Natural Language Engineering (2020), **26**, pp. 137–161
doi:10.1017/S1351324919000469

CAMBRIDGE UNIVERSITY PRESS

SURVEY PAPER

How to evaluate machine translation: A review of automated and human metrics

Eirini Chatzikoumi*

Instituto de Literatura y Ciencias del Lenguaje, Pontificia Universidad Católica de Valparaíso, Av. El Bosque 1290, Viña del Mar, Chile
*Corresponding author. Emails: chatzikoumi@gmail.com, eirini.chatzikoumi@mail.pucv.cl

(Received 30 November 2017; revised 5 August 2019; accepted 5 August 2019; first published online 11 September 2019)

Abstract

This article presents the most up-to-date, influential automated, semiautomated and human metrics used to evaluate the quality of machine translation (MT) output and provides the necessary background for MT evaluation projects. Evaluation is, as repeatedly admitted, highly relevant for the improvement of MT. This article is divided into three parts: the first one is dedicated to automated metrics; the second, to human metrics; and the last, to the challenges posed by neural machine translation (NMT) regarding its evaluation. The first part includes reference translation–based metrics; confidence or quality estimation (QE) metrics, which are used as alternatives for quality assessment; and diagnostic evaluation based on linguistic checkpoints. Human evaluation metrics are classified according to the criterion of whether human judges directly express a so-called subjective evaluation judgment, such as 'good' or 'better than', or not, as is the case in error classification. The former methods are based on directly expressed judgment (DEJ); therefore, they are called 'DEJ-based evaluation methods', while the latter are called 'non-DEJ-based evaluation methods'. In the DEJ-based evaluation section, tasks such as fluency and adequacy annotation, ranking and direct assessment (DA) are presented, whereas in the non-DEJ-based evaluation section, tasks such as error classification and postediting are detailed, with definitions and guidelines, thus rendering this article a useful guide for evaluation projects. Following the detailed presentation of the previously mentioned metrics, the specificities of NMT are set forth along with suggestions for its evaluation, according to the latest studies. As human translators are the most adequate judges of the quality of a translation, emphasis is placed on the human metrics seen from a translator-judge perspective to provide useful methodology tools for interdisciplinary research groups that evaluate MT systems.

Keywords: Machine translation; Machine translation evaluation; Human metrics; Automated metrics; Machine translation quality

1. Introduction

This article is a review designed to be used in machine translation (MT) evaluation projects by interdisciplinary teams made up of MT developers, linguists and translators. The crucial importance of the MT evaluation has been highlighted by a series of researchers (Zhou *et al.* 2008; Gonzàlez and Giménez 2014; Graham *et al.* 2015; Bentivogli *et al.* 2018), as it is used not only to compare different systems but also to identify a system's weaknesses and refine it (Gonzàlez and Giménez 2014). The latest paradigm in the field, neural machine translation (NMT), has brought about a radical improvement in the MT quality (Hassan *et al.* 2018) but poses new challenges to evaluation, which is still of growing importance 'due to its potential to reduce post-editing human effort in disruptive ways' (Martins *et al.* 2017).

The aim of this article is to offer a compact presentation of an array of evaluation methods, including information on their implementations, advantages and disadvantages, from the translator–evaluator perspective. This lattermost perspective is what has been conspicuously absent in the few existing works exclusively devoted to the MT evaluation (Euromatrix 2007; Han 2018). On the one hand, the Euromatrix (2007) survey provides a thorough review of automated evaluation up to its year of publication but is less detailed as to human evaluation. On the other hand, the survey article by Han (2018) provides a balanced review of human and automated metrics; however, our work consists of a more detailed survey based on a different classification, which aims at a theory of the MT evaluation as suggested in Euromatrix (2007); it presents information and recommendations on the implementation of different metrics from the translator–evaluator perspective and introduces the recent challenges posed by the neural paradigm in MT and their impact in the field of evaluation.

To set up a solid-quality evaluation project, quality must be defined, which is done in Section 2 of this article. In Section 3, the classification of methods is introduced, that is the way in which the different evaluation methods are discussed in the rest of the article. This classification is based on already-existing as well as newly coined categories, where the main dichotomy is between automated and human methods. The former are examined in Section 4 and the latter in Section 5. Section 6 provides a glimpse of the challenges posed by NMT systems in the field of evaluation. Finally, conclusions are presented in Section 7, along with specific recommendations as derived from the review of the existing evaluation metrics.

2. MT quality

The evaluation of MT systems is highly important, since its results show the degree of output reliability and are exploited for system improvements (Dorr, Snover and Madnani 2011). In this article, the evaluation metrics reviewed are those related to the MT output; other parameters, such as the speed or usability, have been left out.

Before discussing the MT quality, a brief review of translation quality definitions is necessary. A succinct definition, which is also widely used in the MT field, comes from Koby *et al.* (2014): 'A quality translation demonstrates accuracy and fluency required for the audience and purpose and complies with all other specifications negotiated between the requester and provider, taking into account end-user needs'. Of course, in the case of MT, the end user can vary from an occasional Internet user to a professional translator (Specia, Raj and Turchi 2010), and, therefore, the desirable result shall also vary from gisting to human-like translation. As far as the purpose is concerned, according to Skopos theory (Nord 1997), translation is an act with a specific purpose, the result of which is a target text. The purpose that governs the translation process is defined by the requester, and the translation should work in such a way that the purpose is fulfilled. Finally, according to House (2014), 'an adequate translation text is a pragmatically and semantically equivalent one', highlighting the importance of equivalence between source and target text functions. The core criteria can, therefore, be summed up as (i) fluency in the target language, which includes grammaticality and naturalness; (ii) adequacy as in semantic and pragmatic equivalence between the source and the target text; and (iii) compliance with possible requester specifications.

As far as the MT quality is concerned, it is roughly guided by the same definitions as those for human translation quality; indeed, its uppermost aim is to reach a human-like level (Papineni *et al.* 2002). Recent advances in the field have brought about the issue of human parity of MT, for which Hassan *et al.* (2018) use the following statistical 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'. From this point of view, the emphasis is put on the statistical indistinguishability between machines and humans, which does not necessarily imply

https://doi.org/10.1017/S1351324919000469 Published online by Cambridge University Press

equivalence (Läubli, Sennrich and Volk 2018) or the qualitative indistinguishability in the terms of an imitation game (Turing 1950). It should also be remarked that the conclusions of Hassan *et al.* (2018) have been refuted by Toral *et al.* (2018) and that human parity has been questioned regarding document-level evaluation by Läubli *et al.* (2018).

3. Classification of metrics

During recent years, many evaluation techniques have been developed, which have been used for all types of MT systems (rule-based, statistical, hybrid and neural). In international literature, evaluation techniques are classified either as automated or as human (or manual) metrics. These two categories seem distinguishable enough, almost self-defined, and any reader can easily perceive that the first one refers to an evaluation which is performed automatically, that is by a machine, while the latter refers to an evaluation performed by humans. However, in practice, these two categories are not that distinguishable. On the one hand, automated evaluation has, up to now, used either human translations or human annotations, and what is automatically performed is a calculation; on the other hand, human evaluation makes use of several computational tools and automated processes. In between these two categories are the so-called semiautomated metrics, which, in this article, are included in the human evaluation category.

The metrics reviewed are presented on the basis of the two aforementioned categories: automated and human evaluation. In the category of automated evaluation, the techniques presented are those in which human intervention is limited to the development of the evaluation system itself, not taking place during the evaluation process. Likewise, in the category of human evaluation, the techniques reviewed are those in which humans intervene manually during the evaluation phase, regardless of possible interventions in previous phases; for this reason, semiautomated metrics are also included in this category. Automated versions of human metrics are presented in the same section as their human counterparts for coherence purposes.

The human evaluation category is further divided into two subcategories based on the following criterion: whether directly expressed judgment (DEJ) is used or not. These two subcategories are thus called 'DEJ-based' and 'non-DEJ-based'. This distinction is due to the substantial differences between methods in which, for instance, humans state that a translation or the language in a translation is good, fair or bad and methods in which humans are asked to classify and/or correct errors or to complete a gap-filling task based on the comprehension of a machine-translated text. Tasks such as gap filling and error analysis do not imply a direct judgment. In fact, the only element in correcting—or better still, in requesting a correction that involves judgment—is the insinuation that the translation is not perfect. In the field of translation, such imperfection is expected. According to Ricoeur (2003), there is no perfect translation, and even Newmark (1998), when defining translation, states that each act of translation involves some loss of meaning, a loss which 'is on a continuum between over-translation and under-translation'. Moreover, there are a variety of critical comments against translation, such as the ones attributed to Widmer by Newmark (1998), according to whom the quality of many published translations is appalling, and mistake-free translations are rare. However, this 'handicap' has never inhibited the practice of translation. As Ricoeur (2003) concludes, it is the actual practice of translation, which has been around for quite some centuries now, that comes to prove that translation can indeed be performed (no matter how imperfectly).

DEJ-based methods can be considered prone to a higher degree of subjectivity than non-DEJ-based metrics. This is because, even if there are guidelines to follow, there is more room for subjectivity when one is asked to assess the quality of a text, as it will also depend on the degree of (linguistic) lenience of the judge, in contrast to being asked to classify errors according to a taxonomy or answer questions on the content of a text. DEJ-based metrics are also more sensitive to the drawbacks of indirect comparison: determining the degree of fluency of one segment can be

https://doi.org/10.1017/S1351324919000469 Published online by Cambridge University Press

highly influenced by the fluency of the previous segment(s). It is plausible to think that the cognitive processing involved in the two types of evaluation is quite different: in DEJ-based metrics, the judges have to make an assessment, while, in non-DEJ-based metrics, the process is much more task oriented (classification, postediting, question answering, gap filling, etc.). We trust that this distinction can contribute to the theoretical discussion at the background of the evaluation and that it has a practical impact in the implementation of the metrics.

4. Automated metrics for evaluation of MT quality

Under the label of the automated evaluation of MT, we classify the systems which score MT outputs without any human involvement; human involvement in automated metrics takes place in the set-up of the task, for example data collection, annotations or reference translations production.

Nowadays, there are three types of automated evaluation: (i) metrics that yield a score for the MT output based on the degree of similarity to reference translations; (ii) confidence or quality estimation (QE) metrics, that is systems that classify the MT output by quality levels, which are not evaluation metrics per se but are considered as proxies for them (Specia *et al.* 2009); and (iii) diagnostic evaluation based on checkpoints.

Just like any other system, automated evaluation systems must be evaluated with specific criteria. Such criteria, though first developed for reference translation–based metrics, are also used for QE metrics. According to Banerjee and Lavie (2005), a satisfactory automated evaluation system should meet the following conditions: (i) high correlation with human judgments quantified in relation to translation quality, (ii) sensitivity to nuances in quality among systems or outputs of the same system in different stages of its development, (iii) result consistency (similar results for similar texts translated by the same system), (iv) reliability (assured correspondence between evaluation scores and performance), (v) a great range of fields and (vi) speed and usability. These conditions are complemented by Koehn's (2010) suggestions with some overlap: (i) low cost, (ii) possibility of direct system performance optimisation in regard to a given metric, (iii) possibility of intuitive interpretation of the scores, (iv) consistency and (v) correctness. Correlation with human judgment is considered the most important criterion (Specia *et al.* 2010).

4.1 Reference translation–based metrics

These metrics yield a score to the MT output, based on the degree of similarity to reference translations, that is quality human translations (Papineni *et al.* 2002), also called gold translations. To calculate the score, a series, or a combination, of techniques are used. Several implementations have been developed based on each technique, which, in this article, are classified generically with references to the more impactful implementations in international literature.

4.1.1 Edit distance

One of the automated evaluation techniques used is based on the edit distance – to be more specific, Levenshtein's distance (Levenshtein 1966). The edit distance between string *a* and string *b* is the minimum number of operations needed for converting string *a* to string *b* (Navarro 2001). Edit operations are insertion, elimination and substitution of a character by another one (Levenshtein 1966). In the use of Levenshtein's distance in translation evaluation, edit operations are the steps that lead us from the MT output to the reference translation, and they do not concern characters but words (Euromatrix 2007).

Such implementations in the field of MT evaluation are the Word Error Rate (WER) (Niessen *et al.* 2000) and its variations and extensions. Word Error Rate is the ratio of the sum of the edit

https://doi.org/10.1017/S1351324919000469 Published online by Cambridge University Press

operations in an MT output to the number of words in the reference translation (Niessen *et al.* 2000).

Some of the variations and extensions of WER are WERg (Blatz *et al.* 2004); Translation Edit Rate (TER) (Snover *et al.* 2006); Multiple Reference WER (MWER) (Niessen *et al.* 2000); inversion WER (invWER) (Leusch, Ueffing, and Ney 2003); sentence error rate (SER) (Tomás, Mas and Casacuberta 2003); cover disjoint error rate (CDER) (Leusch, Ueffing, and Ney 2006); and hybrid TER (HyTER) (Dreyer and Marcu 2012). It bears mentioning that WER has also been used for several grammatical categories (Popović and Ney 2007). The differences among these variations and extensions are mostly to do with whether they use one or more reference translations and whether they also consider the movements of words and phrases as an edit operation or not.

4.1.2 Precision and recall

Precision and recall make up another set of widely used metrics in automated measurement techniques. In the case of translation, precision is the ratio between acceptable *n*-grams in the MT output (i.e. the *n*-grams also found in at least one of the reference translations) to the number of *n*-grams in the same MT output. In this sense, in a 10 one-gram sentence, that is a sentence that consists of 10 words, and in which six are acceptable, precision is 6/10. Recall is the ratio of acceptable *n*-grams in the MT output (i.e. the *n*-grams also found in at least one of the reference translations) to the number of *n*-grams of the reference translation (the ideal number of *n*-grams). In the first case, the percentage of the correct words in the translation is calculated, and, in the latter, how many of the ideal words in the translation. In the previous example of 6/10 precision, if in the reference translation there are 15 one-grams (i.e. the MT output is quite shorter than the ideal translation), the recall would be 6/15.

Among precision-based implementations of these techniques in the MT evaluation, the most widely used metric is the bilingual Evaluation Understudy (BLEU) (Papineni *et al.* 2002). It was first introduced in 2001 (Euromatrix 2007) and was suggested to meet the needs for fast and low-cost MT evaluation for any given language pair, both for general use and for system tuning. With this metric the MT output is evaluated as to adequacy and fluency by comparing it with reference translations (Papineni *et al.* 2002). As the MT quality is judged according to its closeness to the human translation, this degree of closeness is calculated, and the greater it is the better the MT output is considered. Closeness between the MT output and reference translations is expressed in a 0–1 scale, 0 being the minimum score (Papineni *et al.* 2002). Thus, to perform this comparison, two basic components are needed: the algorithm with which closeness is computed and reference translations.

For the calculation of the degree of closeness, *n*-grams are compared, precisely 1–4 grams, as 4 is considered the number which yields the greatest correlation with monolingual judges' evaluation (Papineni *et al.* 2002). In this metric, modified *n*-gram precision (also called 'clipped precision') is used, to ensure that high scores are avoided in outputs with too many occurrences of the same word that is present in the reference translations. For example, in the MT output *the the the the the the the*, if there is a reference translation *the cat is on the mat* and if standard precision is computed, the 1-gram *the* is present in the reference translation and, therefore, is correct, so the number of correct 1-grams is 7, while the total number of 1-grams is also 7, yielding a 7/7 precision score, which is not enough for the purposes of this kind of evaluation (Papineni *et al.* 2002). Modified precision is the ratio of the maximum number of *n*-grams in the MT output present in any one of the reference translations to the total number of *n*-grams of the MT output; that is instead of counting the number of 'correct' *n*-grams, one counts the maximum number of occurrences of these *n*-grams in any one of the reference translations. This way, in the example of the MT output *the the the the the the the* with a reference translation *the cat is on the mat*, the maximum number of occurrences in a reference translation of the *n*-gram *the* is 2, so the modified precision is 2/7 (Papineni *et al.* 2002).

https://doi.org/10.1017/S1351324919000469 Published online by Cambridge University Press

What is not directly covered by this technique is recall, since in translation there are more than one right answers. This means that an output which is too short in relation to the reference translations is not penalised. The lack of recall is compensated by the brevity penalty, with which precision results are multiplied. To compute this penalty, the best match length is detected in the reference translations (the length of the reference sentence that is closest to the MT sentence length), and the sum of the best match lengths is calculated for every sentence of the MT output (which yields the reference length). When the total length of the MT output is longer than the reference length, the penalty is equal to 1, while when it is smaller than or equal to the reference length, the penalty is a decaying exponential and is multiplied by the geometric mean of the scores of the *n*-grams; the product of this multiplication is the final score (Papineni *et al.* 2002).

Table 1 shows examples of BLEU scores of MT outputs with three reference translations, accompanied by short comments.

A close observation of specific segments, their scores and their reference segments highlights the importance of the following factors: the quality of the alignment; the freedom in some human translations, which also allows for splitting or merging sentences; and the number of reference translations used [the more the reference translations, the higher the score (Papineni *et al.* 2002)], apart from the well-known lack of the use of synonyms and constituents' order.

A very similar metric, which was actually developed as a variation of BLEU, is the one created by the US Department of Commerce, National Institute of Standards and Technology (NIST) (Doddington 2002). Its main difference is that it gives more weight to more informative *n*-grams (Euromatrix 2007)[a].

A combination of precision and recall is used in the F-measure metric (Melamed, Green and Turian 2003); the Character n-gram F-score (CHRF) score,[b] which stands for F-score based on character *n*-grams (Popović 2015); the Recall-Oriented Understudy for Gisting Evaluation (ROUGE) implementations[c] (Lin and Och 2004), which are adaptations of BLEU for the evaluation of automatic summarisation; the General Text Matcher metrics (GTM)[d] (Melamed *et al.* 2003); weighted *n*-gram model (WNM) (Babych and Hartley 2004); Metric for Evaluation of Translation with Explicit Ordering (METEOR)[e] (Banerjee and Lavie 2005); ParaEval (Zhou, Lin and Hovy 2016); the Broad Learning and Adaptation for Numeric Criteria (BLANC) metrics family (Lita, Rogatti and Lavie 2005); and the Length Penalty, Precision, n-gram Position difference Penalty and Recall (LEPOR) metric[f] (Han, Wong, and Chao 2012).

4.1.3 Advantages of reference translation metrics

The advantages of reference translation metrics, in comparison to human evaluation, are their speed, low cost, fewer human resources and reusability (Banerjee and Lavie 2005; Lavie 2011). Moreover, they do not require bilingual judges (Banerjee and Lavie 2005; Lavie 2011). Nonetheless, they do require translators, a fact which, although rarely mentioned, does not allow the full automation of the process and cost and time minimisation. The most important advantage is the metrics' reusability during the development of MT systems, which allows for modifications, improvements and re-evaluation (Banerjee and Lavie 2005; Koehn 2010; Lavie 2011). Finally, the argument of objectivity is widely used, to the extent that this class of evaluations is referred to as 'objective evaluation' metrics in contrast to human ('subjective') evaluation (Euromatrix 2007). Nonetheless, we consider that there is a confusion between objectivity and consistency of a system in comparison to a person. One cannot assure that a person will give the same score to the

[a]The script for BLEU and NIST evaluation is available at https://www.nist.gov/itl/iad/mig/tools (last access 09/02/2019).
[b]Available at https://github.com/m-popovic/chrF (last access 05/08/2019).
[c]Available at https://github.com/ng-j-p/rouge-we (last access 05/08/2019).
[d]Available at https://nlp.cs.nyu.edu/GTM/ (last access 05/08/2019).
[e]Available at https://www.cs.cmu.edu/∼alavie/METEOR/README.html (last access 05/08/2019).
[f]Available at https://github.com/aaronlifenghan/aaron-project-lepor (last access 05/08/2019).

https://doi.org/10.1017/S1351324919000469 Published online by Cambridge University Press

Table 1. BLEU examples

	0.7825 score
MT	The eggs are fertilized in the laboratory and transferred into the woman's uterus
Reference 1	Eggs are fertilised in the laboratory and transferred to the woman's uterus
Reference 2	The eggs are fertilised in the laboratory and then transferred into the woman's uterus
Reference 3	The eggs are fertilised in the laboratory and transferred into the woman's uterus
A high score for a segment which is very similar to the reference translations and can be considered an adequate and fluent translation.	
	0.4792 score
MT	Similarly, in cases where the woman does not produce eggs, can be used eggs from another woman and fertilized with sperm of her partner
Reference 1	Likewise, in cases where the female partner produces no eggs, IVF allows fertilisation of eggs from a female donor with sperm from the male partner
Reference 2	Similarly, in cases when a female partner does not produce eggs, then the eggs of a donor woman may be fertilised by the partner's sperm
Reference 3	Respectively, in cases the woman does not produce eggs, there is the possibility of using eggs of another woman and fertilise them with the partner's sperm
A segment that can easily be fixed. It is comprehensible and accurate though not very fluent.	
	0.3967 score
MT	Moreover, the hospital can become magnetic heart tomography dimensional display for full control of all congenital heart disease
Reference 1	Moreover, the hospital provides cardiac magnetic resonance imaging (MRI) with 3D imaging of the heart for the complete checking of all congenital heart diseases
Reference 2	3D Heart Magnetic Resonance Imaging can also be performed in the hospital to identify congenital heart conditions
Reference 3	At our hospital you may also have a heart MRI with 3D imaging for a full control of all congenital heart diseases
This MT output cannot be understood without turning to the source text or a reference translation, and it needs major modifications to be usable.	
	0.1999 score
MT	So the couple in collaboration with the specialist gynecologist and after completion of all necessary tests and have found that they can not have children, we proceed to the choice of method
Reference 1	A couple shall first contact the doctor and proceed with all necessary tests and examinations. After establishing that natural conception is impossible the couple and the doctor shall proceed with the selection of the appropriate method
Reference 2	When all the necessary tests have been carried out and it has been established that the couple cannot conceive, they proceed to choose the method of assisted reproduction in close cooperation with a specialist gynaecologist
Reference 3	So the couple, in collaboration with the specialised gynaecologist and following all the required examinations showing that they cannot procreate, proceed to choose a method
A quite long sentence with a very low score, which, however, can easily be fixed and presents no comprehension impediments. The low score is due to the lack of similarity with the reference translations, which, in this case, does not accurately reflect the MT quality, especially if one compares it with the previous example (0.3967).	

https://doi.org/10.1017/S1351324919000469 Published online by Cambridge University Press

same text twice; neither can one assure that two or more people will agree on the evaluation. An automated metric, however, will always give the same score to the same text, given that all the evaluation parameters are maintained unaltered. In this regard, the consistency of automated systems is a clear advantage (Euromatrix 2007; Koehn 2010). Nevertheless, it is not a matter of objectivity, since automated metrics use reference translations, which are products of human intellect and humans are subjects. Moreover, the ideal result of 'objective' metrics is precisely the closest one to this subjective product.

4.1.4 Disadvantages of reference translation metrics

As far as the disadvantages of these metrics are concerned, the most common in the literature are the fact that the need for reference translations limits the quantity of data that can be evaluated (Specia *et al.* 2010), the lack of distinction between nuances (Lavie 2011) the lack of reliable segment-level evaluation (Lavie 2011), the difficulty in interpreting the evaluation scores (Koehn 2010), as well as the inability to provide information as to the exact strengths and drawbacks of an MT system (Zhou *et al.* 2008). When reference translations are used – both in automated and in human metrics – MT outputs that are very similar to the reference translation are boosted and not similar MT outputs are penalised even if they are good; this is the so-called reference bias (Bentivogli *et al.* 2018).

Moreover, most of the disadvantages of BLEU are considered disadvantages in other metrics (Callison-Burch, Osborne and Koehn 2006). These include lack of stemming, lemmatisation, synonyms and paraphrase use (Callison-Burch *et al.* 2006; Lavie 2011) – apart from paraphrases used in reference translations (Callison-Burch *et al.* 2006) – the fact that all *n*-grams have the same weight for score calculation, thus treating high- and low-semantic-level lexical units the same way (Doddington 2002; Callison-Burch *et al.* 2006; Lavie 2011); long-distance linguistic relations are not captured as only consecutive grams are used (Zhou *et al.* 2008); no distinction between very low scores in very low quality or very free translations (Coughlin 2003); the fact that fluency is measured merely indirectly by large *n*-grams (Banerjee and Lavie 2005); and low performance in short texts (Euromatrix 2007) and in comparisons between human and MTs (Euromatrix 2007), as well as between statistical and rule-based systems (Coughlin 2003; Euromatrix 2007). It should be noted, however, that some of these weaknesses have been addressed by other metrics, such as the addition of 'information gain' by NIST or of syntactic dependency trees (Amigo *et al.* 2006; Koehn and Monz 2006). As far as neural MT systems are concerned, they put pressure on automated metrics due to their 'surface-matching heuristics that are relatively insensitive to subtle differences' (Isabelle, Cherry and Foster 2017). According to the findings of the experiment carried out by Isabelle *et al.* (2017), NMT errors correspond to subtleties, such as specific cases of agreement features and subjunctive mood triggers, or syntactically flexible idioms. These specific cases fail to be reflected by the BLEU metric, which indeed shows a poor correlation with the challenge-set evaluation performed by Isabelle *et al.* (2017), an evaluation based on especially difficult phenomena for an MT system to handle. Evaluation projects on NMT are further discussed in Section 6.

4.2 Confidence or quality estimation

QE is used to predict 'the quality of a system's output for a given input, without any information about the expected output' (Specia *et al.* 2009). Although QE metrics are not evaluation metrics per se, they are considered as a proxy for them (Specia *et al.* 2009) or an alternative way of assessment (Bojar *et al.* 2016) and are recommended both for the quality evaluation and for the selection of the best among several MT systems (Specia *et al.* 2013), a task for which the reference-based metrics cannot be used. QE metrics do not have the same goal as the reference-based ones, nor do they mean to replace them; what they aim to do is, first, fill the gap created in cases where no reference translations exist and, second, meet the needs of segment-level QE where other metrics have poor results (Specia *et al.* 2010).

https://doi.org/10.1017/S1351324919000469 Published online by Cambridge University Press

QE metrics entered the MT field as binary classification systems (Blatz *et al.* 2004) that could distinguish between 'good' and 'bad' translations, evolving into classifiers of more than two categories (Specia *et al.* 2009, 2013) and also producing continuous ratings in regression settings (Wisniewski, Kumar Singh and Yvon 2012). When first used for MT, they worked on word level (Gandrabur and Foster 2003; Ueffing and Ney 2005), only to later expand their scope to sentences (Quirk 2004). In the first attempts at QE, reference translations were still needed. However, now there are systems that require only manual annotation of translations for already-defined quality levels (Specia *et al.* 2013). To illustrate the kind of features used, the QE platform presented by Specia *et al.* (2013) is briefly described in the following paragraph.

This QE platform consists of two independent modules: a feature extraction module and a machine learning module. The latter uses the features extracted from the source and the target text by the former, to create QE models with the use of regression and classification algorithms. There are three types of features used: (i) complexity, (ii) fluency and (iii) adequacy features. Complexity refers to the complexity of translation, and this type of features includes the number of tokens in the source sentence and the language model probability of the source sentence. Fluency features are extracted from the translations and include the number of tokens in the target sentence, the average number of occurrences of the target word in the target sentence and the language model probability of the target sentence. Finally, the adequacy features are used for computing the degree to which the structure and meaning of the source text is maintained in the translation. They include the ratio of the number of tokens in the source and target text, the ratio of percentages of numbers and content and non-content words in the source and target text, the ratio of various parts of speech in the source and target text, the proportion of dependency relations between aligned constituents in the source and target text, the difference between the numbers of named entities in the source and target text and so on (Specia *et al.* 2013).

Recent improvements have been reported due to the combination of word-level QE and automatic posediting[g] (Martins *et al.* 2017). The QE module of Martins *et al.* (2017) consists of a neural model incorporated in a binary linear classifier. With a word-to-sentence conversion, the system can work on sentence level. The binary labels are obtained automatically by aligning the MT and the postedited sentences, thus avoiding the time-consuming and expensive manual annotation.

4.3 An automated metric repository: Asiya

Asiya[h] is an open toolkit that provides an interface to a collection of both reference-based and QE metrics (Gonzàlez and Giménez 2014). It includes reference-based metrics based on different similarity measures, such as precision, recall and edit rate, as well as metrics operating at lexical, syntactic and semantic dimensions, apart from providing schemes for metric combination and a mechanism to determine optimal metric sets. It is complemented by the Asiya tSearch tool, which can be used for translation error analysis and system comparison. The outputs of the Asiya toolkit are evaluation reports, metric scores and linguistic annotations.

4.4 Diagnostic evaluation based on checkpoints

This type of evaluation is based on linguistically motivated features, such as ambiguous words and noun or prepositional phrases – called 'checkpoints' – which have been predefined and automatically extracted from parallel sentences (Zhou *et al.* 2008). The checkpoints are then used to monitor the translation of important linguistic phenomena and, thus, provide diagnostic evaluation.

[g]Modifications of the MT output so that it can be used; the concept is further discussed in Section 5.2.4.
[h]http://asiya.lsi.upc.edu/ (last access 09/02/2019).

https://doi.org/10.1017/S1351324919000469 Published online by Cambridge University Press

The method proposed by Zhou *et al.* (2008) consists of the following steps: first, creating the checkpoint database by building a corpus of parallel sentences, parsing both source and target sentences, aligning words in sentence pairs, extracting the checkpoints of each category and, finally, determining the references of the checkpoints in the source sentences. Then, the evaluation is performed by selecting from the database the test sentences on the basis of the categories to be evaluated and calculating the number of *n*-grams of the references matched with the MT sentences. The calculation provides the credit of the MT system in the translation of a specific checkpoint and, based on that, the credit of each category and, finally, of the MT system. This method can be implemented in any pair of languages for which there are available word aligners and parsers, the precision of which actually determines the quality of the evaluation.

5. Human evaluation techniques of MT quality

For the purposes of this article, we consider human evaluation of the MT quality as that in which humans intervene at the evaluation stage itself (again, in contrast to automated evaluation where they only intervene in previous stages). Moreover, the distinction between DEJ-based and non-DEJ-based evaluation methods is made.

A highly important factor in the process of human evaluation is the judges, also called annotators, who have to meet certain criteria so that reliability is assured. Depending on the type of evaluation, judges can either be monolingual or bilingual, that is native or near-native speakers of the target language or of both source and target languages. Judge training, evaluation guidelines with examples, as well as the familiarity of the judge with the field to which the texts belong are prerequisites for the evaluation project. Regarding the judges' mother tongues, the usual principle is the same as that applied to translation: one translates into one's mother tongue, by virtue of the Recommendation on the Legal Protection of Translators and Translations and the Practical Means to Improve the Status of Translators, adopted by the General Conference of UNESCO in 1976. This principle has since been questioned during the last decades (Sánchez-Gijón and Torres-Hostench 2014). As far as postediting is concerned, translation principles tend to apply, and relevant research is being carried out (Lacruz, Denkowski and Lavie 2014). The ideal procedure includes more than one judge and an interannotator agreement calculation. In practice, the existence of judges is a thorny issue due to the additional cost incurred, thus being substituted by the researchers participating in the evaluation tasks of the annual MT workshops (Graham *et al.* 2015; Bojar *et al.* 2016) despite the findings that suggest experienced translators should be preferred (Läubli *et al.* 2018). The evaluation conditions are also highly important; they require factors like reasonable text volume and uninterrupted task performance (Przybocki *et al.* 2011).

However, human MT evaluation is not always performed by a small number of judges; it can also be partly or entirely crowdsourced (Graham *et al.* 2015), which considerably reduces the cost, one of the main drawbacks of manual evaluation (Callison-Burch 2009; Graham *et al.* 2015), as detailed in Section 5.5. Crowdsourcing consists of getting a large number of people to perform simple tasks (human intelligence tasks or HITs) that cannot be sufficiently dealt with by computers, for a small sum of money, and this is usually carried out through the website of Amazon's Mechanical Turk (Callison-Burch 2009). Crowdsourcing has been used for a variety of MT-related tasks, such as human-mediated translation edit rate (HTER), reading comprehension tasks, creation of reference translations (Callison-Burch 2009), fluency and adequacy assessments and ranking (Graham *et al.* 2015), which are described in the next sections. In some cases, the interannotator agreement has been very low, but Graham *et al.* (2015) achieve improvements in this direction and conclude that MT systems can be reliably evaluated only by crowdsourcing, as described in Section 5.1.3.

https://doi.org/10.1017/S1351324919000469 Published online by Cambridge University Press

5.1 Evaluation methods based on directly expressed judgment (DEJ-based evaluation methods)

In DEJ-based evaluation methods, judges directly express judgment on the translation quality. Such assessments are usually made on accuracy (also called adequacy in this field) and on fluency and are performed by comparing either the source text with the target text or the target text with a reference translation. As the most common method, there is a series of tools that can be used for its undertaking. It is usually carried out on a five-point scale regarding adequacy and fluency (Callison-Burch *et al*. 2007), although there also other scales such as the four-point scale used by Translation Automation User Society (TAUS)[i] or seven-point ones (Przybocki *et al*. 2009). In terms of accuracy, judges determine how much of the content in the reference translation or in the source text is transmitted, for example everything (5), most of (4), a big part of (3), a small part of (2) or none (1) (Callison-Burch *et al*. 2007). As regards fluency, judges determine whether the language in the MT output is perfect (5), good (4), not natural (3), ungrammatical (2) or unintelligible (1) (Callison-Burch *et al*. 2007). The adequacy and fluency measures have been, however, altogether abandoned in the Workshop on Statistical Machine Translation (WMT) evaluations due to inconsistencies of the five-point scale (Bojar *et al*. 2016). Other methods of DEJ-based evaluation consist of ranking several MT systems [either by ranking sentences or constituents (Bojar 2011)], comparing two MT systems or making a general or specific judgment on the translation, such as a constituent judgment and direct assessment (DA), an improvement of the five-point adequacy and fluency scale that uses an analogue scale, which maps to a 100-point one (Bojar *et al*. 2016). In the next sections, some of the most commonly used tasks in DEJ-based evaluation are presented. Tools that can be used for the performance of the tasks are also presented.

5.1.1 Adequacy and fluency annotation tasks

Dynamic Quality Framework (DQF)[j] is a platform developed by TAUS in 2011, free for academics since 2014, that looks to standardise the evaluation of human and machine translation. Tools, good practices, metrics, reports and data to be used in the translation quality evaluation can be found on this platform. Users fill in the required information (content, purpose, communication channel, etc.) and choose an evaluation task. Texts to be evaluated are uploaded in the form of spreadsheets, and the results are exported in the same form. In the result sheet, source and target sentences can also be consulted.

When using the DQF tool, users choose whether they wish to evaluate adequacy, fluency, or both, which is the most common practice. There is a four-point scale for both adequacy (everything/most/little/none of the content transmitted) and fluency (flawless/good/disfluent/incomprehensible language in the target text). Judges should have very clear criteria about the limits between the four levels and be sufficiently trained so that the results are as reliable as possible.

Another approach to adequacy evaluation is Human UCCA-Based MT Evaluation (HUME), proposed by Birch *et al*. (2016) as a semantic evaluation measure. The Universal Conceptual Cognitive Annotation (UCCA)[k] is a 'cross-linguistically applicable scheme for semantic annotation', developed by (Abend and Rappoport 2013). It requires only a short training and has proven stable across translations (Birch *et al*. 2016). Once the UCCA annotation is performed, the HUME annotation takes place, which consists of going through the annotated semantic units of the source sentence and marking the extent to which its arguments and relations are expressed in the target sentence. The annotator has to decide whether a unit is atomic or structural, that is contains subunits, and mark the atomic ones as correct, partially correct or incorrect and the structural ones as

[i] https://www.taus.net (last access 10/02/2019).
[j] https://www.taus.net/evaluate/about (last access 09/02/2019).
[k] http://vm-05.cs.huji.ac.il/mteval (last access 09/02/2019).

https://doi.org/10.1017/S1351324919000469 Published online by Cambridge University Press

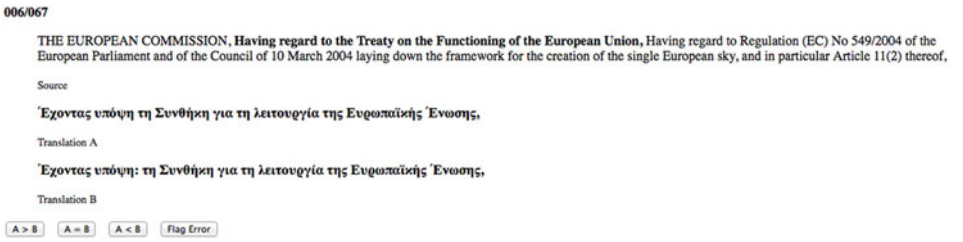

Figure 1. Appraise 3-Way Ranking task.

adequate or bad. Adequacy is evaluated regardless of the fluency, which could actually be flawed in an adequate sentence.

5.1.2 Ranking

Ranking is a comparison task for choosing the best among several systems. Görög (2014) suggests the maximum number of systems to be evaluated should be three, as 'research has shown that an evaluator's ability to make robust judgments is impaired if he or she has to score more than 3 options segment-by-segment'. Nonetheless, in the WMT evaluation campaigns, the standard practice has been five MT outputs at a time, a number found to be 'a good compromise between efficiency and reliability' (Bojar *et al.* 2016). In the DQF environment, there are two kinds of ranking tasks: Quick Comparison, in which the judge chooses the best translation among a maximum number of three systems' outputs, and Rank Translations, in which translated segments are ranked from best (1) to worst (3). Translated segments of each system are not always presented in the same order.

Appraise (Federmann 2010)[1] is an open source tool with which MT evaluation annotation tasks can be performed, with XML as the supported format for importing and exporting files. There are two ranking tasks available in Appraise: 3-Way Ranking and Ranking. In 3-Way Ranking (Figure 1), segments are annotated in pairs of systems by choosing one of the three labels: A > B, A = B and A < B, where A and B stand for the two systems. Users can view the number of the current segment, the original segment in context and the two segments that are being ranked. In the Ranking task, the segments appear in the same way and users can evaluate two or more translations by assigning a value of 1, 2 or 3 to each segment; Appraise allows the same score for two different systems.

A big advantage of sentence ranking is its conceptual simplicity, which makes it easy and straightforward to explain to annotators (Bojar *et al.* 2016). Its main disadvantages include 'the relatively low annotator agreement rates, the immense amount of annotator time required, and the difficulty of scaling the sentence ranking task to many systems' (Bojar *et al.* 2016).

5.1.3 Direct Assessment

DA, one of the most prominent methodologies used nowadays along with postediting (Bentivogli *et al.* 2018), consists of the expression of a judgment of the quality of the MT output in a continuous rating scale (Graham *et al.* 2015), which captures the degree to which one translation is better than another as opposed to ranking interval-level scales (Graham *et al.* 2013). Although DA has been used for both adequacy and fluency, it is now mainly focused on adequacy (Bentivogli *et al.* 2018) and it can be reference-based or source-based. Graham *et al.* (2013; 2015) suggest a method of DA through crowdsourcing, in which adequacy and fluency are assessed on a 100-point scale

[1] https://github.com/cfedermann/Appraise (last access 09/02/2019).

https://doi.org/10.1017/S1351324919000469 Published online by Cambridge University Press

with a moving slider.[m] This kind of selection of a rate allows for fine-grained statistical analysis. As in crowdsourcing, the assessors are not experts; in this project (Graham *et al.* 2015) reliability is assured by the use of quality control items, which intervene between the MT outputs under evaluation and are reference translations, bad reference translations or repeated MT outputs. This yields intraannotator agreement, that is whether assessors are consistent with their previous judgments. Moreover, the approach suggested by Graham *et al.* (2015), apart from removing several sources of bias and assuring interannotator agreement, reduces the cognitive burden by separating the adequacy and the fluency task.

5.1.4 Quality-checking annotation tasks

In this task, also provided by Appraise, the available tags are *acceptable*, *can easily be fixed* and *none of both*. No guidelines are offered regarding the categories and the levels. It should be highlighted that the ease with which a segment can be fixed also depends on factors that affect the degree of translation difficulty of the source text, such as complexity, ambiguity, clarity and/or terminology.

5.2 Evaluation methods not based on directly expressed judgment (non-DEJ-based evaluation methods)

In these techniques, human judgment is only indirectly expressed. This may be by using semiautomated metrics; by performing tasks which require the comprehension of a machine-translated text; or by classifying, analysing and correcting MT outputs. In all cases, annotators use their intellect, but they do not directly express an evaluation judgment on the MT output or the system. There is, however, an evaluation method used at the WMT in 2009 and 2010, which can be considered as a combination of DEJ- and non-DEJ-based metrics: sentence comprehension. It consists of postediting for fluency with no reference translation provided and determining whether the edits performed result in a good translation (Bojar *et al.* 2016).

5.2.1 Semiautomated metrics

Semiautomated metrics, also known as human-in-the-loop evaluation , are variations of automated metrics with the intervention of annotators, such as in the cases of HTER, HBLEU and HMETEOR (Snover *et al.* 2006).

A brief description of HTER follows as an example of semiautomated metrics. The TER automated metric calculates the number of edits that would be required for an MT output to become identical to a reference translation. When using HTER, the annotator also creates a new reference translation by making the least possible edits to either the MT or the existing reference translation. This method has achieved high correlation with human judgments but is also considered not to be indicative of the annotator's effort; indeed, there is no record of edits made and then eliminated due to a changed mind (Lacruz *et al.* 2014).

5.2.2 Task-based evaluation

Another way of evaluating an MT system is by evaluating the efficacy of its output when it comes to performing a particular task. Some examples of such evaluation tasks are the following: (i) asking people to detect the most relevant information in a text; (ii) asking people to answer questions on the text's content (Sanders *et al.* 2011); and (iii) gap filling, that is restoring keywords in reference translations (Ageeva *et al.* 2015). This way, the humans involved indirectly evaluate the degree to which the source text's concepts are expressed in the MT output, without making a judgment on the quality of the output language. This kind of evaluation is mostly used for gisting

[m] The aforementioned Appraise tool also has an implementation for DA (Federmann 2018).

https://doi.org/10.1017/S1351324919000469 Published online by Cambridge University Press

translation, in which only the basic concepts of the source text are mentioned. A translation may receive a very low score in another type of method and a very high one in the task-based, and vice versa (Dorr *et al*. 2011).

5.2.3 Error classification and analysis

A widely used method for human evaluation is error classification, ideally accompanied by error analysis. In this section, the harmonised Multidimensional Quality Metrics (MQM) and DQF error typology is described, followed by a presentation of the Appraise annotation tool typology and the error taxonomy proposed by Popović (2018).

MQM[n] was developed by QTLaunchPad,[o] which defines human and machine translation quality and outlines its evaluation process with specific standards. The steps for the evaluation process are the following: (i) detecting the important features of the translation (called parameters or dimensions), (ii) selecting the relevant error categories and (iii) annotating them with the *translate5* tool to get a score. In its latest version (30 December 2015),[p] users should define 12 parameters before the evaluation task. These are language/locale (e.g. a text to be used by French-speaking readers in Canada), subject field/domain (e.g. law or pharmacology), terminology, text type (e.g. a manual), audience (e.g. the users of a washing machine), purpose (the aim of the text), register (e.g. formal or neutral), style (e.g. compliance with a style guide), content correspondence (e.g. whether a summary or a full translation should be provided), output modality (e.g. subtitles), file format (e.g. html) and production technology (e.g. use of translation memories). After defining the parameters, users select related issues (possible errors) (Figure 2), which can also be detected in the source text, to evaluate it and consider its quality in the final score (the translation can, in fact, improve the source text). Then, annotators select the segment that they consider includes an issue and the closest possible subcategory for its classification. If an issue cannot be classified in one of the subcategories, it is annotated in the more generic category. QTLaunchPad provides detailed annotation guidelines, with examples, specific cases and an algorithm for category selection. The annotation tool provided is *translate5*,[q] an open source tool, which takes CSV files as an input and exports results in the same file type. Although scoring is not necessary in MQM, a scoring mechanism is provided to achieve consistency; it includes weights according to error severity on a four-level scale (none, minor, major and critical) and a scoring algorithm.

MQM revision 0.9 allows for the harmonisation of MQM and DQF typology (developed by TAUS), thus creating a subset of MQM (Figure 3). The only exception to the harmonisation is the DQF kudos feature, a category used for extra points for an exceptionally good translation, which has not yet been included in the new scheme. The typology consists of seven categories (accuracy, fluency, terminology, style, locale convention, design and verity), of which the first two shall be presented in more detail.

The category of accuracy consists of five subcategories: addition (any word[s] or character[s] added to the translation with no reference to the source text); improper exact translation memory match; untranslated (words, usually acronyms, that should have been translated but have not); omission (any content or function word omitted in the translation), which also has the subcategory of omitted variable; and mistranslation, which is further subdivided into the categories of ambiguous translation, mistranslation of technical relationship and overly literal (e.g. in literal translation of idioms). On the other hand, the category of fluency consists of seven subcategories: character encoding, spelling, punctuation, link or cross-reference, grammatical register, inconsistency (with a further subcategory of inconsistency with external reference) and grammar (e.g. agreement errors).

[n] http://www.qt21.eu/quality-metrics/ (last access 09/02/2019).
[o] http://www.qt21.eu/launchpad/ (last access 09/02/2019).
[p] http://www.qt21.eu/mqm-definition/definition-2015-12-30.html (last access 09/02/2019).
[q] https://www.translate5.net (last access 09/02/2019).

https://doi.org/10.1017/S1351324919000469 Published online by Cambridge University Press

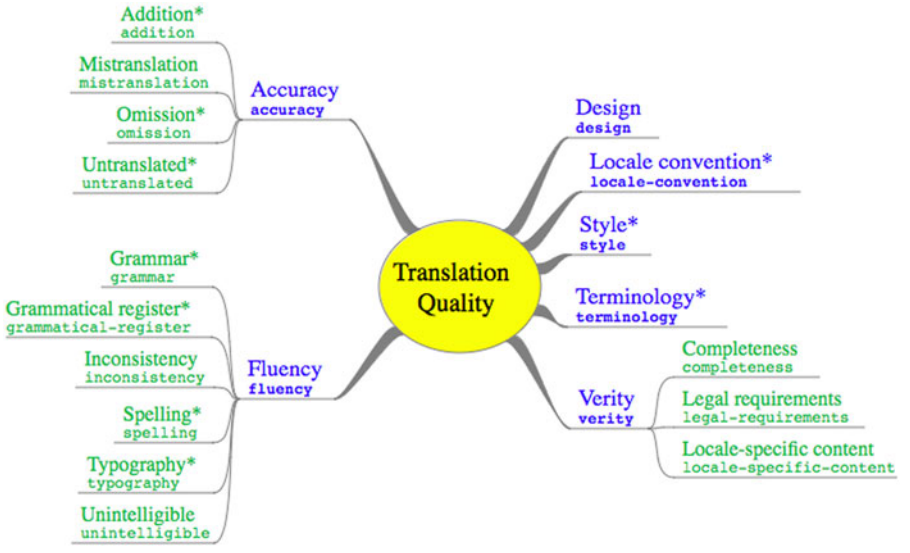

Figure 2. MQM core (http://www.qt21.eu/mqm-definition/definition-2015-06-16.html (last access 09/02/2019)).

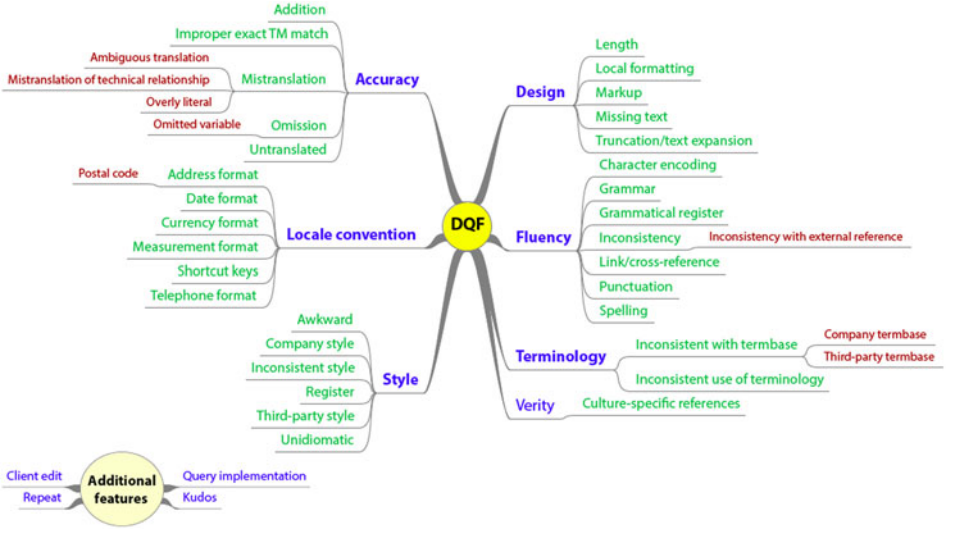

Figure 3. DQF subset (http://www.qt21.eu/mqm-definition/definition-2015-06-16.html (last access 09/02/2019)).

The last annotation tool presented in this section is Appraise (Figure 4), which provides an error typology, though without specifications, definitions or examples of each error type. The Appraise categories are *missing words*, *too many errors*, *terminology*, *lexical choice*, *syntax (ordering)*, *insertion (extra word)*, *morphology*, *misspelling*, *punctuation* and *other (idiom, etc.)*. It should be noted that the misspelling category has long been unnecessary for the MT output evaluation. This tool provides a two-level-scale error severity (minor and severe), which is not defined. Just like the other Appraise tasks, annotators can view the segment number, the source and target segments, as well as the *missing words* and *too many errors* options. Results are exported in XML files in the same format as they are imported.

Popović (2018) suggests a general taxonomy based on other existing taxonomies and the observation that, to improve the process, a set of broad categories with possible expansions should

https://doi.org/10.1017/S1351324919000469 Published online by Cambridge University Press

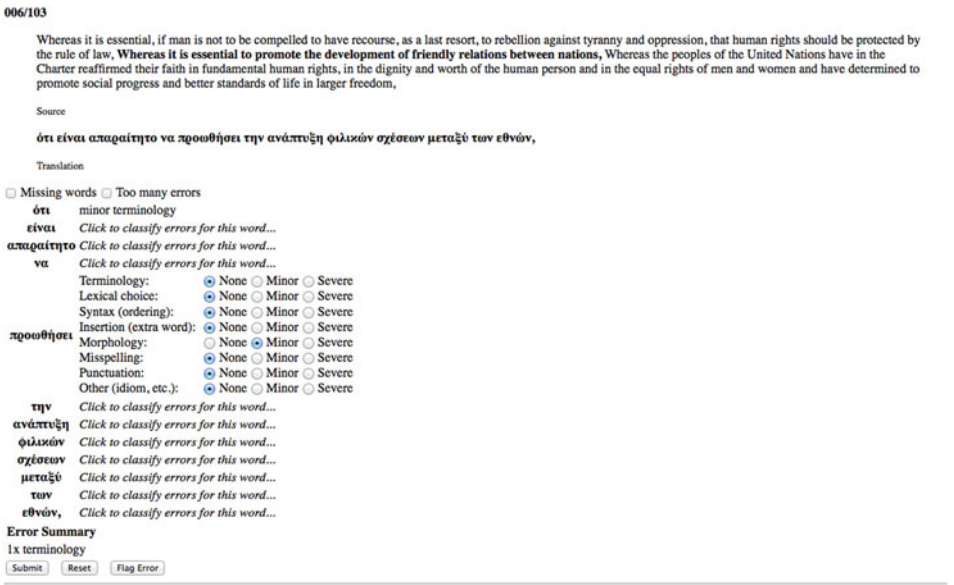

Figure 4. Appraise error classification environment.

be used. The broad categories are lexis, morphology, syntax, semantic, orthography and too many errors, a category that is useful for low-quality outputs but should be used with caution. The whole taxonomy with two more levels of expansion can be found in Table 2.

Further observations regarding error typologies have to do with what can be extracted and exploited to improve an MT system (Popović 2018). A distinction that might be useful would be between errors in function and content words as well as between errors in punctuation and grammar. On the other hand, it might be irrelevant whether a mistranslation is a lexical error or an overly literal translation, although this would be relevant for an evaluation designed to be used by translation services providers. This only comes to highlight the fact that the metric(s) to be chosen depends on the nature and the aims of the evaluation project. Other potentially important features of the typologies and their corresponding tools are whether the annotation can be done on a single word (e.g. in Appraise) or on a segment to be selected by the annotator (e.g. in the MQM environment) and whether, in the exported results, the exact word or segment where an error was detected is mentioned or not. Finally, the number of categories and subcategories as well as the number of error severity levels will also depend on the needs of the evaluation project; it should, however, be stressed that subtle differences between categories and levels require more cognitive effort, which is directly associated to the reliability of the evaluation, and, according to Popović (2018), a large number of categories can affect the consistency of a classification.

These observations are closely related to the low interannotator agreement of error classification tasks (Lommel, Popović and Burchardt 2014; Popović 2018), which can be attributed to factors such as disagreement as to error spans, ambiguity between categories and disagreement as to the existence and the severity of errors (Lommel, Popović and Burchardt 2014; Popović 2018). As Lommel, Popović and Burchardt (2014) point out, the deficiencies reported in their work have allowed the improvement of the annotation guidelines; nonetheless, they cannot be expected to be thoroughly eradicated, as they are 'inherent in the quality assessment task'.

Recent advances in error classification include automated and semiautomated approaches (Popović 2018). A complete scheme of automatic classification elaborated by Popović and Ney (2011) was merged with the Addicter tool in 2012 (Berka *et al.* 2012). Although automatic error

https://doi.org/10.1017/S1351324919000469 Published online by Cambridge University Press

Table 2. Error taxonomy by Popović (2018)

Level 1	Level 2	Level 3
Lexis	Mistranslation	Terminology
	Addition	
	Omission	
	Untranslated	
	Should not be translated	
Morphology	Inflection	Tense, number, person
		Case, number, gender
	Derivation	POS
		Verb aspect
	Composition	
Syntax	Word order	Range
	Phrase order	Range
Semantic	Multi-word expressions	
	Collocations	
	Disambiguation	
Orthography	Capitalisation	
	Punctuation	
	Spelling	
Too many errors		

classification tools still lack in precision, tend to assign incorrect error tags, strongly depend on reference translations and cannot provide as detailed annotations as humans, they do present a series of advantages: they can be used for preannotation to facilitate manual classification and are faster, cheaper and more consistent than manual annotation (Popović 2018).

5.2.4 Postediting

Postediting is defined as the task by which the MT output is transformed into a deliverable translation (Lacruz *et al.* 2014). It usually has to be defined as full or light. Full postediting renders the MT output human-like. According to Massardo *et al.* (2016), the human translation quality refers to a text which is comprehensible, accurate (i.e. it transmits the meaning of the source text) and stylistically fine, and in which 'syntax is normal, grammar and punctuation are correct'. The authors do mention, however, that the style may not be as good as the one of a native-speaking human translator. On the other hand, in light postediting, only the necessary changes are made so that the MT output can be comprehensible. The level of quality achieved by light postediting is called 'good enough' quality by Massardo *et al.* (2016), and it is defined as comprehensible and accurate but not 'stylistically compelling'. They go on to add that syntax can be unusual and grammar not perfect; therefore, the reader can tell that the text is machine generated.

Although it is mostly performed by professional translators, postediting training has only recently started to be a part of translation studies curricula (Lacruz *et al.* 2014). According to

https://doi.org/10.1017/S1351324919000469 Published online by Cambridge University Press

Lacruz *et al.* (2014), postediting differs greatly from translating and, therefore, the cognitive processes involved are also very different; Lacruz *et al.* (2014) conclude then that traditional translator training might not be ideal for performing postediting. Although they are indeed two distinct processes, they do share certain stages. When postediting is performed by comparing to the source text and not a gold (i.e. human) translation, it consists of the following steps: (i) detecting translation errors by contrasting the source and target text, (ii) detecting linguistic errors in the target language, (iii) fixing the errors and (iv) proofreading the edited segment. These subprocesses – or at least some of them – are not necessarily distinguishable; that is they can be – and usually are – performed in parallel and are often repeated until the desired output is achieved. A bilingual person who is not a translator can effectively detect translation errors, but translators have already developed skills in detecting them; it is a part of the translation process and also an independent task performed by many translators, as experts usually review, edit and/or proofread others' translations. What differs between the two procedures is the type of errors. For linguistic errors in the target language, the same argument could be used; it is still another subtask of the translator's work. Correcting these errors is a task which is very similar to translation; it requires skills such as searching for the adequate term, word or collocation by using the same tools which translators are already familiar with. Moreover, translators, although specialised in specific domains, are usually familiar with a range of different fields. In the correction phase, postediting normally looks to modify or edit as little as possible. Though this is indeed a skill not necessarily developed by a translator (although recommended in human translation editing), it is still a skill based on the other language processing skills developed by translators. Undoubtedly, one should not underestimate the differences between correcting a human translation and correcting the MT output, as they involve different phenomena. MT is also responsible for the posteditor's exposure to toxic texts, which include severe word order errors at the phrase level, the most demanding type of errors, according to Temnikova (2010). However, one cannot question the fact that there are language professionals who can be properly trained in postediting: translators. Their academic curriculum is the closest to postediting as currently exists, and professional translators are increasingly receiving training in postediting for professional purposes.

Postediting is also used as a quality metric by calculating the required temporal and cognitive effort (Lacruz *et al.* 2014), and there are currently publicly available tools that yield relevant statistical information on the postedits, such as Translog-II,[r] CASMACAT[s] and PET.[t] Research on postediting has not yet yielded sufficient results (Lacruz *et al.* 2014), and the growing need for this task makes it a highly relevant issue of scientific interest (Lacruz *et al.* 2014).

Postediting guidelines have been suggested, among others, by TAUS as well as in the frameworks of evaluation projects. The TAUS guidelines differ depending on the type of quality one wishes to achieve. For good enough quality, they suggest the following guidelines (Massardo *et al.* 2016):

(1) Aim for semantically correct translation.
(2) Ensure that no information has been accidentally added or omitted.
(3) Edit any offensive, inappropriate or culturally unacceptable content.
(4) Use as much of the raw MT output as possible.
(5) Basic rules regarding spelling apply.
(6) No need to implement corrections that are of a stylistic nature only.
(7) No need to restructure sentences solely to improve the natural flow of the text.

[r] http://www.translog.dk (last access 10/02/2019)
[s] http://www.caitra.org/index.php?n=Workbench.Workbench (last access 10/02/2019)
[t] http://www.clg.wlv.ac.uk/projects/PET/ (last access 10/02/2019)

https://doi.org/10.1017/S1351324919000469 Published online by Cambridge University Press

The TAUS guidelines for human translation quality are the following (Massardo *et al.* 2016, p. 18):

(1) Aim for grammatically, syntactically and semantically correct translation.
(2) Ensure that the key terminology is correctly translated and that untranslated terms belong to the client's list of "Do Not Translate" terms.
(3) Ensure that no information has been accidentally added or omitted.
(4) Edit any offensive, inappropriate or culturally unacceptable content.
(5) Use as much of the raw MT output as possible.
(6) Basic rules regarding spelling, punctuation and hyphenation apply.
(7) Ensure that the formatting is correct.

A different set of guidelines for light postediting are the ones used in 2007 at the GALE MT evaluation programme (Przybocki *et al.* 2011, p. 840):

(1) Make the MT output have the same meaning as the reference human translation: no more and no less.
(2) Make the MT output be as understandable as the reference. Similarly, try to make the MT output not be more or less ambiguous than the reference.
(3) Punctuation must be understandable, and sentence-like units must have a sentence-ending punctuation and proper capitalisation. Do not insert, delete or change punctuation merely to follow traditional rules about what is 'proper'.
(4) Capture the meaning in as few edits as possible using understandable English. If words/phrases/punctuation in the MT output are completely acceptable, use them (unmodified) rather than substituting something new and different.

In case of conflicts among these four rules, consider them to be ordered by importance.

The best practice is to adapt guidelines to the language pair, the translation direction and the parameters of each project; provide to annotators detailed instructions with examples; train them sufficiently; perform a pre-evaluation postediting task, which will allow the detection of possible flaws or issues not taken under consideration; and, finally, proceed to a final adjustment of the guidelines.[u]

As far as the postediting tool is concerned, in the Appraise environment, the MT output is presented in a box where annotators can intervene to perform the necessary edits. There is a *translate from scratch* option, and the import and export file type is XML.

The error classification and the postediting processes are intertwined. This is due to the fact that the first step of classification is the detection of an error. If the editor knows there is an error, it means he/she knows a correct alternative, so the cognitive process of detecting an error includes the process of correcting. Bojar (2011) concludes that annotators implicitly use an acceptable translation and annotate the necessary changes so that the MT output transforms into that acceptable translation. Moreover, when classification and postediting are performed simultaneously, the visual contact with the postedited segment contributes to optimising error detection and to keeping track of the number of annotated errors; that is the annotators ensure that the minimum number of errors is being annotated, if this is required. To illustrate this, a mistranslation error example can be used. If, for instance, only error classification is used and there is a lexical error in the MT output, the annotator can detect it and mark it as such but might fail to see the possible agreement error generated by the change of the mistranslated word. This scenario is avoided with the parallel use of postediting. Moreover, the number of errors and edits will coincide this way. Snover *et al.* (2006) conclude that the creation of a new reference translation and

[u]For an analysis of the postediting cognitive effort and the ways to reduce it, see Przybocki *et al.* (2011).

https://doi.org/10.1017/S1351324919000469 Published online by Cambridge University Press

error counting is preferable to the expression of subjective judgments. Popović (2018) also suggests merging error classification and postediting to facilitate the annotation task and improve the interannotator agreement.

5.3 An integrated environment for human evaluation: MT-EquAl

MT-EquAl (Machine Translation Errors, Quality, Alignment)[v] is a toolkit with three human evaluation tasks: error annotation; rating tasks, such as adequacy and fluency and ranking; and word alignment (Girard *et al*. 2014). Its main features are that it is an open-source, web-based and multiuser tool; it provides project management and progresses monitoring functions; and its tasks can be adapted to specific needs. It is currently incorporated in the MateCat project, a web-based computer-assisted translation tool.

5.4 Advantages of human evaluation

The advantages of human evaluation include the following: (i) given that translations are generated for human use, human judgment is considered to be the most adequate criterion (Sanders *et al*. 2011); (ii) human comprehension of the real world allows the judges to estimate the practical importance of translation errors (Sanders *et al*. 2011); and last but not least, (iii) it is considered that there is no substitute for human judgment in the case of translation and, therefore, this is the reference for quality in translation (Sanders *et al*. 2011). In the words of Graham *et al*. (2013), human annotations in natural language processing are required 'in order to estimate how well a given system mimics activities traditionally performed by humans'. Human metrics are still an important component of evaluation in the annual MT workshops, such as the WMT (Bojar 2011) and International Workshop on Spoken Language Translation (IWSLT) (Bentivogli *et al*. 2018).

5.5 Disadvantages of human evaluation

By far the most-mentioned disadvantage of human evaluation is its subjectiveness (Euromatrix 2007; Dorr *et al*. 2011). Nonetheless, it is precisely its so-called negative subjectiveness that constitutes the reference for the automated metrics quality. Other disadvantages include high cost, lack of repeatability and its time-consuming character, as well as low interannotator agreement (Dorr *et al*. 2011). The latter can be dealt with via statistical significance controls and independent evaluator teams (Dorr *et al*. 2011). Moreover, all of these disadvantages mentioned by Dorr *et al*. (2011) have been addressed by the use of crowdsourced DA, as described in Section 5.1.3. The process of evaluation is not exclusive to MT; for all human evaluation methods, the appropriate techniques are developed to minimise the adverse effects of interannotator agreement. One could pose the question: how is this problem dealt with in the language evaluation, such as in mother tongue writing tests or in foreign language written or oral discourse tests? The answer is: with strict criteria and evaluators' training. Advance has been made to this direction, but still important issues, such as the number of reference translations, ratings and postedits required for a reliable evaluation, remains unclear (Lommel, Popović and Burchardt 2014). Last but not least, Läubli *et al*. (2018) underscore the fact that human evaluation tasks are predominantly performed on the sentence level, which lacks in perception of intersentence cohesion. The authors claim that, although MT outputs are generated on the sentence level, their current state of fluency requires a document-level evaluation to detect flaws. Judges cannot evaluate textual cohesion and coherence if they are provided only with out-of-context sentences.

6. NMT challenges

The reported improvement in MT brought about by neural systems (Kalchbrenner and Blunsom 2013; Cho *et al*. 2014; Sutskever, Vinyals and Le 2014; Bahdanau, Cho and Bengio 2015; Wu *et al*.

[v] http://www.mt4cat.org/software/mt-equal (last access 09/02/2019).

https://doi.org/10.1017/S1351324919000469 Published online by Cambridge University Press

2016) poses new challenges for MT evaluation. Human parity is claimed to have been reached in a specific task, namely Chinese to English news texts (Hassan *et al.* 2018). Although the researchers underscore that the results cannot be generalised to other languages and domains, the results at the aforementioned WMT 2017 task achieve human parity in the statistical sense mentioned in the discussion on the MT quality in Section 2, that is statistical indistinguishability from human translations, and exceed the quality of crowdsourced translations (Hassan *et al.* 2018). However, human error analysis 'indicates that there is still room to improve machine translation quality', and the focus should be now posed on languages and domains which lack large amounts of data (Hassan *et al.* 2018). Although human parity, even in the sense of statistical indistinguishability, is a milestone for the MT quality, the same-level quality does not necessarily mean indistinguishability in terms of an imitation game. Moreover, crowdsourced translations, which are used for the comparison of quality in the work of Hassan *et al.* (2018), should not be the yardstick for the quality of translation, since it is not a professional work. Non-professional translations present significant differences compared to professionals', as suggested by the existence of translation learner corpora, such as MeLLANGE (Castagnoli *et al.* 2010), and the research on translation students' performance, which indicates less fluency in students' work than in professionals' (Carl and Buch-Kromann 2010). This point has been highlighted by Toral *et al.* (2018), who showed that the original language of the source text, the translation proficiency of the evaluators and the context are important elements that were not taken into consideration in the conclusions of Hassan *et al.* (2018).

On the one hand, the focus seems to be shifting towards specific linguistic phenomena, regardless of whether human or automated metrics or a combination of both is used. In this framework, Isabelle *et al.* (2017) suggest the evaluation of a challenge set of sentences, that is a set of sentences with linguistically demanding features, especially designed to challenge NMT systems. In the same line, Sennrich (2017) also focuses on 'linguistically interesting phenomena that have previously been found to be challenging for machine translation', such as agreement over long distances, transliteration of names and polarity. Klubièka, Tora, and Sánchez-Cartagena (2018) suggest a fine-grained manual evaluation based on MQM, to compare between statistical (pure and factored phrase-based) and neural MT systems. Their approach includes the adaptation of the MQM taxonomy to the linguistic phenomena of the languages they use in their evaluation project. Their results show that this kind of metric can capture the relevant features that pose challenges to NMT evaluation, precisely because of the detailed feedback on the linguistic phenomena involved.

On the other hand, Koehn and Knowles (2017) list the following challenges for NMT: domain mismatch, the amount of training data, rare words, long sentences, word alignment and beam search. To evaluate NMT outputs regarding these domains, they use BLEU but adapt the task for each domain, for example by modifying the conditions of systems' training.

7. Conclusions

Summarising, MT developers have an array of evaluation methods from which to select the adequate method for each project; there are also toolkits with a variety of metrics, such as Asiya and MT-EquAl. The most widely used automated metrics, that is the ones based on reference translations, present the main advantages of speed, low cost, consistency and reusability (Banerjee and Lavie 2005; Lavie 2011). However, they also have certain disadvantages: they do not distinguish between nuances, they are not considered reliable for segment-level evaluation (Lavie 2011), their scores are not very intuitive (Koehn 2010) and they do not provide feedback about the strengths and drawbacks of the system under evaluation (Zhou *et al.* 2008). Human judges, on the other hand, can evaluate the severity of MT errors and address some of the disadvantages of the automated metrics, such as distinguishing between nuances (Sanders *et al.* 2011); however, they present the disadvantages of higher cost, requiring more time, lack of repeatability and lack of consistency (Dorr *et al.* 2011). Crowdsourcing has, however, addressed the cost, time and consistency issues (Graham *et al.* 2015).

https://doi.org/10.1017/S1351324919000469 Published online by Cambridge University Press

Taking the aforementioned advantages and drawbacks of each type of evaluation under consideration, a combination of automated and human metrics is suggested as the most reliable method for evaluating the MT output. Automated metrics can be used as a guide in large corpora to select a percentage of the texts to be evaluated by humans, for instance, proceeding with human evaluations in the case of very low or very high scores, or in the case of very similar scores between two different MT systems. However, as already discussed, the best practice is to adapt the evaluation to the project's goals. Therefore, a task-based evaluation could be adequate for gisting purposes, whereas more fine-grained metrics should be used for the improvement of systems, in which case a combination of metrics, such as error classification and postediting, would be adequate.

As far as the judges' profile is concerned, according to our review, we suggest the following recommendations. First, to avoid the reference bias, use bilingual judges. Second, if a small number of judges are to be used, these should be trained translators with clear guidelines according to the project parameters. Finally, interannotator agreement tests are indispensable. If crowdsourcing is to be used, it has been shown (Graham *et al*. 2015) that non-experts yield reliable results, given that intra- and interannotator agreement is covered.

As far as current needs are concerned, given the recent advances in terms of the MT quality (Kalchbrenner and Blunsom 2013; Cho *et al*. 2014; Sutskever *et al*. 2014; Bahdanau *et al*. 2015; Wu *et al*. 2016; Hassan *et al*. 2018), evaluation methods are to be sensitive to nuances; therefore, the focus is on specific linguistic phenomena of specific language pairs and directions as well as specific demanding domains. In this regard, evaluation projects need to be elaborated on the basis of challenge sets of these specific features. Lastly, the text span used for evaluation has also been questioned under this light, and document-level evaluation is now recommended (Läubli *et al*. 2018) and could be considered in future evaluation projects.

Acknowledgments. I would like to acknowledge the contribution of all the academics of ILCL who participated in the project 'Fortalecimiento de la publicación de los Departamentos de Didáctica, Inglés y Traducción del ILCL'. Special thanks to Ricardo Benítez Figari, Stephanie Díaz-Galaz and Rogelio Nazar for their fruitful comments, as well as to Stelios Piperidis, who triggered this work.

Financial support. This work was partially supported by the project 'Fortalecimiento de la publicación de los Departamentos de Didáctica, Inglés y Traducción del ILCL', Pontificia Universidad Católica de Valparaíso.

References

Abend O. and Rappoport A. (2013). Universal conceptual cognitive annotation (UCCA). In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Sofia, Bulgaria , pp. 228–238.

Ageeva E., Tyers F., Forcada M. and Perez-Ortiz J. (2015). Evaluating machine translation for assimilation via a gap-filling task. In *Proceedings of the Conference of the European Association for Machine Translation*, Antalya, Turkey, pp. 137–144.

Amigo E., Giménez J., Gonzalo J. and Màrquez L. (2006). MT evaluation: Human-like vs. human acceptable. In *Proceedings of COLING-ACL06, Joint Conference of the International Committee on Computational Linguistics and the Association for Computational Linguistics*, Sydney, Australia.

Babych B. and Hartley A. (2004). Extending BLEU MT evaluation method with frequency weighting. In *Proceedings of ACL (Association for Computational Linguistics)*, Barcelona, Spain.

Bahdanau D., Cho K. and Bengio Y. (2015). Neural machine translation by jointly learning to align and translate. In *Proceedings of ICRL 2015*, San Diego, USA.

Banerjee S. and Lavie A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*, Ann Arbor, Michigan, pp. 65–72.

Bentivogli L., Cettolo M., Federico M. and Federmann C. (2018). Machine translation human evaluation: An investigation of evaluation based on Post-editing and its relation with Direct Assessment. In *Proceedings of the International Workshop on Spoken Language Translation*, Bruges, Belgium, pp. 62–69.

Berka J., Bojar O., Fishel M., Popovic M. and Zeman D. (2012). Automatic MT error analysis: Hjerson helping Addicter. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC'12*, Istanbul, Turkey, pp. 2158–2163.

Birch A., Abend O., Bojar O. and 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, pp. 1264–1274.

Blatz J., Fitzgerald E., Foster G., Gandraburn S., Goutte C., Kulesza A., Sanchis A. and Ueffing N. (2004). Confidence estimation for machine translation. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING)*, Geneva, Switzerland.

Bojar O. (2011). Analyzing error types in English-Czech machine translation. *Prague Bulletin of Mathematical Linguistics* **95**, 63–76.

Bojar O., Federmann C., Haddow B., Koehn P., Post M. and 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"*. Available at http://www.cracking-the-language-barrier.eu/wp-content/uploads/Bojar-Federmann-etal.pdf.

Callison-Burch C. (2009). Fast, cheap, and creative: Evaluating translation quality using Amazon's Mechanical Turk. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, Suntec, Singapore, pp. 286–295.

Callison-Burch C., Fordyce C., Koehn P., Monz C. and Schroeder J. (2007). (Meta-)evaluation of machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation '07*, Prague, Czech Republic, pp. 136–158.

Callison-Burch C., Osborne M. and Koehn P. (2006). Re-evaluating the role of BLEU in machine translation research. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguisics (EACL)*, Trento, Italy, pp. 249–256.

Carl M. and Buch-Kromann M. (2010). Correlating translation product and translation process data of professional and student translators. In *Proceedings of the Annual Conference of the European Association for Machine Translation*, Saint-Raphaél, France.

Castagnoli S., Ciobanu D., Kunz K., Volanschi A. and Kübler N. (2010). Designing a learner translator corpus for training purposes. In Kübler N. (ed), *Corpora, Language, Teaching and Resources: From Theory to Practice*. Bern, Switzerland: Peter Lang.

Cho K., Van Merriënboer B., Bahdanau B. and Bengio Y. (2014). On the properties of neural machine translation: Encoder-decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, Doha, Qatar, pp. 103–111.

Coughlin D. (2003). Correlating automated and human assessments of machine translation quality. In *Proceedings of MT Summit IX*, New Orleans, LA, USA, pp. 63–70.

Doddington G. (2002). Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the 2nd Human Language Technologies Conference (HLT-02)*, San Diego, CA, USA, pp. 128–132.

Dorr B., Snover M. and Madnani N. (2011). Chapter 5.1 introduction. In Olive J., McCary J. and Christianson C. (eds), *Handbook of Natural Language Processing and Machine Translation. DARPA Global Autonomous Language Exploitation*. New York: Springer, pp. 801–803.

Dreyer M. and Marcu D. (2012). HyTER: Meaning-equivalent semantics for translation evaluation. In *Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Montreal, Canada, pp. 162–171.

Euromatrix (2007). Survey of machine translation evaluation. *Statistical and Hybrid Machine Translation Between All European Languages, IST 034291, Deliverable 1.3*.

Federmann C. (2010). Appraise: An open-source toolkit for manual phrase-based evaluation of translations. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC)*, Valletta, Malta.

Federmann C. (2018). Appraise evaluation framework for machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, Santa Fe, New Mexico, USA, pp. 86–88.

Gandrabur S. and Foster G. (2003). Confidence estimation for translation prediction. In *Proceedings of the 7th Conference on Natural Language Learning at HLT-NAACL (CONLL)*, Edmonton, Canada.

Girardi C., Bentivogli L., Farajian M. and Federico M. (2014). MT-EquAl: A toolkit for human assessment of machine translation output. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations*, Dublin, Ireland, pp. 120–123.

Gonzàlez M. and Giménez J. (2014). *Asiya. An open toolkit for automatic machine translation (meta-)evaluation*. Technical Manual, version 3.0. TALP Research Center, LSI Department, Universitat Politècnica de Catalunya.

Görög A. (2014). Quantifying and benchmarking quality: The TAUS Dynamic Quality Framework. *Revista Tradumàtica: tecnologies de la traducció, Traducció i qualitat* 12. ISSN: 1578–7559. Available at http://revistes.uab.cat/tradumatica.

Graham, Y., Baldwin, T., Moffat, A. and Zobel, J. (2013). Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop & Interoperability with Discourse, Sofia*, Bulgaria, pp. 33–41.

Graham Y., Baldwin T., Moffat A. and Zobel J. (2015). Can machine translation systems be evaluated by the crowd alone. *Natural Language Engineering* **23**(1), 3–30.

Han A.L.F., Wong D.F. and Chao L.S. (2012). LEPOR: A robust evaluation metric for machine translation with augmented factors. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012): Posters*, Mumbai, India, pp. 441–450.

Han L. (2018). Machine translation evaluation resources and methods: A survey. arXiv:1605.04515v8. Cornell University Library.

https://doi.org/10.1017/S1351324919000469 Published online by Cambridge University Press

Hassan H., Aue A., Chen C., Chowdhary V., Clark J., Federmann C., Huang X., Junczys-Dowmunt M., Lewis W., Li M., Liu S., Liu T., Luo R., Menezes A., Qin T., Seide F., Tan X., Tian F., Wu L., Wu S., Xia Y., Zhang D., Zhang Z. and Zhou M. (2018). Achieving human parity on automatic Chinese to English news translation. arXiv:1803.05567.

House J. (2014). *Translation Quality Assessment: Past and Present*. New York: Routledge.

Isabelle P., Cherry C. and Foster G. (2017). A challenge set approach to evaluating machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark, pp. 2486–2496.

Kalchbrenner N. and Blunsom P. (2013). Recurrent continuous translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, Seattle, Washington, USA, pp. 1700–1709.

Klubièka F., Toral A. and Sánchez-Cartagena V. (2018). Quantitative fine-grained human evaluation of machine translation systems: A case study on English to Croatian. arXiv:1802.01451v1.

Koby G.S., Fields P., Hague D., Lommel A. and Melby A. (2014). Defining translation quality. *Tradumàtica* **12**, 413–420.

Koehn P. (2010). *Statistical Machine Translation*. Cambridge: Cambridge University Press.

Koehn P. and Knowles R. (2017). Six challenges for neural machine translation. arXiv:1706.03872v1.

Koehn P. and Monz C. (2006). Manual and automatic evaluation of machine translation between European languages. In *Proceedings of the 2006 Workshop on Statistical Machine Translation*, New York, USA.

Lacruz I., Denkowski M. and Lavie A. (2014). Cognitive demand and cognitive effort in post-editing. In *Proceedings of the Third Workshop on Post-Editing Technology and Practice. 11th Conference of the Association for Machine Translation in the Americas*, Vancouver, BC, Canada.

Läubli S., Sennrich R. and Volk M. (2018). Has machine translation achieved human parity? A case for document-level evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium, pp. 4791–4796.

Lavie A. (2011). Evaluating the output of machine translation systems. In *Proceedings of the 13th MT Summit*, Xiamen, China.

Leusch G, Ueffing N. and Ney H. (2003). A novel string-to-string distance measure with applications to machine translation evaluation. In *Proceedings of MT Summit IX*, New Orleans, LA, USA.

Leusch G, Ueffing N. and Ney H. (2006). CDER: Efficient MT evaluation using block movements. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy.

Levenshtein V. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics – Doklady* **10**(8), 707–710. Original in Russian 1965.

Lin C.Y. and Och F.J. (2004). Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL-04)*, Main Volume, Barcelona, Spain, pp. 605–612.

Lita L.V., Rogatti M. and Lavie A. (2005). BLANC: Learning evaluation metrics for MT. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP)*, Vancouver, Canada, pp. 740–747.

Lommel A., Popović M. and Burchardt A. (2014). Assessing inter-annotator agreement for translation error annotation. In *Proceedings of LREC Workshop on Automatic and manual Metrics for Operational Translation Evaluation*, Reykjavik, Iceland.

Martins A., Junczys-Dowmunt M., Kepler F., Astudillo R., Hokamp C. and Grundkiewicz R. (2017). Pushing the limits of translation quality estimation. *Transactions of the Association for Computational Linguistics* **5**, 205–218.

Massardo I., Van der Meer J., O'Brien S., Hollowood F., Aranberri N. and Drescher K. (2016). *MT Post-Editing Guidelines*. The Netherlands: TAUS Signature Editions.

Melamed I., Green R. and Turian J. (2003). Precision and recall of machine translation. In *Proceedings of the HLT-NAACL 2003*, Edmonton, Canada.

Navarro G. (2001). A guided tour to approximate string matching. *ACM Computing Surveys* **33**(1), 31–88.

Newmark P. (1988). *A Textbook of Translation.* Essex: Pearson Education Limited.

Niessen S., Och F., Leusch G. and Ney H. (2000). An evaluation tool for machine translation: Fast evaluation for MT research. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation*, Athens, Greece.

Nord C. (1997). *Translating as a Purposeful Activity: Functionalist Approaches Explained*. Manchester: St. Jerome.

Papineni K., Roukos S., Ward T. and Zhu W.J. (2002). BLEU: A method for automatic evaluation of machine translation. In *Proceedings of ACL-2002: 40th Annual meeting of the Association for Computational Linguistics*, Philadelphia, Pennsylvania, USA, pp. 311–318. CiteSeerX: 10.1.1.19.9416

Popović M. (2015). CHRF: Character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, Lisbon, Portugal, pp. 392–395..

Popović M. (2018). Error classification and analysis for machine translation quality assessment. In Moorkens J., Castilho S., Gaspari F. and Doherty S. (eds), *Translation Quality Assessment. From Principles to Practice*. Cham, Switzerland: Springer.

Popović M. and Ney H. (2007). Word error rates: Decomposition over POS classes and applications for error analysis. In *Proceedings of the Second Workshop on Statistical Machine Translation,* Association for Computational Linguistics, Prague, Czech Republic, pp. 48–55.

Popović M. and Ney H. (2011). Towards automatic error analysis of machine translation output. *Computational Linguistics* **37**(1), 657–688.

Przybocki M., Le A., Sanders G., Bronsart S., Strassel S. and Glenn M. (2011). Chapter 5.4.3 Post-editing. In Olive J., McCary J. and Christianson C. (eds), *Handbook of Natural Language Processing and Machine Translation. DARPA Global Autonomous Language Exploitation*. New York: Springer.

Przybocki M., Peterson K., Bronsart S. and Sanders G. (2009). *The NIST 2008 Metrics for Machine Translation Challenge – Overview, Methodology, Metrics, and Results*. Gaithersburg MD, USA: Multimodal Information Group, National Institute of Standards and Technology.

Quirk C.B. (2004). Training a sentence-level machine translation confidence measure. In *Proceedings of the 4th Conference on Language Resources and Evaluation*, Lisbon, Portugal, pp. 825–828.

Ricoeur P. (2003). *Sur la traduction*. Paris: Bayard.

Sánchez-Gijón, P. and Torres-Hostench, O. (2014). MT post-editing into the mother tongue or into a foreign language? Spanish-to-English MT translation output post-edited by translation trainees. In *Proceedings of the Third Workshop on Post-Editing Technology and Practice, 11th Conference of the Association for Machine Translation in the Americas (AMTA)*, Vancouver, Canada, pp. 5–19.

Sanders G., Przybocki M., Madnani N. and Snover M. (2011). Chapter 5.1.2 human subjective judgments. In Olive J., McCary J. and Christianson C. (eds), *Handbook of Natural Language Processing and Machine Translation. DARPA Global Autonomous Language Exploitation*. New York: Springer, pp. 806–807.

Sennrich R. (2017). How grammatical is character-level neural machine translation? Assessing MT quality with contrastive translation pairs. arXiv:1612.04629v3.

Snover M., Dorr B., Schwartz R., Micciulla L. and Makhoul J. (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*, Boston Marriott, Cambridge, Massachusetts, USA.

Specia L., Raj D. and Turchi M. (2010). Machine translation evaluation versus quality estimation. *Machine Translation* **24**, 39–50. Springer Science+Business Media B.V. doi:10.1007/s10590-010-9077-2.

Specia L., Shah K., De Souza J.G.C. and Cohn T. (2013). QuEst – A translation quality estimation framework. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, Sofia, Bulgaria, pp. 79–84.

Specia L., Turchi M., Cancedda N., Dymetman M. and Cristianini N. (2009). Estimating the Sentence Level Quality of Machine Translation Systems. In *EAMT09*, Barcelona, Spain, pp. 28–37.

Sutskever I., Vinyals O. and Le Q. (2014). Sequence to sequence learning with neural networks. In *Proceedings of Advances in Neural Information Processing Systems*, Montreal, Canada, pp. 3104–3112.

Temnikova I. (2010). A cognitive evaluation approach for a controlled language post-editing experiment. In *Proceedings of International Conference Language Resources and Evaluation (LREC2010)*, Valletta, Malta.

Tomás J., Mas J.A. and Casacuberta F. (2003). A quantitative method for machine translation evaluation. In *Proceedings of the EACL 2003 Workshop on Evaluation Initiatives in Natural Language Processing: Are Evaluation Methods, Metrics and Resources Reusable?*, Budapest, Hungary.

Toral A., Castilho S., Hu K. and Way A. (2018). Attaining the unattainable? Reassessing claims of human parity in neural machine translation. In *Proceedings of the Third Conference on Machine Translation (WMT), Volume 1: Research Papers, Association for Computational Linguistics*, Brussels, Belgium, pp. 113–123.

Turing A. (1950). Computing machinery and intelligence. *Mind* **49**, 433–460.

Ueffing N. and Ney H. (2005). Application of word-level confidence measures in interactive statistical machine translation. In *Proceedings of the 10th Conference of the European Association for Machine Translation*, Budapest, Hungary, pp. 262–270.

Wisniewski G., Kumar Singh A. and Yvon F. (2012). Quality estimation for machine translation: Some lessons learned. *Machine Translation* **27**(3–4), 213–238. doi:10.1007/s10590-013-9141-9.

Wu Y., Schuster M., Chen Z., Le Q.V., Norouzi M., Macherey W., Krikun M., Cao Y., Gao Q., Macherey K., Klingner J., Shah A., Johnson M., Liu X., Kaiser L., Gouws S., Kato Y., Kudo T., Kazawa H., Stevens K., Kurian G., Patil N., Wang W., Young C., Smith J., Riesa J., Rudnick A., Vinyals O., Corrado G., Hughes M. and Dean J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR* abs/1609.08144. Available at: http://arxiv.org/abs/1609.08144.

Zhou L., Lin C.-Y. and Hovy E. (2006). Re-evaluating machine translation results with paraphrase support. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Sydney, Australia.

Zhou M., Wang B., Liu S., Li M., Zhang D. and Zhao T. (2008). Diagnostic evaluation of machine translation systems using automatically constructed linguistic check-points. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, Manchester, United Kingdom, pp. 1121–1128.

Cite this article: Chatzikoumi E (2020). How to evaluate machine translation: A review of automated and human metrics. *Natural Language Engineering* **26**, 137–161. https://doi.org/10.1017/S1351324919000469

https://doi.org/10.1017/S1351324919000469 Published online by Cambridge University Press