 million entity mentions containing 425k sentences in 18 languages into a Github repository (Gekhman et al., 2020).

3.2.2 Quality estimation of machine translation using machine learning: COMET and Memsource MTQE

As mentioned in the definitions section, machine translation quality estimation (MTQE) is an alternative way of assessing machine translation quality in a completely automated way. The predictions or estimations can be provided at different levels: word, phrase, sentence, paragraph, or document. However, even though the first studies on MTQE were published more than 20 years ago, the use of MT was not really widespread back then,

so the interest from academia (and the industry) for this type of automated metric was quite marginal.

Despite the efforts from WMT and other research institutions to disseminate and apply the results of the MTQE shared tasks and experiments to real projects in the translation industry, only a very limited number of companies in the industry are actively using or at least experimenting with MTQE systems or metrics. However, this might change in the next few years, as we have recently seen the introduction of some relatively stable commercial MTQE solutions that will probably contribute to lower the barrier for the industry to experiment and refine this technology.

The first of these stable commercial solutions (an automatic MTQE metric) was developed by Unbabel, a translation company specialised in AI-powered, human-refined machine translation technology. This metric, called COMET, has shown since then a great level of correlation with human evaluation and metrics, such as the Multidimensional Quality Metrics (MQM), edit-distance, and direct assessment scores. The system is also able to identify incorrect words and provides an automatic quality score for a translated sentence, which helps the human post-editors to easily detect the parts of sentences that might need to be fixed according to that quality score. COMET was presented in 2016 to the WMT annual campaign and was ranked as the best MTQE system of that year with a score of 49.5%, against the 41.1% obtained by the best non-Unbabel system (Rei et al., 2020)

Another promising (and more recent) proposal was presented in 2018 by Memsource (Memsource, 2018) as an AI-powered feature integrated into their cloud-based translation management system and aimed at providing MT quality scores before post-editing to improve post-editing efficiency. This feature provides segment-level quality estimates for MT suggestions in the form of percentages, giving an automatic indication of how much editing might be required (Memsource, 2020).

There is also a very interesting debate among different researchers to try to ascertain the real level of accuracy of the automatic estimations and metrics, and some authors such as Sun, Guzmán, and Specia have recently questioned whether the promising quality estimation results obtained by some NMT engines without any human supervision are indeed accurate when it comes to measuring complex indicators such as the adequacy of translations (Sun et al., 2020). This has opened the door to several studies aimed at comparing the automatic results with the human results (Bojar et al., 2016; Freitag et al., 2021; Sun et al., 2020), or establishing a human assessment step to confirm that the results of an automatic MTQE system (such as the one presented by Memsource) are indeed accurate (Ziganshina et al., 2021).

4 Looking into the future

There is no question in the translation industry that the new AI-based technologies and tools which improve, measure, and assess the quality of the translations provided by humans and machines alike will continue expanding their reach and sophistication, although it is still not clear whether these new capabilities will have more positive or

negative effects for the various actors involved in the translation process. On the one hand, when applied to quality assurance workflows for human translations, machine learning tools seem to be very valuable for translators and revisers, especially if the claims regarding the ability of AI-based tools to learn and detect more complex translation errors than the traditional quality control tools can be validated through further research studies. On the other hand, when applied to quality assessment workflows in conjunction with standardised quality metrics, AI-based technologies can take care of collecting different sets of data and present them in a clear and structured way, allowing evaluators, project managers, and quality managers to make informed and data-driven decisions. However, even if technological innovation seems to be a key factor to achieve optimal performance in translation and quality management workflows, it is also true that these innovations could have a negative impact in the professional landscape, were they to be used as a "lever" to tighten the industry requirements in terms of quality (in the case of translators), and productivity (in the case of revisers and evaluators). While it might still be too soon to draw conclusions in this respect, there is no doubt that this topic will be very relevant in the years to come. And, as authors like Moorkens (2017) rightly point out, there is certainly scope for the language experts to take an "advisory role" as to what technology, tool, or workflow to choose according to the needs of a specific project or client to plan, estimate, control, evaluate and deliver the quality desired.

5 Conclusions

With new developments and opportunities come new challenges that require innovative solutions and approaches to human and machine evaluation methodologies. One of the clearest examples of this conundrum of opportunities vs challenges is the improved efficiency and quality of NMT engines. As these engines continue to increase their performance year after year, the evaluation of the quality provided becomes more complex: human evaluation is still considered the "gold standard" to measure the final quality of the translation, but it presents some limitations that are particularly problematic when it comes to the evaluation of MT engines, such as the relatively low agreement rates among different evaluators in comparison with the automatic evaluation (as shown by the different publications on this topic released in connection with the WMT Shared Tasks), the amount of time required to conduct evaluations of big sets of data, the long-standing debate of how to evaluate the quality and skillsets of the evaluators themselves, how to assess human-machine parity in language translation (Läubli et al., 2020), as well as the standardisation of the characteristics of each type of evaluation and the required skills and to perform translation quality assessment, as stated by academics such as Doherty, Moorkens or Gaspari (Doherty et al., 2018).

A few examples have also been reviewed of the applications of AI and machine learning to improve the efficiency and objectivity of the quality assessments performed: from the automatic evaluations to assess the quality of MT outputs (with the consequent reduction in the amount of time required to conduct evaluations of big sets of data and the increase in the objectivity and homogenisation of the scores), to technologies such

as ILVS, which can be integrated with translation management systems and quality modules, and support translators and revisers to automatise certain quality assurance tasks.

With the notable exception of the WMT conference and its Shared Tasks, academia and industry tend to work in their own silos. However, there is a very promising trend that could help to change this disconnection, as some big corporations of the likes of Microsoft, Google, Facebook, Adobe, and Salesforce, invest large sums in their private research teams and publish their results in open repositories, which can be accessed by other research groups and academics. There are also quite exciting new academic research venues in other areas: on the one hand, some technologies (such as ILVS) have been developed entirely by medium-sized companies with more limited research resources and are still in their early stages, which makes them ideal candidates to establish projects in collaboration with other research groups (public and private) that might be interested in advancing the research and test the capabilities of this new technology in different quality assessment tasks.

Bibliography

ASTM. (2006). ASTM F2575-06: Standard Guide for Quality Assurance in Translation. *Annual Book of ASTM Standards,* June. <https://doi.org/10.1520/F2575-14>. [Accessed: 20211110].

ASTM. (2021). WK46396: *Standard Practice for Analytic Translation Quality Evaluation.* Unpublished manuscript.

Birch, A.; Abend, O., Bojar, O.; Haddow, B. (2016). Hume: Human UCCA-based evaluation of machine translation. In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing.* Austin, Texas: Association for Computacional Linguistics, pp. 1264-1274. <https://doi.org/10.18653/v1/d16-1134>. [Accessed: 20211116].

Bojar, O.; Federmann, C.; Haddow, B.; Koehn, P.; Post, M; Specia, L. (2016). Ten Years of WMT Evaluation Campaigns: Lessons Learnt. In: *Proceedings of the LREC 2016 Workshop "Translation Evaluation: From Fragmented Tools and Data Sets to an Integrated Ecosystem,* pp. 27-34. <http://www.cracking-the-language-barrier.eu/wp-content/uploads/Bojar-Federmann-etal.pdf>. [Accessed: 20211116].

Castilho, S.; Doherty, S.; Gaspari, F.; Moorkens, J. (2018). Approaches to Human and Machine Translation Quality Assessment. In: Moorkens J.; Castilho S.; Gaspari F.; Doherty S. (eds). *Translation Quality Assessment: from principles to practice*. Cham: Springer. <https://doi.org/10.1007/978-3-319-91241-7_2>. [Accessed: 20211116].

Colina, S. (2008). Translation quality evaluation: empirical evidence for a functionalist approach. *The Translator,* v. 14, n. 1, pp. 97–134. <https://doi.org/10.1080/13556509.2008.10799251>. [Accessed: 20211115].

Deutsches Forschungszentrum für Künstliche Intelligenz, DFKI (2015). *Multidimensional Quality Metrics Definition.* <http://www.qt21.eu/mqm-definition/definition-2015-06-16.html#dqf-mapping>. [Accessed: 20210613].

Devlin, J.; Chang, M. W.; Lee, K.; Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In: *NAACL HLT 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Proceedings of the Conference.* <https://arxiv.org/abs/1810.04805>. [Accessed: 20211115].

Dickson, B. (2019). *What is artificial narrow intelligence (Narrow AI)?* TechTalks. <https://bdtechtalks.com/2020/04/09/what-is-narrow-artificial-intelligence-ani/>. [Accessed: 20210613].

Doherty, S. (2016). The impact of translation technologies on the process and product of translation. *International Journal of Communication,* v. 10, pp. 947–969. <https://ijoc.org/index.php/ijoc/article/view/3499>. [Accessed: 20210119].

Doherty, S.; Moorkens, J.; Gaspari, F.; Castilho, S. (2018). On Education and Training in Translation Quality Assessment. In: Moorkens, J.; Castilho, S.; Gaspari, F.; Doherty, S. (eds). *Translation Quality Assessment: from principles to practice.* Cham: Springer, pp. 95-106. <https://doi.org/10.1007/978-3-319-91241-7_5>. [Accessed: 20211116].

Esselink, B. (2019). Multinational language service provider as user. In: M. O'Hagan (ed.). *The Routledge Handbook of Translation and Technology.* Abingdon, Oxon [etc.]: Routledge, pp. 109-126. <https://doi.org/10.4324/9781315311258-7>. [Accessed: 20211116].

Freitag, M.; Foster, G.; Grangier, D.; Ratnakar, V.; Tan, Q.; Macherey, W. (2021). E*xperts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation.* <http://arxiv.org/abs/2104.14478>. [Accessed: 20210604].

Gekhman, Z.; Aharoni, R.; Beryozkin, G.; Freitag, M.; Macherey, W. (2020). KoBE: Knowledge-Based Machine Translation Evaluation. In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.findings-emnlp.287>. [Accessed: 20201115].

Görög, A. (2017). The 8 most used standards and metrics for Translation Quality Evaluation. In: *TAUS Blog*. <https://blog.taus.net/the-8-most-used-standards-and-metrics-for-translation-quality-evaluation>. [Accessed: 20210119].

Grave, E.; Bojanowski, P.; Gupta, P.; Joulin, A.; & Mikolov, T. (2019). Learning word vectors for 157 languages. In: *LREC 2018: 1th International Conference on Language Resources and Evaluation*, pp. 3483–3487. <https://www.aclweb.org/anthology/L18-1550.pdf>. [Accessed: 20201115].

ISO. (2015a). *ISO 17100:2015 (en) Translation services: Requirements for translation services.* <https://www.iso.org/obp/ui/#iso:std:iso:17100:ed-1:v1:en>. [Accessed: 20211116].

ISO. (2015b). *ISO 9000:2015 (en) Quality management systems: Fundamentals and vocabulary*. <https://www.iso.org/obp/ui/#iso:std:iso:9000:ed-4:v1:en>. [Accessed: 20211116].

Jaworski, R. (2020a). Unlocking the Secrets of Language AI With Inter-language Vector Space. *Slator news* <https://slator.com/sponsored-content/unlocking-secrets-language-ai-inter-language-vector-space/>. [Accessed: 20211116].

Jaworski, R. (2020b). Assessing Cross-lingual Word Similarities Using Neural Networks Main challenges. *Translating and the Computer TC42*. Unpublished manuscript.

Jiménez-Crespo, M. A. (2009). The evaluation of pragmatic and functionalist aspects in localization: towards a holistic approach to Quality Assurance. *The Journal of Internationalization and Localization,* v. 1, n. 1, pp. 60–93. <https://doi.org/10.1075/jial.1.03jim>. [Accessed: 20211116].

Läubli, S.; Castilho, S.; Neubig, G.; Sennrich, R.; Shen, Q.; Toral, A. (2020). A set of recommendations for assessing human-machine parity in language translation. *JAIR, Journal of Artificial Intelligence Research,* v. 67, pp. 653–672). <https://doi.org/10.1613/JAIR.1.11371>. [Accessed: 20211116].

Lavie, A.; Agarwal, A. (2007). METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In: *Proceedings of the Second Workshop on Statistical Machine Translation, June*. Association for Computational Linguistics, pp. 228-231. <https://aclanthology.org/W07-0734/>. [Accessed: 20210613].

Lommel, A.; Melby, A. K. (2018). MQM-DQF: *A Good Marriage (Translation Quality for the 21st Century).* <https://www.aclweb.org/anthology/W18-1925.pdf>. [Accessed: 20210503].

Marr, B. (2020). What Is The Artificial Intelligence Revolution And Why Does It Matter To Your Business? *Forbes.* <https://www.forbes.com/sites/bernardmarr/2020/08/10/what-is-the-artificial-intelligence-revolution-and-why-does-it-matter-to-your-business/>. [Accessed: 20211112].

Martínez, Roberto (2014). A deeper look into metrics for translation quality assessment (TQA): A case study. *Miscelanea,* v. 49, pp. 73–94. <https://www.miscelaneajournal.net/index.php/misc/article/view/170>. [Accessed: 20211116].

Massardo, I. (2018). *Business Analytics for Translation: Client Ratings.* Wordbee. <https://www.wordbee.com/blog/wordbee-features/business-analytics-for-translation-and-localization-client-ratings/>. [Accessed: 20211115].

Memsource. (2018). Machine Translation Quality Estimation: Memsource's Latest AI-powered Feature. *Memsource blog.* <https://www.memsource.com/blog/machine-translation-quality-estimation-memsources-latest-ai-powered-feature/>. [Accessed: 20210622].

Memsource. (2020). *MT Quality Estimation*. Memsource Help Center. <https://help.memsource.com/hc/en-us/articles/360012527380>. [Accessed: 20210622].

Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. (2013). Efficient estimation of word representations in vector space. In: *1st International Conference on Learning Representations, ICLR 2013: Workshop Track Proceedings*. <https://arxiv.org/abs/1301.3781>. [Accessed: 20211108].

Moorkens, J. (2017). Under pressure: translation in times of austerity. *Perspectives: Studies in Translation Theory and Practice,* v. 25, n. 3, pp. 464–477. <https://doi.org/10.1080/0907676X.2017.1285331>. [Accessed: 20210507].

Muzii, L. (2014). The red-pen syndrome. *Tradumàtica, Tecnologies de la Traducció,* n. 12, pp. 421-429. <https://doi.org/10.5565/rev/tradumatica.68>. [Accessed: 20211116].

Németh, G. D. (2019). Machine Translation: Compare to SOTA. *Towards Data Science.* <https://towardsdatascience.com/machine-translation-compare-to-sota-6f71cb2cd784>. [Accessed: 20201108].

O'Brien, S. (2012). Towards a dynamic quality evaluation model for translation. *Jostrans, The Journal of Specialised Translation,* n. 17, pp. 55-77. <https://www.jostrans.org/issue17/art_obrien.pdf>. [Accessed: 20211116].

Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. (2001). BLEU: a method for automatic evaluation of machine translation. In: *Proceedings of the 40th Annual Meeting of the Association for computational Linguistics,* pp. 311-318. <https://aclanthology.org/P02-1040/>. [Accessed: 20211116].

Petrova, V. (2019). Translation Quality Assessment Tools and Processes in Relation to CAT Tools. In: *Proceedings of the Human-Informed Translation and Interpreting Technology Workshop (HiT-IT) 2019,* pp. 89–97. <https://doi.org/10.26615/issn.2683-0078.2019_011>. [Accessed: 20210329].

Pielmeier, H. (2017). Using Big Data to Save Money on Translations. *Our Analysts' Insights Blog.* CSA Research. <https://csa-research.com/Insights/ArticleID/167/Using-Big-Data-to-Save-Money-on-Translations>. [Accessed: 20211119].

Reddy, B. P.; Bhusan, S.; Sarkar, S.; Mukherjee, A. (2020). NwQM: A neural quality assessment framework for Wikipedia. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP*). Association for computational Linguistics, pp. 8396-8406. <https://doi.org/10.18653/v1/2020.emnlp-main.674>. [Accessed: 20201115].

Rei, R.; Stewart, C.; Farinha, A. C.; Lavie, A. (2020). COMET: A Neural Framework for MT Evaluation. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP*). Association for computational Linguistics, pp. 2685–2702. <https://doi.org/10.18653/v1/2020.emnlp-main.213>. [Accessed: 20210622].

SAE International. (2005). *SAE J2450 Translation Quality Metric Task Force.* <https://www.sae.org/standardsdev/j2450p1.htm>. [Accessed: 20210622].

Sargent, B. B.; DePalma, D. A.; Toon, A. (2019). *MarketFlex for Language-Oriented TMS: Systems for Translating Multiple Content Streams.* CSA Research. <https://insights.csa-research.com/reportaction/305013021/Marketing>. [Accessed: 20210507].

Schwab, K. (2016). *The Fourth Industrial Revolution: what it means and how to respond.* World Economic Forum. <https://www.weforum.org/agenda/2016/01/the-fourth-industrial-revolution-what-it-means-and-how-to-respond/>. [Accessed: 20201116].

Specia, L.; Raj, D.; Turchi, M. (2010). Machine translation evaluation versus quality estimation. *Machine Translation,* v. 24, n. 1, pp. 39–50. <https://doi.org/10.1007/s10590-010-9077-2>. [Accessed: 20210621].

Sun, S.; Guzmán, F.; Specia, L. (2020). Are we Estimating or Guesstimating Translation Quality? In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics of the Association for Computational Linguistics.* Association for computational Lingustics, pp. 6262–6267. <https://doi.org/10.18653/v1/2020.acl-main.558>. [Accessed: 20201227].

Vandenberg, R. (2018). How AI is changing the future of translation management. *Multilingual,* v. 29, n. 8, pp. 32–35. <https://multilingual.com/articles/how-ai-is-changing-the-future-of-translation-management/>. [Accessed: 20211116].

Way, A. (2018). Quality Expectations of Machine Translation. In: Translation Quality Assessment. In: Moorkens, J.; Castilho, S.; Gaspari, F.; Doherty, S. (eds). *Translation Quality Assessment: from principles to practice.* Cham: Springer, pp. 159-178. <https://doi.org/10.1007/978-3-319-91241-7_8>. [Accessed: 20210613].

Ziganshina, L. E.; Yudina, E. V., Gabdrakhmanov, A. I.; Ried, J. (2021). Assessing Human Post-Editing Efforts to Compare the Performance of Three Machine Translation Engines for English to Russian Translation of Cochrane Plain Language Health Information: Results of a Randomised Comparison. *Informatics,* v. 8, n. 1. <https://doi.org/10.3390/informatics8010009>. [Accessed: 20210622].