Lang Resources & Evaluation (2022) 56:593–619
https://doi.org/10.1007/s10579-021-09537-5

SURVEY

Machine translation systems and quality assessment: a systematic review

Irene Rivera-Trigueros[1]

Accepted: 6 March 2021 / Published online: 10 April 2021

Abstract Nowadays, in the globalised context in which we find ourselves, language barriers can still be an obstacle to accessing information. On occasions, it is impossible to satisfy the demand for translation by relying only in human translators, therefore, tools such as Machine Translation (MT) are gaining popularity due to their potential to overcome this problem. Consequently, research in this field is constantly growing and new MT paradigms are emerging. In this paper, a systematic literature review has been carried out in order to identify what MT systems are currently most employed, their architecture, the quality assessment procedures applied to determine how they work, and which of these systems offer the best results. The study is focused on the specialised literature produced by translation experts, linguists, and specialists in related fields that include the English–Spanish language combination. Research findings show that neural MT is the predominant paradigm in the current MT scenario, being Google Translator the most used system. Moreover, most of the analysed works used one type of evaluation—either automatic or human—to assess machine translation and only 22% of the works combined these two types of evaluation. However, more than a half of the works included error classification and analysis, an essential aspect for identifying flaws and improving the performance of MT systems.

Keywords Evaluation · Machine translation · Systematic review · Quality

✉ Irene Rivera-Trigueros
irenerivera@ugr.es

[1] Department of Translation and Interpreting, Universidad de Granada, C/Buensuceso, 11, 18002 Granada, Spain

1 Introduction

Language barriers can be an obstacle to accessing information in the globalised context in which we find ourselves. Such is the abundance of information generated that it is on occasions impossible to satisfy the demand for translations by relying solely on professional human translators (Lagarda et al., 2015; Way, 2018). One of the implications of this situation is the growing demand for tools that provide different types of audiences with multilingual access to information. Machine translation (MT) is therefore profiled as one of the resources with the greatest potential for solving this problem and has been a point of focus both in terms of research and from the perspective of professional settings. Consequently, new MT paradigms and systems frequently emerge that could also integrate other resources such as translation memories or terminology databases to optimise the effectiveness of the professional translation process (Koponen, 2016).

One of the MT battlefields, however, refers to the quality of the product it creates, which is generally inferior to that reached by professional human translations. In this regard, measuring the quality of an MT system can present great difficulties, given that in the majority of cases there is not just one translation for an original text that may be considered correct (Mauser et al., 2008; Shaw & Gros, 2007). Despite this, it should be possible to determine the quality of how an MT system works, and its impact on the workflow of professional translators, in an objective manner. This will require both automated and human metrics that need, in addition, to take into account the human post-editing that is usually necessary for MT. Moreover, the annotation and classification of translation errors is fundamental for contributing to the improvement of MT systems, in order to understand the criteria of human metrics—given that this type of assessment has an element of subjectivity—and to optimise the post-editing process (Costa et al., 2015; Popović, 2018).

The aim of this study is to identify what MT systems are currently most employed, their architecture, the quality assessment procedures applied to determine how they work, and which of these systems offer the best results. The methodology is based on a systematic review of the specialised literature created by translation experts, linguists and specialists in related fields. Thus, the approach from which this study is tackled seeks to complement others that are frequently centred on the sphere of computational sciences. We start out from the consideration that, in order to determine translation quality, it is essential to incorporate the perspective of translation experts or areas related to language study because human evaluation and error annotation are extremely relevant when measuring MT quality—they are both processes that must be carried out by evaluators trained in the field of translation—.

The state of the art is developed below, which includes the evolution that machine translation has undergone, along with the main proposals concerning its assessment. The next section details the methodology employed for the systematic literature review process, specifying both the sample selection process and the analysis thereof. Following on, the results obtained are presented and discussed. Lastly, the conclusions that can be extracted from the study are formulated.

1.1 Evolution of MT

MT has traditionally presented two different approaches: the approach based on rules (RBMT, Rule-Based Machine Translation), and the corpus-based approach (Hutchins, 2007). Nevertheless, the last few years have seen the development of new architectures, giving rise to hybrid approaches and, most recently, neural MT (Castilho et al., 2017; España-Bonet & Costa-jussà, 2016).

RBMT systems use bilingual and monolingual dictionaries, grammars and transfer rules to create translations (Castilho et al., 2017; España-Bonet & Costa-jussà, 2016). The problem with these systems is that they are extremely costly to maintain and update and, furthermore, due to language ambiguity they can create problems when translating, for example, idiomatic expressions (Charoenpornsawat et al., 2002). At the end of the 1980s, corpus-based systems began to gain in popularity (Hutchins, 2007). These machine translators employ bilingual corpora of parallel texts to create translations (Hutchins, 1995). Corpus-based automatic systems are divided into statistical MT systems (SMT) and example-based systems (EBMT); despite this, both approaches converge on many aspects and isolating the distinctive characteristics of each one is very complicated (Hutchins, 2007). Moreover, the statistical approach was the predominant model until the recent emergence of neural MT systems (Bojar et al., 2015; España-Bonet & Costa-jussà, 2016; Hutchins, 2007). The advantage of these compared to RBMT systems is their solid performance when selecting the lexicon—especially if focusing on a thematic domain—and the little human effort they require in order to be trained automatically (Hutchins, 2007; Koehn, 2010). However, they can sometimes produce translations that are badly structured or have grammatical errors, added to which is the difficulty in finding corpora of certain thematic domains or language pairs (España-Bonet & Costa-jussà, 2016; Habash et al., 2009).

The hybrid approaches arose with the objective of attempting to overcome the problems caused by the RBMT and SMT systems and combine the advantages of both, in order to improve translation quality and precision (Hunsicker et al., 2012; Tambouratzis et al., 2014; Thurmair, 2009). The hybrid approach can be implemented in different ways and, generally speaking, a distinction can be made between those architectures with an RBMT system at their core or, in contrast, an SMT system. Thus, in some cases the output of an RBMT system is adjusted and corrected using statistical information, while others see rules being employed to process both the input and the output of an SMT system (España-Bonet & Costa-jussà, 2016).

Neural MT is currently dominating the paradigms of machine translation, this kind of MT "attempts to build and train a single, large neural network that read a sentence and outputs a correct translation" (Bahdanau et al., 2015, p.1).These systems are based on neural networks to create translations thanks to a recurrent neural architecture, based on the encoder-decoder model in which the encoder reads and encodes the source sentence into a fixed-length vector while the decoder produces a translation output from the encoded vector (Bahdanau et al., 2015; Cho et al., 2014). Consequently, this architecture implies a simplification regarding previous paradigms, given that they use less components and processing steps,

moreover, they require much less memory than SMT and allow to use human and data resources more efficiently than RBMT (Bentivogli et al., 2016; Cho et al., 2014). Such has been the success of these models that the main MT companies—Google, Systran, Microsoft, etc.—have already integrated them into the technologies of their machine translators.

1.2 MT quality assessment

It is essential to measure the quality of an MT system to improve how it performs. There is, however, a great lack of consensus and standardisation in relation to translation quality assessment—both human and machine—given the complicated cognitive, linguistic, social, cultural and technical process this supposes (Castilho et al., 2018). According to House (2014) translation quality assessment will mean a constant to and from a macro-analytic approach, wherein questions of ideology, function, gender or register are considered, to a micro analytical one in which the value of collocations and individual linguistic units are considered. Nevertheless, it should be taken into account that these approaches can differ enormously according to the individuals, groups or contexts in which quality is assessed. Thus, quality assessment in the industry is normally focused on the final product or customer, whereas in the field of research the purpose can be to demonstrate significant improvements over prior studies or different translation processes (Castilho et al., 2018).

Taking into consideration this panorama, together with the difficulty and lack of consensus regarding MT quality assessment, general distinctions between human (or manual) and automated metrics can be made. It is worth mentioning, though, that there are other ways to evaluate quality, focused on the human revision process rather than on the translation output, for example by measuring the post-editing effort in temporal, technical and cognitive terms.

1.2.1 Automated and human metrics

Automated metrics in general compare the output of an MT system with one or more reference translations (Castilho et al., 2018; Han, 2016). One of the first metrics used, Word Error Rate (WER) was based on the Levenshtein or edit distance (Levenshtein, 1966; Nießen et al., 2000). This measurement does not admit the reordering of words and substitutions, deletions and insertions are equal. The number of edit operations is divided between the number of words in the reference translation (Castilho et al., 2018; Han, 2016; Mauser et al., 2008). The PER (Position-Independent Word Error Rate, Tillmann et al., 1997) and TER (Translation Error Rate, Snover et al., 2006) metrics attempt to solve the problem created by WER by not allowing the reordering of words. Thus, PER compares the words in the two sentences without taking into account the order and TER counts the reordering of words as a further edit (Han, 2016; Mauser et al., 2008; Nießen et al., 2000; Snover et al., 2006). The most popular metric is Bilingual Evaluation Understudy (BLEU), a precision measurement carried out at the level of n-grams,

indivisible language units. It employs a modified precision that takes into account the maximum number of each n-gram appearance in the reference translation and applies a brevity penalty that is added to the measurement calculation (Papineni et al., 2002). This measurement became very popular as it showed good correlations with human evaluations and its usage extended amongst different MT evaluation workshops (Castilho et al., 2018). There are also other precision-centred metrics such as NIST (Doddington, 2002), ROUGE (Lin & Hovy, 2003), F-measure (Turian et al., 2003) and METEOR (Banerjee & Lavie, 2005), amongst others.

In relation to human evaluation, this usually occurs in terms of adequacy and fluency. Adequacy evaluates semantic quality, that is, if the information has been correctly transmitted or not, which requires comparison with reference translations (monolingual) or with the original text (bilingual). For its part, fluency evaluates syntactic quality; in this case, comparison with the original text is unnecessary and evaluation is monolingual. Moreover, other methods can be employed to evaluate the legibility, comprehension, usability and acceptability of translations (Castilho et al., 2018). A number of instruments can be employed to measure these aspects, such as for example the Likert-type ordinal scales, rankings—either selecting the best translation from various, or ordering the different options from better to worse according to specific criteria—or employing other methods such as error correction or gap-filling tasks, the latter not involving direct judgement from the evaluator (Castilho et al., 2018; Chatzikoumi, 2020). Further, it is important to mention that error identification, annotation and classification is another widely used human evaluation method, which shall be looked at in more detail in the following section.

There are pros and cons to both human and automatic evaluation. On the one hand, automatic evaluation requires less human effort, is more objective and less costly than human evaluation. Nevertheless, it must be taken into account that the majority of automated metrics require reference translations created by humans and, in many cases, the quality of these translations is assumed, but not verified, which could introduce an element of subjectivity and variability (Castilho et al., 2018). In addition, these metrics evaluate translation in relation to its similarity to the reference translations—despite there being no one single correct translation—and its capacity to evaluate syntactic and semantic equivalence is extremely limited (Castilho et al., 2018; Han, 2016). Finally, the large majority of these metrics arose from systems with outdated architectures, to which on occasions they are not adjusted to current paradigms (Way, 2018).

In contrast, although the results of human metrics are considered to be more reliable than those provided by automated metrics, the disadvantages of this method include large demands on time and resources, and it cannot be reproduced (Han, 2016). Furthermore, human evaluators—or annotators—must fulfil certain criteria in order to assure the reliability of the results. In the same vein, there is a need for training, evaluation criteria and familiarity with the subject area of the texts on the part of evaluators. Additionally, the ideal procedure includes more than one evaluator and calculating the inter-annotator agreement (Chatzikoumi, 2020; Han, 2016). It is worth mentioning that some authors have addressed the drawbacks on time and cost of human evaluation by means of crowdsourcing (Graham et al., , 2013, 2015). In the face of this situation, the combination of both human and

automatic metrics appears to be one of the most reliable methods for MT evaluation (Chatzikoumi, 2020).

1.2.2 Classification and analysis of MT errors

It will frequently be necessary to investigate the strengths and weaknesses of MT systems and the errors they produce, along with their impact on the post-editing process. In this regard it is very difficult to find a relationship between these aspects and the quality scores obtained both by automated and human metrics (Popović, 2018). Thus, the identification, classification and analysis of errors are fundamental to determining the failures of an MT system and being able to improve the performance thereof.

Error classification can be implemented automatically, manually or using combined methods. These methods have advantages and drawbacks in the same way as human and automated metrics do. Manual error classification, as in the case of human metrics, is costly and requires a lot of time and effort and additionally, normally presents problems regarding inter-annotator agreement. In contrast, automatic methods can overcome these problems, but they give rise to confusion between the different types of errors, especially for very detailed typologies, and they also require human reference translations (Popović, 2018).

There has traditionally been, as in the case of that which occurs with quality assessment, a lack of standardisation for MT error analysis and classification (Lommel, 2018). In this regard, many authors (Costa-Jussà & Farrús, 2015; Costa et al., 2015; Farrús et al., 2010; Gutiérrez-Artacho et al., 2019; Krings, 2001; Laurian, 1984; Schäfer, 2003; Vilar et al., 2006) have proposed different typologies and classifications for errors related to MT and, generally speaking, the majority of these distinguish errors at different levels (spelling, vocabulary, grammar, discourse), divided into subcategories that include, for example, errors relating to concordance, style, confusion in word meaning with various exceptions, etc. Nevertheless, the last decade has seen the appearance of projects seeking to standardise these methods with the objective of facilitating the adaptation of different tasks and language pairs to reduce effort and inconsistencies when developing an error typology (Popović, 2018). This is the case for the Multidimensional Quality Metrics (MQM) frameworks, created by the QTLaunchPad and Dynamic Quality Framework (DQF) project, developed by the Translation Automation User Society (TAUS), which started independently and were integrated in 2014 in "DQF/MQM Error Typology" (Görög, 2014; Lommel, 2018).

2 Methodology

The research methodology is based on the systematic review of specialised literature from 2016 onwards. This methodology consists in the analysis of scientific journals using explicit and rigorous methods that allow the summarising of the results, with the aim of responding to specific research questions (Gough et al. 2012).

The study undertaken endeavours to respond to the following research questions:

1. What MT systems that include the English–Spanish language combination are the most analysed in the specialised literature?
2. What procedures are being applied to measure MT quality in the field of translation?
3. What MT systems are obtaining the best results?

The publications that comprise the study sample, which formed the basis for the analysis carried out, originate from different bibliographical databases to which the queries were put. The methodology therefore has different stages:

1. Selection of bibliographical databases
2. Undertaking of bibliographical queries to determine the final sample
    a. Identification of search keywords—terms, synonyms, variants—.
    b. Creation of the search string—Boolean operators—.
    c. Filtering of results—document type, publication date, amongst others—.
3. Analysis of documents from the sample with the NVivo software package (Release 1.0).

The procedure applied is set out below.

2.1 Database selection

The study sample was obtained from different bibliographical databases, both general and specialised. The typology and number of databases queries permitted the guarantee of an adequate representation of articles on MT published by translation specialists, linguists and experts in related fields.

Searches were carried out on 10 specialised databases:

- *Dialnet* is a bibliographical database focused on Hispanic scientific literature in the spheres of Human, Legal and Social Sciences.
- *Hispanic American Periodical Index (HAPI)* includes bibliographical references on political, economic, social, art and humanities subjects in scientific publications from Latin America and the Caribbean from 1960 onwards.
- *Humanities Full Text:* includes complete texts from the Humanities field.
- *InDICEs* is a bibliographical resource that compiles research articles published in Spanish scientific journals.
- *International Bibliography of the Social Sciences (IBSS)—Proquest* includes bibliographical references from the field of Social Sciences from 1951 onwards.
- *Library and Information Science Abstracts (LISA)—Proquest* includes bibliographical references from Library and Information Science and other related fields.

- *Library, Information Science and Technology Abstracts (LISTA)* is a bibliographical database developed by EBSCO that includes references from the fields of Library and Information Science.
- *Linguistics Collection—Proquest*. This database also includes the *Linguistics and Language Behavior Abstracts (LLBA)* collection and compiles bibliographical references related to all aspects of the study of language.
- *MLA International Bibliography* is a bibliographical database developed by EBSCO that includes references from all fields relating to modern languages and literature.
- *Social Science Database—Proquest* is a database that includes the comprehensive text of scientific and academic documents relating to the Social Sciences disciplines.

The search was also carried out on two of the main multidisciplinary databases:

- *Scopus* is a database edited by Elsevier that includes bibliographical references from scientific literature belonging to all fields of science, including Social Sciences, Art and Humanities. Scopus is, according to the information posted in its official blog,[1] "the largest abstract and citation database of peer-reviewed literature".
- *Web of Science* is a platform managed by Clarivate Analytics that includes references from the main scientific publications in all fields of knowledge from 1945 onwards.

2.2 Keywords and search string

Taking into account the research questions and objectives set by our study, the main search terms were identified, both in Spanish and English, which best represent the concepts involved in our analysis (Table 1). These are:

Table 1 Query keywords

Keywords (Spanish)	Keywords (English)
traducción automática	machine translation, automated translation, automatic translation
evaluación, calidad, errores	evaluation, assessment, quality, errors
español, inglés	Spanish, English

The search string was then created, which was adapted to the characteristics of each of the databases with the objective of recovering all of the relevant documents possible (Table 2):

[1] https://blog.scopus.com/about (last accessed 17 February 2021).

Table 2 Search string

Search string
("machine translation" OR "automated translation" OR "automatic translation" OR "traducción automática") AND (quality OR error* OR evaluat* OR assess* OR evaluación OR calidad OR) AND ((Spanish AND English) OR (español AND inglés))

Table 3 Documents recovered in the databases queried

Database	*N*
Dialnet	21
Hispanic American Periodical Index	0
Humanities Full Text	4
InDICEs	12
International Bibliography of the Social Sciences	28
Library and Information Science Abstracts	33
Library, Information Science and Technology Abstracts	11
Linguistics Collection	55
MLA International Bibliography	2
Scopus	92
Social Science Database	23
Web of Science	37
Total	318

Table 3 shows the number of documents recovered in the different databases queried. The search string used offered 318 results and permitted Scopus to be identified as the database with the greatest index of exhaustivity in relation to the subject of our study.

The results were filtered by applying the following inclusion criteria: (i) publication language, (ii) publication date, (iii) document type and (iv) speciality of the authors.

Thus, the publications in Spanish and English were considered, in line with the language combination contemplated in the objectives of this work. Moreover, it should be borne in mind that the evolution of MT technologies are in constant development to which recently published documents were included (2016 onwards) to guarantee that the MT systems were up to date. Reviews and essays were also rejected, selecting empirical research papers. Lastly, the speciality of the authors was taken into account in a way that at least one author was required to be from the field of languages (translation, linguistics or similar), given that human evaluation and error annotation should be carried out by evaluators withs specific training in the field of translation. The application of these criteria resulted in 19 documents that were all relevant for our study and permitted the identification of the most prominent authors in relation to the subject in question. Hence, a second query on

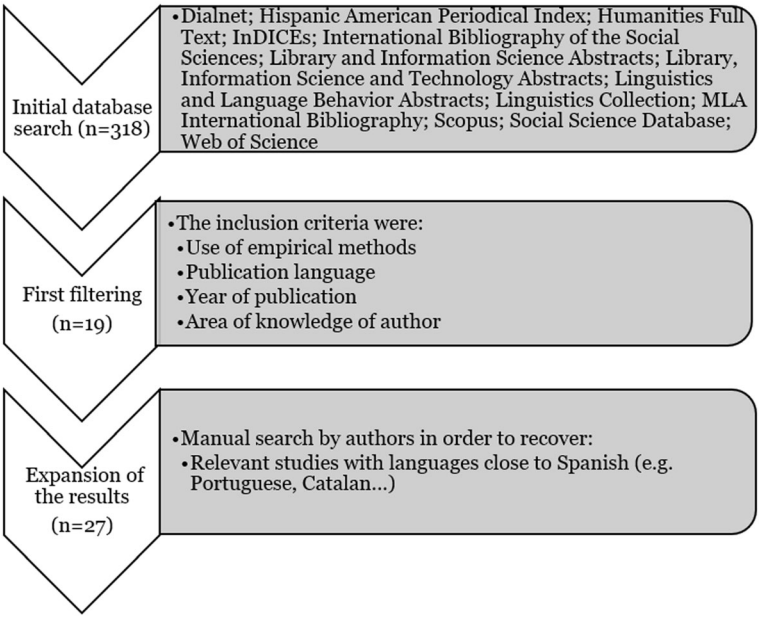


Fig. 1 Document search and selection process

authors in Scopus allowed us to complete the initial sample with a further 8 documents refer to languages close to Spanish—Portuguese and Catalan—that could enrich the study. The final sample was therefore comprised of 27 documents (Fig. 1).

2.3 Qualitative analysis process

The study sample documents were ordered alphabetically and identified by an ID composed by the word *Item* followed by its corresponding number, e.g., Item 12 (Annex 1). All the documents were then stored on a bibliographical reference manager—Mendeley—and it was verified that their metadata were correct. The documents were then exported together with their metadata in order to facilitate a qualitative analysis of the content via the NVivo software package (Release 1.0). The analysis of the content allows for the application of systematic and objective procedures for describing the content of the messages (Bardin, 1996; Mayring, 2000).

To afford a greater rigour to the analysis, two researchers with experience in qualitative research using the NVivo package identified and defined the categories. Being an analysis of content, it was determined that the categories were exclusive, as they were required to have been formed by stable units of meaning (Trigueros-Cervantes et al., 2018; Weber, 1990). The objective of this initial coding was to identify what systems and architectures, evaluation measurements and MT error analysis processes were employed in the different studies, and to determine whether

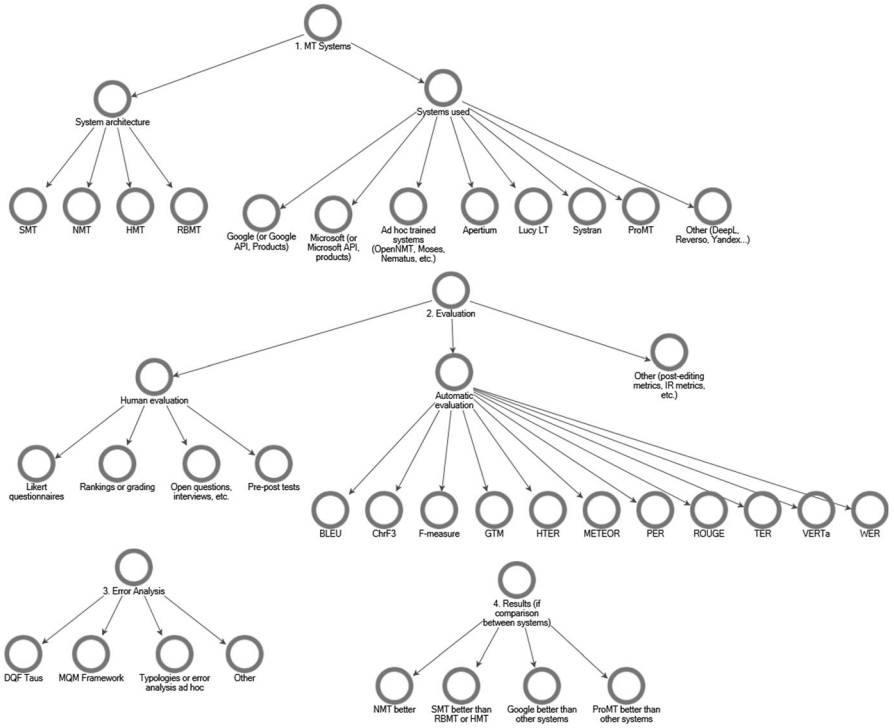
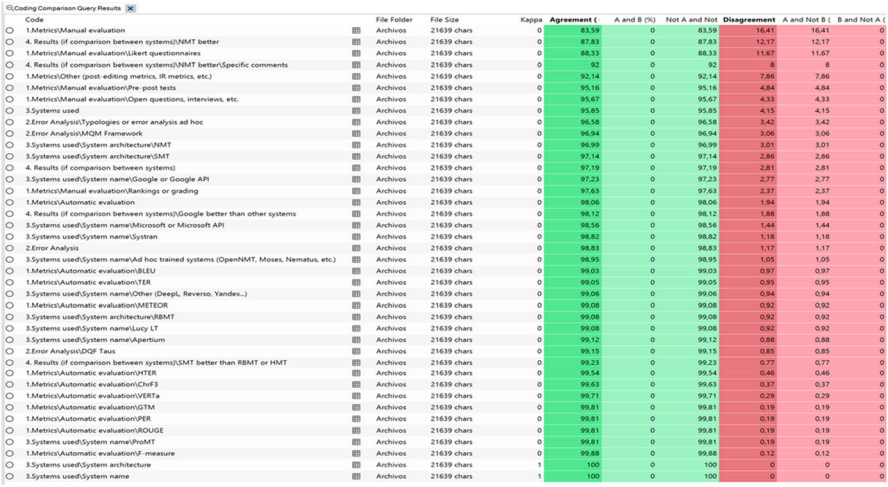


Fig. 2 Category system

there had been comparisons between MT systems or architectures in any of the studies analysed. Following a consultation of experts in qualitative research, a representative sample of the documents was selected (approximately 20%), which were independently coded by both researchers to identify the underlying categories. After agreement was reached, the definitive category system was created (Fig. 2).

All of the documents were subsequently analysed and coded from their in-depth reading in categories or nodes by both researchers, in accordance with the previously established category system. The coding was carried out in both independent NVivo projects that when merged allowed a comparison of coding to be carried out in accordance with the Kappa index, which permits the calculation of the inter-annotator agreement. In this regard, as shown by Fig. 3, there is a very high level of agreement in the large majority of categories. Those categories with a percentage over 10% of disagreement were reviewed and agreement was reached on their coding. The high percentage of agreement is due to the use of very specific concept and exclusive categories, as recommended for this type of analysis.

Once the final categorisation was complete, different coding matrices were generated to carry out a meticulous analysis of all of the coded references in the different categories comprising the object of the study. The use of these matrices permits the exploration of the relationships between different categories and the



Coding Comparison Query Results

Code	File Folder	File Size	Kappa	Agreement (	A and B (%)	Not A and Not	Disagreement	A and Not B (	B and Not A (
1.Metrics\Manual evaluation	Archivos	21639 chars	0	83,59	0	83,59	16,41	16,41	0
4. Results (if comparison between systems)\NMT better	Archivos	21639 chars	0	87,83	0	87,83	12,17	12,17	0
1.Metrics\Manual evaluation\Likert questionnaires	Archivos	21639 chars	0	88,33	0	88,33	11,67	11,67	0
4. Results (if comparison between systems)\NMT better\Specific comments	Archivos	21639 chars	0	92	0	92	8	8	0
1.Metrics\Other (post-editing metrics, IR metrics, etc.)	Archivos	21639 chars	0	92,14	0	92,14	7,86	7,86	0
1.Metrics\Manual evaluation\Pre-post tests	Archivos	21639 chars	0	95,16	0	95,16	4,84	4,84	0
1.Metrics\Manual evaluation\Open questions, interviews, etc.	Archivos	21639 chars	0	95,67	0	95,67	4,33	4,33	0
3.Systems used	Archivos	21639 chars	0	95,85	0	95,85	4,15	4,15	0
2.Error Analysis\Typologies or error analysis ad hoc	Archivos	21639 chars	0	96,58	0	96,58	3,42	3,42	0
2.Error Analysis\MQM Framework	Archivos	21639 chars	0	96,94	0	96,94	3,06	3,06	0
3.Systems used\System architecture\NMT	Archivos	21639 chars	0	96,99	0	96,99	3,01	3,01	0
3.Systems used\System architecture\SMT	Archivos	21639 chars	0	97,14	0	97,14	2,86	2,86	0
4. Results (if comparison between systems)	Archivos	21639 chars	0	97,19	0	97,19	2,81	2,81	0
3.Systems used\System name\Google or Google API	Archivos	21639 chars	0	97,23	0	97,23	2,77	2,77	0
1.Metrics\Manual evaluation\Rankings or grading	Archivos	21639 chars	0	97,63	0	97,63	2,37	2,37	0
1.Metrics\Automatic evaluation	Archivos	21639 chars	0	98,06	0	98,06	1,94	1,94	0
4. Results (if comparison between systems)\Google better than other systems	Archivos	21639 chars	0	98,12	0	98,12	1,88	1,88	0
3.Systems used\System name\Microsoft or Microsoft API	Archivos	21639 chars	0	98,56	0	98,56	1,44	1,44	0
3.Systems used\System name\Systran	Archivos	21639 chars	0	98,82	0	98,82	1,18	1,18	0
2.Error Analysis	Archivos	21639 chars	0	98,83	0	98,83	1,17	1,17	0
3.Systems used\System name\Ad hoc trained systems (OpenNMT, Moses, Nematus, etc.)	Archivos	21639 chars	0	98,95	0	98,95	1,05	1,05	0
1.Metrics\Automatic evaluation\BLEU	Archivos	21639 chars	0	99,03	0	99,03	0,97	0,97	0
1.Metrics\Automatic evaluation\TER	Archivos	21639 chars	0	99,05	0	99,05	0,95	0,95	0
3.Systems used\System name\Other (DeepL, Reverso, Yandex...)	Archivos	21639 chars	0	99,06	0	99,06	0,94	0,94	0
1.Metrics\Automatic evaluation\METEOR	Archivos	21639 chars	0	99,08	0	99,08	0,92	0,92	0
3.Systems used\System architecture\RBMT	Archivos	21639 chars	0	99,08	0	99,08	0,92	0,92	0
3.Systems used\System name\Lucy LT	Archivos	21639 chars	0	99,08	0	99,08	0,92	0,92	0
3.Systems used\System name\Apertium	Archivos	21639 chars	0	99,12	0	99,12	0,88	0,88	0
2.Error Analysis\DQF Taus	Archivos	21639 chars	0	99,15	0	99,15	0,85	0,85	0
4. Results (if comparison between systems)\SMT better than RBMT or HMT	Archivos	21639 chars	0	99,23	0	99,23	0,77	0,77	0
1.Metrics\Automatic evaluation\HTER	Archivos	21639 chars	0	99,54	0	99,54	0,46	0,46	0
1.Metrics\Automatic evaluation\ChrF3	Archivos	21639 chars	0	99,63	0	99,63	0,37	0,37	0
1.Metrics\Automatic evaluation\VERTa	Archivos	21639 chars	0	99,71	0	99,71	0,29	0,29	0
1.Metrics\Automatic evaluation\GTM	Archivos	21639 chars	0	99,81	0	99,81	0,19	0,19	0
1.Metrics\Automatic evaluation\PER	Archivos	21639 chars	0	99,81	0	99,81	0,19	0,19	0
1.Metrics\Automatic evaluation\ROUGE	Archivos	21639 chars	0	99,81	0	99,81	0,19	0,19	0
3.Systems used\System name\ProMT	Archivos	21639 chars	0	99,81	0	99,81	0,19	0,19	0
1.Metrics\Automatic evaluation\F-measure	Archivos	21639 chars	0	99,88	0	99,88	0,12	0,12	0
3.Systems used\System architecture	Archivos	21639 chars	1	100	0	100	0	0	0
3.Systems used\System name	Archivos	21639 chars	1	100	0	100	0	0	0

Fig. 3 Inter-annotator agreement

studies analysed. These matrices were subsequently exported for the creation of tables and graphs in MS Excel.

2.3.1 Clarifications on the coding

Regarding the MT systems used in the empirical studies from the analysed publications, it should be pointed out that in the category design they were grouped by type. As Fig. 2 shows, different resources belonging to the same company were grouped in a single category. This is the case for the categories *Google (or Google API/products)* and *Microsoft (or Microsoft API/products)*, where not only are their machine translations included, but also other resources offering these companies as application programming interfaces (APIs). All those systems that were specifically trained with systems such as Open NMT, Moses and Nematus to carry out the studies were grouped into the Ad hoc *trained systems* category. Finally, those MT systems with coding frequencies under 2 were included in the *Other* category, and this is the case for the DeepL, Reverso and Yandex systems.

Regarding the MT quality evaluation measurements, three large categories were created, which were then divided into subcategories. On the one hand, the automated metrics (BLEU, METEOR, TER, etc.) were classified in the *Automatic evaluation* category and the manual methods such as questionnaires and interviews were classified in *Human evaluation*. On the other hand, the *Other* category was created to include those measurements not directly related to MT quality such as, for example, the post-editing effort (technical, temporal and cognitive) or measurements orientated towards information retrieval.

Finally, for the typologies and error classifications, despite the fact that the MQM and DQF reference frameworks were integrated into a combined typology in 2014, the difference between both has been maintained as the analysed works referenced

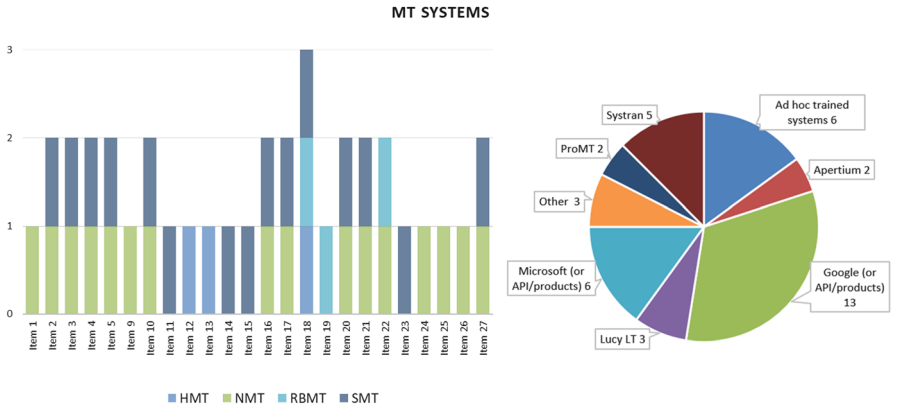


Fig. 4 MT architectures and systems

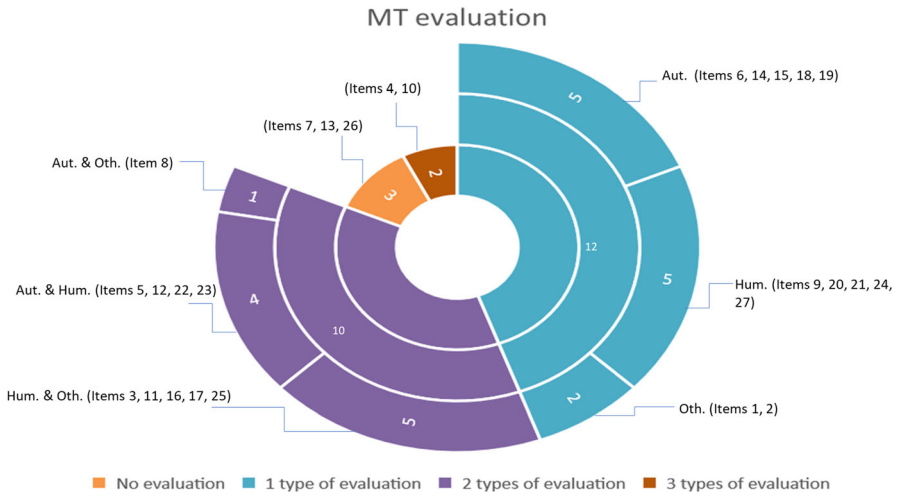


Fig. 5 MT evaluation

them individually. Furthermore, included in the DQF Taus category are those studies employing the DQF platform—despite them not expressly mentioning the error typology—as the aim was to distinguish between those works that used standardised methods and those that did not.

3 Results

3.1 Machine Translation architectures and systems employed

In terms of the architectures employed (Fig. 4) close to 89% of the works—24—used some type of MT system. The 3 remaining studies correspond to Items 6, 7 and 8, and focus on the description and validation of a new MT measurement.

A total of 37 different architectures were employed in the 24 articles, indicating that more than one type of architecture was studied in some of them. Thus, 45% of the studies analysed—Items 2, 3, 4, 5, 10, 16, 17, 20, 21, 22 and 27—used two different types of architecture, whereas only a single study—item 18—used three types. The most used architectures were statistical MT, in 66.7% of the works analysed, and neural MT, in 62.5%. In addition, 41.7% of these works combined both architectures. The use of rule-based or hybrid architectures drops to 12.5% of works in both cases. Finally, it is worth mentioning that in the case of Item 22 it was not possible to accurately define what type of architecture the systems used in the analysis employed. Therefore, although the study publication date was taken as a reference to determine it, given the lack of the study date, the architecture of the systems may have been changed between the analysis and publication dates.

Regarding the MT systems used, the Google translator—or products offered by Google—is the MT system employed by over half of the articles; this is followed by the translators and products offered by Microsoft and MT systems that were specifically trained via Moses, Nematus and OpenNMT, in both cases accounting for 25% of the analysed works.

3.2 Evaluation metrics for MT and error analysis

3.2.1 Evaluation metrics

For the evaluation metrics employed (Fig. 5), again, close to 89% of the works analysed used some type of evaluation metric, either automatic (Aut.), human (Hum.) or other (Oth.) type of metric (measurements related to post-editing effort, information retrieval, etc.). In contrast, Items 7, 13 and 26 did not employ any type of evaluation metric for MT quality, although they did employ error detection and classification methods. 50% of the studies only used one type of evaluation metric. Of these 12 works, five employed automated metrics, five human metrics and two other types of measurement. For the other 50%, between two and three types of evaluation were employed. Of these, 10 studies used two evaluation metrics, of which four combined automated and human metrics, one combined automated metrics with another type of evaluation and five studies combined human metrics with another type. Finally, only two of the works analysed employed the three types of evaluation metric: automated, human and other, in this case focused on the post-editing effort.

3.2.2 Automated evaluation metrics

Automated evaluation metrics (Fig. 6) were employed by 44.4% of the works analysed. On average, these works used three automated evaluation metrics, with Item 12 being the study that used the most—7 metrics—and Items 6, 8 and 19 being those that used the least—1 metric—. The most used metric is BLEU, employed by nine of the 12 works, followed by TER (eight works) and METEOR (seven works); 50% of the works analysed employed these three metrics combined or together with others (Items 4, 10, 12, 14, 15 and 18).

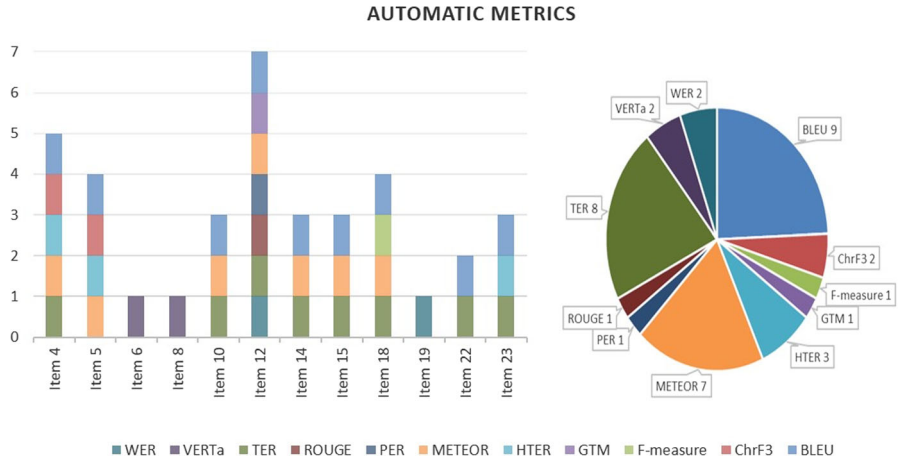


Fig. 6 Automated evaluation metrics

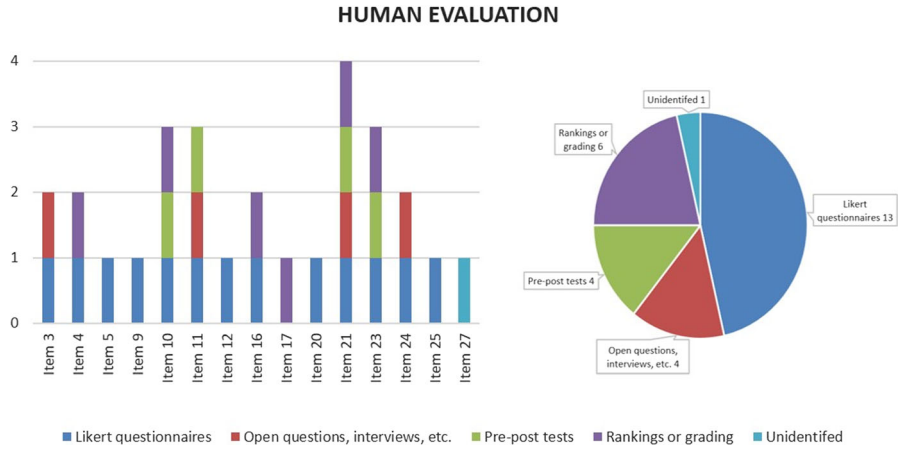


Fig. 7 Human evaluation metrics

Table 4 Classification and analysis of errors

	TAUS DQF framework/guidelines	MQM framework	Typologies or error analysis ad hoc	
Number of works	4	5	5	
Total				14

3.2.3 Human evaluation metrics

Over 55% of the works analysed—15—employed human evaluation metrics (Fig. 7). Of these 15, 86.7% used closed questionnaires with Likert type scales either as a single evaluation method (Items 5, 9, 12, 20 and 15) or combined with others such as ranking or the assignment of scores, pre and post-tests or qualitative methods such as open questionnaires or interviews (Items 3, 4, 10, 11, 16, 21, 23, 24).

3.2.4 Classification and analysis of errors

Regarding the classification and analysis of errors (Table 4), close to 52% of the studies analysed included analyses of errors committed by MT. In this regard, four of the studies (Items 10, 16, 17 and 26) used Dynamic Quality Framework (DQF) and the directives created by the Translation Automation User Society (TAUS), and five (Items 2, 5, 19, 22 and 23) employed the Multidimensional Quality Metrics (MQM) framework, developed by the QTLaunchPad project. In contrast, the five remaining works (Items 1, 4, 7, 13 and 20) developed their own methods of error annotation or typologies.

3.2.5 Combination of evaluation metrics and error analysis

Of the 14 works that included error analysis, all of them apart from 3 (7, 13 and 26) complemented these analyses with automatic, human or other evaluation types. Figure 8 shows that the works that employed the DQF Taus platform, with the exception of Item 26, employed at least human evaluation and other types of measurement. In contrast, all of the articles that employed the MQM framework complemented the error analysis with MT evaluation; in this case, 4 of 5 works that used MQM employed automated metrics together with error analysis. Finally, 3 of the 5 works that developed their own error analyses or typologies combined them with other types of evaluation metrics.

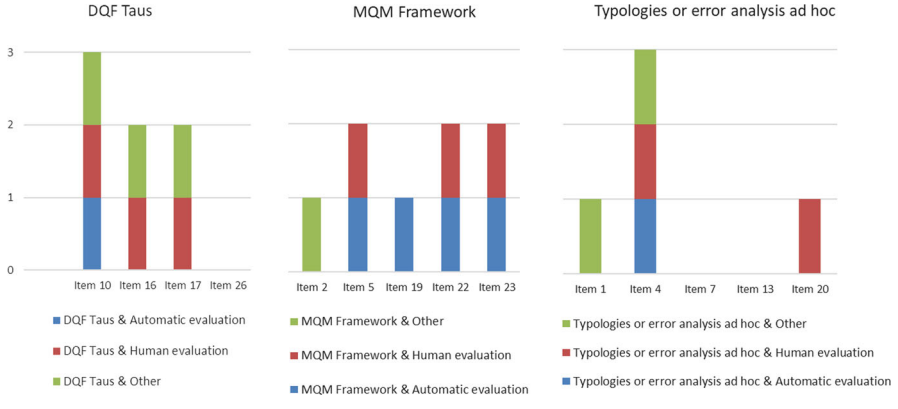


Fig. 8 Combination of evaluation metrics and error classification and analysis

Table 5 Comparison between architectures and systems

	Comparison between architectures		Comparison between systems	
	NMT better than SMT	SMT better than RBMT or HMT	ProMT better than other HMT systems	Google better than other systems
Number of works	10	3	2	3

3.3 Comparison between systems

Regarding the comparison between different systems (Table 5), 59.3% of the studies made comparisons between systems or architectures. Of these, 62.5% (Items 3, 4, 5, 10, 16, 17, 20, 21, 25, 27) established that neural MT was better than statistical, rule-based or hybrid systems. Items 4, 5, 16, 25 and 27, however, offer a number of clarifications regarding the results of the comparison of neural MT with other architectures. Hence, in Item 4, following a description of the results of three different studies, it is concluded that neural MT obtains better results with automated metrics than statistical MT; notwithstanding this, the results of human metrics are not so evident. Item 5 shows that, despite the good general MT compared to statistical MT results, those outcomes in categories for certain errors and for time and post-editing effort are not so evident. This is similar to that described in Items 16, 25 and 27, which highlight a shorter edit distance for neural than for statistical MT; however, their post-editing time is greater. Finally, Item 27 brings attention to the fact that the results for fluency, adequacy and productivity in neural MT were greater in neural than in statistical translation. Conversely, 18.8% of the works—Items 9, 18 and 22—established that statistical MT was better than the rule-based or hybrid kind. It is worth pointing out that in these three cases no comparisons with neural technology systems were made.

As regards MT systems, Items 12 and 13 determined that ProMT obtained better results than another hybrid technology translator—Systran—in terms of automatic and human evaluation, and in error analysis. For their part, Items 1, 9 and 22 highlight the performance of the Google translator compared to other systems. In the case of Items 9 and 22 mention should be made of the fact that Google had not yet adopted the neural system in its machine translation engine.

Finally, given that the majority of the works highlighted the results of neural MT against statistical systems, there was an analysis of what type of evaluation was

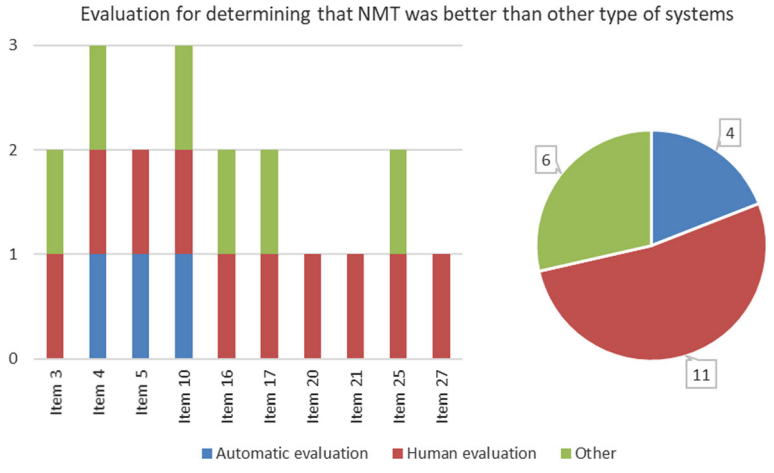


Fig. 9 Types of evaluation employed to determine that NMT obtained better performance

employed in those studies that determined that neural MT was better than other architectures. In Fig. 9 it can be observed that the 10 studies employed human evaluation, and in the case of 3 of them—Items 20, 21 and 27—this was the only evaluation method employed; in 4 of the cases—Items 3, 16, 17 and 25—it was combined with another type of evaluation measurement such as post-editing effort or measurements related to information retrieval effectiveness; in 1 of the cases—Item 5—automatic evaluation was combined with human evaluation; and lastly, in 2 cases—Items 4 and 10—the three evaluation methods were employed.

4 Discussion and conclusions

Following the systematic review of the publications that make up our study sample it is observed, firstly, that neural MT is the predominant model in the current MT scenario. Thus, despite statistical MT being employed in one more study than neural MT, when both architectures were compared the latter obtained better results than the former in all of the studies analysed. These results are along the line of those obtained in one of the main MT evaluation forums (WNT 2015), which confirmed the better performance of neural MT compared to the predominant statistical model up to that point (Bojar et al., 2015). In the same vein, despite the existence of certain clarifications regarding neural MT performance in relation to the order or treatment of long sentences, other later studies confirm this change of paradigm (Bentivogli et al., 2016; Toral & Sánchez-Cartagena, 2017). Moreover, the adoption of neural technologies by the main MT companies such as Google, Systran or Microsoft, among others, confirm that predominance of neural MT in nowadays MT scenario is undoubtable. In relation to the systems employed, Google—or products or APIs offered by Google—is the most used MT system, followed by Microsoft or MT systems that were specifically trained for the objectives of the studies. In this regard, it is noteworthy that despite the current widespread adoption and popularity of DeepL (Schmitt, 2019), only one of these studies employed this machine translator that, furthermore, registered a somewhat lower performance than Google. Therefore, it would be advisable to include DeepL in similar research and to compare its results with those of the nowadays predominant system: Google.

As far as the way of assessing MT is concerned, in spite of the recommendation being to combine both human and automated metrics to obtain the most reliable results possible (Chatzikoumi, 2020; Way, 2018), only 22% of the works analysed combined these two types of measurement, which evidences the fact that the research in the MT field involving translation and language specialists is still scarce and that human evaluation requires a considerable investment of time and resources. Mention should be made of the fact that 2 of these studies, as well as employing both types of evaluation, also utilised other metrics related to post-editing effort. The most used automatic metric is BLEU, which is foreseeable given that it is the most popular automatic metric (Castilho et al., 2018) despite the suggestion on the part of some authors that these metrics may not be adequate for measuring the performances of new neural MT systems, along with the fact that this type of metric does not measure the quality of translations, rather their similarity with reference

translations (Boitet et al., 2006; Castilho et al., 2018; Way, 2018). Thus, given that MT is a constantly evolving sphere, and the possibility of the development of new technologies that go beyond the current paradigms, further research is needed for the development of new MT evaluation methods specifically adapted to these systems.

As regards human evaluation, the majority of the studies employed Likert type questionnaires and rankings; likewise, these results are unsurprising, as they are the most habitual way of carrying out this type of assessment (Castilho et al., 2018; Chatzikoumi, 2020). Nevertheless, it is of note that 4 of the works included open questions or participant interviews, introducing a qualitative analysis approach that is somewhat unusual in this type of evaluation. This type of approach can undoubtedly be interesting, despite its analysis requiring a considerable effort, as on many occasions they offer the possibility of information that is much more enriching and extensive than that provided by the statistical analysis of questionnaire items. Concerning the evaluation techniques employed in the papers which compared neural and statistical MT it is worth mentioning that all of them involved human assessment—along with other kind of procedures or not—. From this fact, it can be concluded that, in the analysed publications, automatic evaluation by itself was not enough to determine which system worked better and, consequently, human evaluations play an essential role in order to determine whether neural techniques have a better performance than statistical ones. On the contrary, the situation changes when there is no comparison between systems or architectures, as in this case, human assessment is included in less than a half of the works. Given that the objective of these works was not to establish comparisons, human involvement does not seem essential to evaluate MT quality with regard to its performance when only one system or architecture is involved. Finally, it should be pointed out that around 52% of the works analysed included error classification and analysis, a fundamental aspect for identifying flaws and improving the performance of MT systems and, in addition, over half of these works carried out this analysis employing standardisation frameworks such as DQF and MQM. Although it is a remarkable amount of works, given the paramount importance of error analysis concerning MT improvement (Popović, 2018), more research is needed including this aspect, specially concerning EN-ES language pair, which necessarily entails the involvement of trained professionals from the translation and languages fields.

The reduced sample size of this study could be its major limitation. However, this limitation is due to the selection criteria involved, wherein two restrictions were applied that affected the final sample substantially. One of these is the chosen language pair—English and Spanish—and the other, which possibly imposed a greater restriction, is the fact that at least one of the authors belongs to the field of translation or similar. Regardless, these results draw attention to the need to involve people with training in translation or related spheres in this type of study given that, as indicated earlier, the most reliable way of evaluating an MT system is by combining automatic and human methods and, as regards the implementation of the latter, there is a need for evaluators who are trained and who have wide knowledge of both languages involved and the translation process. Furthermore, these results could be the product of the traditional rejection of MT on the part of the language

and translation community. Nevertheless, after carrying out this study it can be concluded that MT is a growing area and that, despite its quality not being of the same level as human translation, it is a tool with great potential both for overcoming language barriers and increasing productivity of the translation process.

5 Annex 1: Study sample

ID	References
Item 1	Bowker, L. 2018. Machine translation and author keywords: A viable search strategy for scholars with limited English proficiency? *81st Annual Meeting of the Association for Information Science and Technology*, 13–16. https://doi.org/10.7152/acro.v29i1.15455
Item 2	Carl, M. and Toledo Baez, M. C. 2019. Machine translation errors and the translation process: a study across different languages. *The Journal of Specialised Translation*, 31: 107–132
Item 3	Castilho, S. and Arenas, A. G. 2018. Reading comprehension of machine translation output: What makes for a better read? *EAMT 2018—Proceedings of the 21st Annual Conference of the European Association for Machine Translation*: 79–88
Item 4	Castilho, S., Moorkens, J., Gaspari, F., Calixto, I., Tinsley, J. and Way, A. 2017. Is Neural Machine Translation the New State of the Art? *The Prague Bulletin of Mathematical Linguistics*, 108: 109–120. https://doi.org/10.1515/pralin-2017-0013
Item 5	Castilho, S., Moorkens, J., Gaspari, F., Sennrich, R., Way, A. and Georgakopoulou, P. 2018. Evaluating MT for massive open online courses: A multifaceted comparison between PBSMT and NMT systems. *Machine Translation*, 32(3): 255–278. https://doi.org/10.1007/s10590-018-9221-y
Item 6	Comelles, E., Arranz, V. and Castellon, I. 2017. Guiding automatic MT evaluation by means of linguistic features. *Digital Scholarship in the Humanities*, 32(4): 761–778. https://doi.org/10.1093/llc/fqw042
Item 7	Comelles, E and Atserias, J. 2016. Through the eyes of VERTa. *Procesamiento de Lenguaje Natural,* 57: 181–184
Item 8	Comelles, Elisabet and Atserias, J. 2019. VERTa: a linguistic approach to automatic machine translation evaluation. *Language Resources and Evaluation*, 53: 57–86. https://doi.org/10.1007/s10579-018-9430-2
Item 9	Crespo Miguel, M. and Sánchez-Saus Laserna, M. 2018. Graded Acceptance in Corpus-Based English-to-Spanish Machine Translation Evaluation. In A. Moreno Ortiz andC. Pérez- Hernández (Eds.), *CILC2016 (EPiC Series in Language and Linguistics)* (Vol. 1), pp. 58–70. https://doi.org/10.29007/r819
Item 10	Etchegoyhen, T., Fernández Torné, A., Azpeitia, A., García, E. M. and Matamala, A. 2018. Evaluating Domain Adaptation for Machine Translation Across Scenarios. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 6–15

ID	References
Item 11	Fernández-Torné, A. and Matamala, A. 2016. Machine translation in audio description? Comparing creation, translation and post-editing efforts. *Skase. Journal of Translation and Interpretation*, 9(1): 64–87
Item 12	Gutiérrez-Artacho, J., Olvera-Lobo, M. D. and Rivera-Trigueros, I. 2018. Human post-editing in hybrid machine translation systems: Automatic and manual analysis and evaluation. In A. Rocha, H. Adeli, J. L. Reils and S. Constanzo (Eds.), *Trends and Advances in Information Systems and Technologies*. WorldCIST'18 2018. Advances in Intelligent Systems and Computing (Vol. 745), pp. 254–263. Cham: Springer. https://doi.org/10.1007/978-3-319-77703-0_26
Item 13	Gutiérrez-Artacho, J., Olvera-Lobo, M.-D. and Rivera-Trigueros, I. 2019. Hybrid Machine Translation Oriented to Cross-Language Information Retrieval: English–Spanish Error Analysis. In A. Rocha, H. Adeli, J. L. Reis and S. Constanzo (Eds.), *New Knowledge in Information Systems and Technologies*. WorldCIST'19 2019. Advances in Intelligent Systems and Computing (Vol. 930), pp. 185–194. Cham: Springer. https://doi.org/10.1007/978-3-030-16181-1_18
Item 14	Li, L., Parra Escartín, C. and Liu, Q. 2016. Combining Translation Memories and Syntax-Based SMT Experiments with Real Industrial Data. *Baltic Journal of Modern Computing*, 4(2): 165–177
Item 15	Li, L., Parra Escartín, C., Way, A. and Liu, Q. 2016. Combining translation memories and statistical machine translation using sparse features. *Machine Translation*, 30: 183–202. https://doi.org/10.1007/s10590-016-9187-6
Item 16	López-Pereira, A. 2019. Traducción automática neuronal y traducción automática estadística: percepción y productividad. *Revista Tradumàtica: Tecnologies de la Traducció*, 17: 1–19. https://doi.org/10.5565/rev/tradumatica.235
Item 17	López-Pereira, A. 2018. Determining translators' perception, productivity and post-editing effort when using SMT and NMT systems. In Pérez-Ortiz, Sánchez-Martínez, Esplà-Gomis, Popovic, Rico, Martins, V. den Bogaert and Forcada (Eds.), *EAMT 2018—Proceedings of the 21st Annual Conference of the European Association for Machine Translation*, p. 327
Item 18	Martín-Mor, A. and Sánchez-Gijón, P. 2016. Machine translation and audiovisual products: a case study. *The Journal of Specialised Translation*, 26: 172–186
Item 19	Mercader-Alarcón, J. and Sánchez-Matínez, F. 2016. Analysis of translation errors and evaluation of pre-editing rules for the translation of English news texts into Spanish with Lucy LT. *Revista Tradumàtica: Tecnologies de la Traducció*, 14: 172–186. https://doi.org/10.5565/rev/tradumatica.164
Item 20	Moorkens, J. 2018. What to expect from Neural Machine Translation: a practical in-class translation evaluation exercise. *The Interpreter and Translator Trainer*, 12(4): 375–387. https://doi.org/10.1080/1750399X.2018.1501639
Item 21	Moorkens, J., Toral, A., Castilho, S. and Way, A. 2018. Translators' perceptions of literary post-editing using statistical and neural machine translation. *Translation Spaces*, 7(2): 240–262. https://doi.org/10.1075/ts.18014.moo
Item 22	Ortiz-Boix, C. 2016. Machine Translation and Post-Editing in Wildlife Documentaries: Challenges and Possible Solutions. *Hermēneus. Revista de Traducción e Interpretación*, 18: 269–313
Item 23	Ortiz-Boix, C. and Matamala, A. 2017. Assessing the quality of post-edited wildlife documentaries. *Perspectives: Studies in Translation Theory and Practice*, 25(4): 571–593. https://doi.org/10.1080/0907676X.2016.1245763

ID	References
Item 24	Rossetti, A. and O'Brien, S. 2019. Helping the helpers: Evaluating the impact of a controlled language checker on the intralingual and interlingual translation tasks involving volunteer health professionals. *Translation Studies*, 12(2): 253–271. https://doi.org/10.1080/14781700.2019.1689161
Item 25	Sánchez-Gijón, P., Moorkens, J. and Way, A. 2019. Post-editing neural machine translation versus translation memory segments. *Machine Translation*, 33: 31–59. https://doi.org/10.1007/s10590-019-09232-x
Item 26	Toledo Báez, M. C. 2018. Machine Translation and Post-editing: Impact of Training and Directionality on Quality and Productivity. *Revista Tradumàtica: Tecnologies de La Traducció*, 16: 24–34. https://doi.org/10.5565/rev/tradumatica.215
Item 27	Toral, A., Wieling, M., Castilho, S., Moorkens, J. and Way, A. 2018. Project PiPeNovel: Pilot on Post-editing Novels. In J. A. Pérez-Ortiz andet al. (Eds.), *Proceedings of the 21st Annual Conference of the European Association for Machine Translation*, p. 365

Funding This work was supported by the Spanish Ministry of Science, Innovation and Universities (MCIU) (RTI2018-093348-B-I00, FPU17/00667); the Spanish State Research Agency (AEI) (RTI2018-093348-B-I00); and the European Regional Development Fund (ERDF) (RTI2018-093348-B-I00).

Data availability The references of the literature reviewed are available in Annex 1.

Declarations

Conflict of interest The author declare that there is no conflict of interest.

References

Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *International conference on learning representations*. Retrieved June 11, 2020, from https://arxiv.org/pdf/1409.0473.pdf.

Banerjee, S., & 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 machine translation and/or summarization* (pp. 65–72).

Bardin, L. (1996). *Análisis de contenido*. Akal Ediciones.

Bentivogli, L., Bisazza, A., Cettolo, M., & Federico, M. (2016). Neural versus phrase-based machine translation quality: A case study. In *Proceedings of the 2016 conference on empirical methods in natural language processing* (pp. 257–267).

Boitet, C., Bey, Y., Tomokiyo, M., Cao, W., & Blanchon, H. (2006). IWSLT-06: experiments with commercial MT systems and lessons from subjective evaluations. In *Proceedings of the 2006 international workshop on spoken language translation (IWSLT)* (pp. 23–30).

Bojar, O., In Prague, C. U., Chatterjee, R., Federmann, C., Research, M., Haddow, B., Huck, M., Hokamp, C., Koehn, P., Edinburgh, J., Logacheva, V., Monz, C., Negri, M., Post, M., Scarton, C., & Turchi, M. (2015). Findings of the 2015 Workshop on Statistical Machine Translation. In *Proceedings of the tenth workshop on statistical machine translation* (pp. 1–46).

Castilho, S., Doherty, S., Gaspari, F., & Moorkens, J. (2018). Approaches to human and machine translation quality assessment. In J. Moorkens, S. Castilho, F. Gaspari and S. Doherty (Eds.), *Translation quality assessment*. Cham: Springer, pp. 9–38. https://doi.org/10.1007/978-3-319-91241-7_2

Castilho, S., Moorkens, J., Gaspari, F., Calixto, I., Tinsley, J., & Way, A. (2017). Is Neural Machine Translation the New State of the Art? *The Prague Bulletin of Mathematical Linguistics, 108,* 109–120. https://doi.org/10.1515/pralin-2017-0013

Charoenpornsawat, P., Sornlertlamvanich, V., & Charoenporn, T. (2002). Improving translation quality of rule-based machine translation. *COLING-02 on Machine Translation in Asia, 16*, 1–6. https://doi.org/10.3115/1118794.1118799.

Chatzikoumi, E. (2020). How to evaluate machine translation: A review of automated and human metrics. *Natural Language Engineering, 26*(2), 137–161. https://doi.org/10.1017/S1351324919000469

Cho, K., van Merrienboer, B., Bahdanau, D., & Bengio, Y. (2014). On the properties of neural machine translation: encoder-decoder approaches. In *Eighth workshop on syntax, semantics and structure in statistical translation (SSST-8)*.

Costa, Â., Ling, W., Luís, T., Correia, R., & Coheur, L. (2015). A linguistically motivated taxonomy for Machine Translation error analysis. *Machine Translation, 29*(2), 127–161. https://doi.org/10.1007/s10590-015-9169-0

Costa-Jussà, M. R., & Farrús, M. (2015). Towards human linguistic machine translation evaluation. *Digital Scholarship in the Humanities, 30*(2), 157–166. https://doi.org/10.1007/s10590-015-9169-0

Doddington, G. (2002). Automatic evaluation of machine translation quality using N-gram co-occurrence statistics. In *HLT '02: Proceedings of the second international conference on human language technology* (pp. 138–144).

España-Bonet, C., & Costa-jussà, M. R. (2016). Hybrid machine translation overview. In M. Costa-jussà, R. Rapp, P. Lambert, K. Eberle, R. Banchs, & B. Babych (Eds.), *Hybrid approaches to machine translation. Theory and applications of natural language processing* (pp. 1–24). Cham: Springer. https://doi.org/10.1007/978-3-319-21311-8_1

Farrús, M., Costa-Jussà, M. R., Mariño, J. B., & Fonollosa, J. A. R. (2010). Linguistic-based evaluation criteria to identify statistical machine translation errors. In *EAMT 2010—14th annual conference of the European Association for Machine Translation*.

Görög, A. (2014). Quantifying and benchmarking quality: The TAUS dynamic quality framework. *Revista Tradumàtica: Tecnologies de La Traducció, 12,* 443–453.

Gough, D., Oliver, S., & Thomas, J. (2012). *An Introduction to Systematic Reviews*. SAGE.

Graham, Y., Baldwin, T., Moffat, A., & Zobel, J. (2013). Crowd-sourcing of human judgments of machine translation fluency. In *Proceedings of the Australasian language technology association workshop 2013* (pp. 16–24).

Graham, Y., Baldwin, T., Moffat, A., & Zobel, J. (2015). Can machine translation systems be evaluated by the crowd alone. *Natural Language Engineering, 23*(1), 3–30. https://doi.org/10.1017/S1351324915000339

Gutiérrez-Artacho, J., Olvera-Lobo, M.-D., & Rivera-Trigueros, I. (2019). Hybrid machine translation oriented to cross-language information retrieval: English-Spanish error analysis. In Á. Rocha, H. Adeli, L. Reis, & S. Costanzo (Eds.), *New knowledge in information systems and technologies* (pp. 185–194). Cham: Springer. https://doi.org/10.1007/978-3-030-16181-1_18

Habash, N., Dorr, B., & Monz, C. (2009). Symbolic-to-statistical hybridization: extending generation-heavy machine translation. *Machine Translation, 23*(1), 23–63. https://doi.org/10.1007/s10590-009-9056-7

Han, L. (2016). Machine translation evaluation resources and methods: A survey. ArXiv: Computation and language. Cornell University Library. Retrieved June 11, 2020, from https://arxiv.org/abs/1605.04515.

House, J. (2014). Translation quality assessment: Past and present. *Translation: A multidisciplinary approach* (pp. 241–264). Palgrave Macmillan.

Hunsicker, S., Yu, C., & Federmann, C. (2012). Machine learning for hybrid machine translation. In *Proceedings of the seventh workshop on statistical machine translation*, Montréal, Canada, June 2012.

Hutchins, J. (1995). Machine Translation: A brief History. In E. F. K. Koerner & R. E. Asher (Eds.), *Concise history of the language sciences: From the Sumerians to the cognitivists* (pp. 431–445). Pergamon Press.

Hutchins, J. (2007). Machine translation: A concise history. *Mechanical Translation, 13*(1 & 2), 1–21.

Koehn, P. (2010). *Statistical machine translation*. Cambridge University Press.

Koponen, M. (2016). Is machine translation post-editing worth the effort? A survey of research into post-editing and effort. *JoSTrans: The Journal of Specialised Translation, 25*, 131–148.

Krings, H. (2001). *Repairing texts: Empirical investigations of machine translation post-editing processes*. Kent State University Press.

Lagarda, A. L., Ortiz-Martinez, D., Alabau, V., & Casacuberta, F. (2015). Translating without in-domain corpus: Machine translation post-editing with online learning techniques. *Computer Speech and Language, 32*(1, SI), 109–134. https://doi.org/10.1016/j.csl.2014.10.004

Laurian, A. M. (1984). Machine Translation: What type of post-editing on what type of documents for what type of users. In *Proceedings of the 10th internacional conference on computational linguistics and 22nd annual meeeting on association for computational linguistics* (pp. 236–238). https://doi.org/10.1017/CBO9781107415324.004.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions and re-versals. *Soviet Physics Doklady, 10*(8), 707–710.

Lin, C.-Y., & Hovy, E. (2003). Automatic evaluation of summaries using N-gram co-occurrence statistics. In *Proceedings of the 2003 human language technology conference of the North american chapter of the association for computational linguistics* (pp. 150–157).

Lommel, A. (2018). Metrics for translation quality assessment: A case for standardising error typologies. In J. Moorkens, S. Castilho, F. Gaspari and S. Doherty (Eds.), *Translation quality assessment*. Cham: Springer, pp. 109–127. https://doi.org/10.1007/978-3-319-91241-7_6

Mauser, A., Hasan, S., & Ney, H. (2008). Automatic evaluation measures for statistical machine translation system optimization. In *Proceedings of the sixth international language resources and evaluation (LREC'08)*.

Mayring, P. (2000). Qualitative content analysis. *Forum: Qualitative Social Research, 1*(2), 1–10. https://doi.org/10.1111/j.1365-2648.2007.04569.x

Nießen, S., Och, F. J., Leusch, G., & and Ney, H. (2000). An evaluation tool for machine translation: Fast evaluation for MT research. In *Proceedings of the second international conference on language resources and evaluation (LREC)*.

Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*. https://doi.org/10.3115/1073083.1073135.

Popović, M. (2018). Error classification and analysis for machine translation quality assessment. In J. Moorkens, S. Castilho, F. Gaspari, & S. Doherty (Eds.), *Translation quality assessment* (pp. 129–158). Cham: Springer. https://doi.org/10.1007/978-3-319-91241-7_7.

Schäfer, F. (2003). MT post-editing: How to shed light on the "unknown task"—Experices made at SAP. In *The joint conference of the 8th international workshop of the European Association for machine translation and the 4th controlled language applications workshop* (pp. 133–140).

Schmitt, P. A. (2019). Translation 4.0—evolution, revolution, innovation or disruption? *Lebende Sprachen*, 64(2): 193–229. https://doi.org/10.1515/les-2019-0013

Shaw, F., & Gros, X. (2007). *Survey of machine translation evaluatio*n. Saarbrücken: EuroMatrix. Retrieved June 11, 2020, from http://www.euromatrix.net/deliverables/Euromatrix_D1.3_Revised.pdf.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., & Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In *Proceedings of association for machine translation in the Americas* (pp. 223–231).

Tambouratzis, G. (2014). Comparing CRF and template-matching in phrasing tasks within a Hybrid MT system. In *Proceedings of the 3rd workshop on hybrid approaches to translation (HyTra)* (pp. 7–14).

Thurmair, G. (2009). *Comparing different architectures of hybrid machine translation systems*. In Proceddings of MT Summit XII.

Tillmann, C., Vogel, S., Ney, H., Zubiaga, A., & Sawaf, H. (1997). Accelerated DP based search for statistical translation. In G. Kokkinakis, N. Fakotakis, & E. Dermatas (Eds.), *Proceedings of the fifth European conference on speech communication and technology* (pp. 2667–2670). Rhodos, Greece.

Toral, A., & Sánchez-Cartagena, V. M. (2017). A multifaceted evaluation of neural versus phrase-based machine translation for 9 language directions. In *Proceedings of the 15th conference of the European chapter of the association for computational linguistics* (Vol. 1, pp. 1063–1073).

Trigueros-Cervantes, C., Rivera-García, E., & Rivera-Trigueros, I. (2018). *Técnicas conversacionales y narrativas. Investigación cualitativa con Software NVivo*. Escuela Andaluza de Salud Pública/ Universidad de Granada.

Turian, J. P., Shen, L., & Melamed, I. D. (2003). Evaluation of machine translation and its evaluation. In *Proceedings of MT Summit IX* (pp. 386–393). New Orleans, USA.

Vilar, D., Xu, J., D'Haro, L., & Ney, H. (2006). Error analysis of statistical machine translation output. In *Proceedings of the fifth international conference on language resources and evaluation (LREC'06)* (pp. 697–702).

Way, A. (2018). Quality expectations of machine translation. In J. Moorkens, S. Castilho, F. Gaspari, & S. Doherty (Eds.), *Translation quality assessment* (pp. 159–178). Cham: Springer. https://doi.org/10.1007/978-3-319-91241-7_8.

Weber, R. P. (1990). *Basic content analysis*. SAGE. https://doi.org/10.2307/2289192