Computer Speech and Language 84 (2024) 101582

Contents lists available at ScienceDirect

Computer Speech & Language

journal homepage: www.elsevier.com/locate/csl

Adopting machine translation in the healthcare sector: A methodological multi-criteria review

Marco Zappatore [a,*], Gilda Ruggieri [b]

[a] Department of Engineering for Innovation, University of Salento, via Monteroni, sn, Lecce 73100, LE, Italy
[b] University of Salento, Lecce 73100, LE, Italy

ARTICLE INFO

Keywords:
Machine translation
Automatic translation
Healthcare informatics
Public health
Health communication

ABSTRACT

Background: The recent advances in machine translation (MT) offer an appealing and low-cost solution to overcome language barriers in multiple contexts (e.g., travelling, cultural interaction, digital content localisation). However, highly-technical domains typically exhibiting as long, complex, and specialised texts as the healthcare sector, pose multiple challenges to the effective and risk-safe use of MT.

Methods: To examine how MT nowadays assists written/verbal health communication and because of the existing considerable heterogeneity in technological enablers, language pairs and user groups, training approaches, evaluation processes, and users'' requirements, we propose in this paper a methodological multi-criteria literature review based on current guidelines in computer science research and grounded on a customised configuration of the PRISMA methodology, normally used to perform meta-analyses on clinical trials. The review focuses on language-to-language medical MT, covers the time period January 2015–February 2023, and only refers to articles written in English that are accessible via four scientific online digital libraries. Articles are ranked according to a meta-evaluation scoring method for MT scientific credibility along with a scoring for assessing the scope of MT in healthcare. Finally, a guideline to properly design a study about MT in healthcare is also proposed.

Results: The review included a final set of 58 articles from journals (n = 30) and conference proceedings (n = 28), considering 48 different language combinations. We identified a predominance of English-to-Spanish (n = 19) and English-to-Chinese (n = 16) implementations, mainly tailored to medical staff only (n = 14) or along with patients (n = 12). Included papers addressed clinical communication (n = 21) and health education (n = 37). Unidirectional real-time bilingual MT (n = 24) was the most frequent configuration. MT implementations were dominated by Google Translate (n = 22) often used as baseline, OpenNMT (n = 12), or Moses (n = 11). Training and evaluation approaches varied considerably, while deployment and pre-/post-editing were rarely described with an adequate level of detail.

Conclusion: Even if a significant number of articles reported that the proposed MT solutions were effective when translating (bio)medical texts, only a subset of them complied with rigorous translation quality assessment criteria (e.g., use of automatic metrics better related to human ranking than BLEU or statistical significance testing). Nevertheless, MT can be a valid support/supplement in health communication but to cope with issues in fluency, accuracy, unnatural translations, domain-adequacy, and potential safety risks (for highly-sensitive documents), appropriate MT training is essential, along with in-domain human post-editing. The presence of in-domain training text corpora has also proven to be beneficial. Finally, guidelines

* Corresponding author.
E-mail addresses: marcosalvatore.zappatore@unisalento.it (M. Zappatore), gilda.ruggieri@studenti.unisalento.it (G. Ruggieri).

https://doi.org/10.1016/j.csl.2023.101582
Received 12 October 2022; Received in revised form 4 October 2023; Accepted 24 October 2023
Available online 29 October 2023

about how to design studies on MT in healthcare are also proposed to engage more researchers in this field.

1. Introduction

Machine Translation (MT), intended as the pool of computer-based technologies for translating natural languages without any direct human intervention, is nowadays considerably attracting both research interests and market demands, due to the ground-breaking introduction in the last decade of deep-learning solutions exploiting artificial neural networks (hence the definition of Neural MT or NMT) (Sutskever et al., 2014). NMT rapidly outperformed traditional rule-based MT (RBMT) (Kaji, 1988) and statistical MT (SMT) (Lopez, 2008) approaches in several language pairs, thus also becoming the core technology of various commercial MT systems (Wu et al., 2016; Crego et al., 2016).

The most recent market reports on MT (Technavio, 2022; Global Market Insights, 2022) estimate that its CAGR (compound Annual Growth Rate) will increase of 14.48% from 2021 to 2026, driven by a growing demand in multiple sectors (e.g., automotive, entertainment, e-commerce, and ICT) because of the need for rapid content localisation, services facilitating inter-organisation communication, cost-efficient translation, and solutions improving the customer experience (particularly in non-English speaking countries). The use of MT is also an increasingly appealing solution in many areas where professional translators and interpreters are not readily available and this is especially true in the medical sector, where the aim of providing *"linguistically appropriate care"* (Khoong and Rodriguez, 2022) is considerably challenging.

Typically, from the non-native English-speaking medical student who needs to read a textbook of surgery in its original scientific English version, to the native English-speaking clinician who needs to interact with a foreign patient for whom the health problem adds to the impediment of not knowing the language of her/his doctor, MT apparently promises a convenient solution that facilitates and speeds up communication.

The same holds for more advanced scenarios involving specific technological enablers that could have a relevant integration potential with MT. For instance, the incorporation of MT capabilities in human-to-robot communication already achieved some success with general-domain contents (Manome et al., 2020) and some early applications in the healthcare domain also appeared (Shin et al., 2015). Similarly, with the recent considerable hype surrounding the Metaverse concept and the very first related research works investigating how to promote health and deal with medicine in this virtual environment (Petrigna and Musumeci, 2022), MT could represent a helpful asset to overcome language barriers.

However, MT in the medical sector is very often perceived as still hampered by translation accuracy and reliability issues that could easily lead to safety risks for end users, thus preventing its truly large-scale uptake. The number of potential in-domain applications, along with the typologies of end users, is definitely wide, as it ranges from patient consultations to the dissemination of multilingual public health contents, each one having a different corresponding risk factor (e.g., using MT is definitely riskier when translating a clinical procedure consent rather than a scheduled patient appointment) (Khoong and Rodriguez, 2022).

From a technological standpoint, the number of challenges is also noteworthy. First, it is a matter of fact that MT systems perform differently depending on the source-target language pair, since it cannot be said in advance that a document automatically translated from source language A into target language B1 achieves the same translation accuracy when it is translated into language B2 as well. The same holds for the translation direction, as it cannot be assumed without proper assessment that a given quality level achieved by MT when translating from source language A into target language B is likewise ensured when translating from B into A (Van Der Wees et al., 2019; Intento Inc., 2022). Second, the most recent and best-performing MT models are based on parallel text corpora and, consequently, adequate volumes of specialised texts should be gathered or made available, but the quantification of such volume adequacy is still very debated in terms of text domain, language, and style. Third, MT approaches are suitable to be assessed via multiple translation quality metrics, either manually or automatically computed (Mauser et al., 2008; Lommel et al., 2014), thus making more difficult to compare MT engines (henceforth MTEs) and algorithms against each other.

Even if just briefly sketched, the current landscape of MT in healthcare appears, therefore, extremely variegated. In order to shed light on such complexity, we propose in this paper a methodological literature review (henceforth, MLR) spanning the period January 2015–February 2023, and focusing solely on language-to-language medical MT. We narrowed the scope of the review to journal articles and conference papers, only written in English, which are accessible via four scientific online digital libraries.

Our MLR follows the current guidelines defining how to perform a literature review in the computer science research field, as discussed in Carrera-Rivera et al. (2022), with the aim of offering a rigorous analysis procedure to determine to what extent MT is applied, trained, tested, and perceived in this sector.

More specifically, our analysis is grounded on the meta-evaluation scoring method proposed in Marie et al. (2021) for assessing MT scientific credibility. Furthermore, since that method is valid for any application domain of MT, we also decided to complement it with an additional scoring to assess the scope of MT in healthcare, for every examined MT solution.

Our research goals are the following ones.

1. Investigating what languages are involved and what approaches are applied, depending on specific application scenarios and deployment requirements.
2. Breaking down the current state of the art in terms of adopted MT technologies and training procedures.
3. Examining what are the most referenced target user typologies and language groups when MT is applied to the healthcare domain.

4. Ascertaining how the feasibility of MT to healthcare is evaluated, and with the help of what metrics.
5. Identifying the most recent findings supporting the effective use of MT in healthcare, along with suggested strategies to tackle challenges.
6. Providing end users with useful insights so that more consistent choices can be made when deciding whether adopting (bio)medical MT.
7. Supplying researchers with more rigorous and systematic guidelines to design (bio)medical MT evaluation case studies.
8. Discussing how much the current solutions of (bio)medical MT are actually presented with a perspective clearly oriented to their applicability in this sector, in addition to the traditional perspective of translation quality assessment (conveyed through the classical approaches of manual evaluation and automatic evaluation).

The paper is structured as follows: Section 2 presents the background and related works; the methodology is detailed in Section 3; results are discussed in Section 4, while translation quality assessment scoring, as well as findings about how to use MT in healthcare, and guidelines on how to design case studies in this field are provided in Section 5. Conclusions are drawn in Section 6.

2. Background and related works

2.1. Technological scenario

The shift from traditional RBMT systems to more efficient data-driven approaches (i.e., SMT and, more recently, NMT) has progressively shaped the automatic translation into a viable solution for *"everyday communication needs"* (Yamashita and Ishida, 2006), thanks to various freely-accessible cloud-based systems (e.g., Google Translate, DeepL, Microsoft Bing Translator) alongside professional paid systems (e.g., AWS Amazon Translate, DeepL Pro, Microsoft Language Translator, Systran, IBM Watson, Globalese MT, etc.) (Intento Inc., 2022; Johnson et al., 2017).

In SMT, machine-learning-based statistical methods are trained (at both word and sentence level) on text corpora containing source texts paired (i.e., *aligned*) with already available translations in the target language. Essentially, the *translation model* of a SMT system exploits a probabilistic distribution over strings for a given target language (i.e., *language model*) in order to choose among all possible target strings the one having the highest probability of occurring in the target language (Chen and Goodman, 1999). Since SMT is based on a discrete symbolic representation and learns sentence structures and word collocations directly from text corpora, it suffers from long-distance inter-word dependencies that hamper its output quality (Tan et al., 2020). Moreover, in addition to the need for multiple specialised models (translation, language, word ordering, etc.), SMT requires large text corpora for successful training.[1]

NMT, instead, introduced the ground-breaking innovation of artificial neural networks used to model the entire translation process. This avoids having multiple models and allows end-to-end training approaches. Even if also NMT is a probabilistic data-driven approach, it is based on the assumption that source and target sentences are sequences of words, and that every word has a vectorial representation (with amplitude and direction)[2] (Cho et al., 2014). Therefore, the great majority of NMT models are *sequence-to-sequence encoder–decoder frameworks* that are fed with a source text input, encode it into vectors, estimate how likely the given source sequence corresponds to a target vector (based on pattern-detection performed on the training dataset), and then decode it into the supposedly correct translation. This approach is less hampered by inter-word connections and context awareness and usually provides better translations than SMT (Tan et al., 2020), so that NMT is widely recognised as capable of producing more fluent and idiomatic translations. Nevertheless, also NMT requires large training datasets, thus making it a less-viable solution for low-resource languages (Lakew et al., 2018), and it is sensitive to textual complexity and length. As a consequence, its application to texts featuring long complex technical sentences with rare specialised terms needs appropriate evaluation (Tan et al., 2020).

2.2. MT in healthcare

An end-to-end MT system at no cost is certainly an appealing solution to overcome language barriers, but its reliability and effectiveness depend on many factors. On the one hand, in contexts where safety, reliability, and accuracy do not pose specific high requirements, modern MT represents a successful solution: travelling, communicating with overseas foreign relatives, interacting in multilingual groups, and translating general-domain Web-sourced contents (Kasperė et al., 2021) are just few of the typical application areas that nowadays effectively rely on automatic translation. On the other hand, high-stake domains such as healthcare and law needs for careful use of MT. For instance, the validity and usability of automatically-translated communications in legal contexts is still an open issue (Prieto Ramos, 2015; Wiesmann, 2019) and the concerns about MT misuse in clinical communications are still relevant (Vieira et al., 2021).

From an overall perspective, the risks associated to MT-mediated communication (either written or verbal) frequently come from the absence of shared referring expressions, as MT outputs are inherently not transitive[3] because data-driven MT training are

[1] SMT systems trained with corpora having less than 20k sentences usually provide poor translation results even in closed-domain settings (Costa-jussà et al., 2012).

[2] Since the sequence-to-sequence NMT approach requires persisting the network state for several iterations, not all the available artificial neural network topologies are usable: recurrent neural networks (RNN), gated recurrent unit networks (GRU), and long short term memory networks (LSTM) are those exploited the most.

[3] Automatically translating a text from language A into language B and then back-translating it into language A does not provide the same result.

performed independently.[4] Consequently, speakers might experience difficulties in establishing common grounds where to build their communication upon (Yamashita and Ishida, 2006).

From a more health-oriented standpoint, there exist multiple typologies of medical texts, such as technical documents, regulations, clinical procedures, patient consultations, scientific research articles, drug descriptions, medical equipment manuals, marketing materials, and so on Costa-jussà et al. (2012). Their translation normally requires professional translators. When on-site verbal communication is entailed, in-domain professional interpreters are needed and they are even fewer and more underused than translators, especially in critical contexts. However, when scarcity of professional human resources for translation/interpreting tasks is experienced, this is normally due to a combination of multiple constraints regarding the involved language pair(s), knowledge domain, and setting location (both in time and space). Let us think about how difficult would be to have professional translators/interpreters in contexts (Khoong and Rodriguez, 2022) such as:

- an health emergency occurring in a remote hospital and involving patients (or general practitioners) with limited language proficiency;
- the occurrence of a public health crisis affecting large amounts of people speaking different languages (e.g., natural disasters, epidemics, etc.);
- the need for translating into a low-resource language a recent scientific article (which may deal with a brand new biomedical technique or with the outcomes of a complex clinical trial);
- the necessity of requesting a surgical procedure consent to a not language-proficient patient.

More generally, in some geographical contexts the main challenge to MT in healthcare is posed by the scale of the problem itself. This is especially true in countries where considerable number of additional languages are spoken by language minorities and a non-negligible portion of the population has limited English proficiency (LEP). In the UK, about 68% of the foreign-born population residing in the country for 15 years or more uses English at home, compared with only 28% of those residing there for 0–2 years[5] (The Migration Observatory - University of Oxford, 2019). In the US, according to the interactive government-maintained LEP map, some states reach a 20% of LEP population, with more than 300 different languages spoken (Federal Coordination and Compliance Section (FCS), Civil Rights Division - US Dept. of Justice, 2015).

Therefore, MT is more and more considered either as a replacement for unavailable human resources, as a backup solution or, at least, as the last-standing option when nothing else is at hand. The MT reliability, however, is heavily dependent on the potential safety risks associated to the delivery of wrong MT outputs to intended final users. With source texts such as public health general information, patient-dedicated website contents, and simple patient discharge instructions, MT already proved its suitability (Khoong et al., 2019; Taira et al., 2021), and the same applies for improving doctor-to-patient basic communications (Kaliyadan and Gopinathan Pillai, 2010; Leite et al., 2016). Contrarily, in specific scenarios such as pre-anaesthetic consultations with patients (Beh and Canty, 2015) or when delivering multilingual anticipatory guidance resources (Das et al., 2019), the dangers of inaccurate MT outputs have emerged.

In addition, as it will be clarified in Section 3.5.2, MT is increasingly explored as a novel approach to scenarios that are outside the traditional language-to-language translation, such as cross-lingual medical information retrieval (Rahmani, 2017), semantics identification in clinical reports (Mujjiga et al., 2019), resolution of biomedical acronyms and abbreviations (Kirchhoff and Turner, 2016), text simplification of clinical language (Weng et al., 2019), or multilingual mapping of ICD-10 codes (Falissard et al., 2022).

2.3. Previous studies on MT in healthcare

The challenges for MT in healthcare have been investigated multiple times in scientific literature during the recent years. However, because of the large amount of aspects to take into account, the majority of the research works addressed only specific combinations of settings and scenarios. Typically, a study about MT introduces a custom model (in terms of design, implementation, and evaluation), which is then compared against a baseline of freely available online alternatives (e.g., Google Translate) or different, progressively improved versions of itself. However, training aspects are not always described in details and validation procedures greatly differ in terms of adopted protocols and methodologies. Usually, only a given language pair is considered (as various language pairs would require an equivalent number of MT models) as well as a single specific subfield (e.g., electronic prescriptions, patient guidelines, general information on public health, etc.). Moreover, the current coexistence of multiple technologies (e.g., RBMT, SMT, NMT, etc.) further increases the variety of these studies.

It is also worth to point out that very few literature reviews have addressed this area recently (although some previous works dating back one decade or more are available Costa-jussà et al., 2012). Indeed, while overall MT has been largely examined and multiple literature reviews and surveys have been published since several decades (Kasperė et al., 2021; Van Der Wees et al., 2019; Tan et al., 2020), the same does not apply to the health sector.

In Dew et al. (2018), a large study spanning the 2006–2016 time period was presented: it focused on health communication and examined language pairs, MT approaches, enabling technologies, and quality concerns. However, NMT was not considered (because the very first studies on NMT started to appear only from 2014 onward) and some analysis criteria as important as MT training, target user groups, and the breakdown of evaluation procedures, were not considered.

[4] As anticipated in Section 1, A→B and B→A translations produce different outputs: this is because the corresponding training processes are performed separately.

[5] At the moment of writing, UK Census 2021 data have not been released yet.

A more recent work (Vieira et al., 2021) proposed a structured qualitative meta-analysis of official documents on MT use in medical and legal case studies. The research offered several interesting elements of assessment and referred to numerous real-world examples, but the contemporary focus on medical and legal context limited its scope, preventing a deeper analysis of its implications in healthcare.

Qualitative studies based on interviews to MT users (e.g., doctors, patients, general public) are also available in the literature. In Mehandru et al. (2022), 20 medical staff members were asked about typical language barriers they face at work and about the challenges they encounter in using MT. In Almahasees and Jaccomard (2020), a survey was conducted about the perceived usefulness of the Facebook translation service among Jordanians during the COVID-19 outbreak. These works provide interesting perspectives but lack of a structured multi-criteria analysis approach.

A relevant stream of works focusing on MT and (bio)medical MT comes from WMT, the flagship annual conference on MT research, also connected with similar events in natural language processing. WMT participants are involved in team-based competitions on specific aspects of MT such as translation and evaluation, called *shared tasks*, and are required to implement MT solutions whose output is then evaluated and presented during the conference, while datasets and instructions are provided by the conference organisers. Several tasks are proposed yearly (i.e., *recurrent tasks*), as the *general machine translation task* and the *biomedical translation task*. While the former is considered the main WMT shared task, the latter has been attracting increasing attention from the participants and, consequently, the amount of works dealing with (bio)medical MT presented at the WMT conference is growing. Noteworthy, a summarising article dealing with all the proposed solutions for a given shared task at a given WMT edition is always published at the conference, thus representing a valuable yearly snapshot of the trends in that sector.

Finally, an interesting *research agenda for MT in clinical medicine* was presented in Khoong and Rodriguez (2022), where four interrelated analysis domains were considered (i.e., communication scenarios, target populations, MT algorithms, and translation outcomes) with the aim of improving the research in MT for clinical care.

From such premises, we propose in this paper a MLR compliant with Carrera-Rivera et al. (2022) that extends the analysis criteria presented in Dew et al. (2018) by adding the methodology proposed in Marie et al. (2021), entails the relevant case studies described in Vieira et al. (2021), complies and widens further the suggestions from Khoong and Rodriguez (2022), and considers WMT shared tasks on biomedical translation. The aim is to offer a thorough and structured appraisal of use patterns, technologies, and challenges of MT in the healthcare sector, covering the period January 2015–February 2023, which is not examined at this level of detail in any other similar literature review at this time of writing.

3. Methodology

3.1. Initial assumptions

In evidence-based medicine, systematic literature reviews (SLRs) are very often proposed as an important resource for researchers and practitioners. According to the Cochrane centre, a SLR helps to "[...] *make sense of many kinds of data* [and it is] *a way of summarising the results from all the research that exists about a particular question in an objective, transparent and systematic way.*" (Cochrane Consumers and Communication, 2023). More specifically, a Cochrane SLR (or, for brevity, *Cochrane review*) is entirely focused on the rigorous and reproducible analysis of the research results of a given healthcare protocol/treatment/intervention in a given healthcare-related scenario.[6]

Similarly, SLRs are adopted in the computer science domain (Kitchenham and Charters, 2007) as "[...] *a secondary study with the objective to identify, analyse and interpret all available evidences from primary studies related to a specific research question,* [whose activities involve] *planning, conducting and reporting the review*". Also in this case, being systematic, rigorous and transparent are the pivotal aspects that the SLR must feature, even if SLRs in medicine and computer science exhibit different features.

Since in our work we do actually have neither a therapeutic treatment to examine across the stream of scientific literature related to it nor a purely clinical research question to be addressed, but we rather focus on the application of a computer science topic (i.e., MT) to the healthcare domain, we decided not to adopt the scope of a typical Cochrane SLR but to encompass the methodological rigorousness of the guidelines about SLR in computer science, thus characterising this work as a methodological literature review (MLR).

Therefore, in order to propose such a methodically organised and rigorous review, we complied with the checklist proposed in Carrera-Rivera et al. (2022) and the two-phase pipeline *planning-conducting* described in Kitchenham and Charters (2007). Consequently, the first stage (i.e., *MLR planning*) involves: 1. definition of PICOC[7] keywords and synonyms (Section 3.2); 2. formulation of research questions (already listed at the end of Section 1); 3. selection of DLs (Sections 3.2 and 3.4); 4. definition of inclusion/exclusion criteria (Section 3.5); 5. definition of Quality Assessment and definition of a data-extraction form template (Section 3.6). The second stage, (i.e., *MLR conducting*), involves: 1. applying the DL query string; 2. retrieving the articles returned by the query; 3. refining the study; 4. extracting the data; 5. analysing the final dataset and reporting the analysis outcomes (steps from 1 to 4 are discussed in Section 4, while steps 5 and 6 in Section 5).

Finally, to ensure adequate transparency, the entire dataset of the articles retrieved from the queried DLs (except for the automatically removed duplicates), as well as the full multi-criteria analysis performed on the final subset of selected articles, are made accessible as detailed in Section 3.3.

[6] "[...] *A systematic review summarises the results of available carefully designed healthcare studies (controlled trials) and provides a high level of evidence on the effectiveness of healthcare interventions.* [...] *The aim of a systematic review is to thoroughly assess, by means of a set procedure, the best possible evidence about the effects of a healthcare intervention or treatment in a particular healthcare situation.*" (Cochrane Consumer Network, 2023).

[7] PICOC: Population, Intervention, Comparison, Outcome, Context.

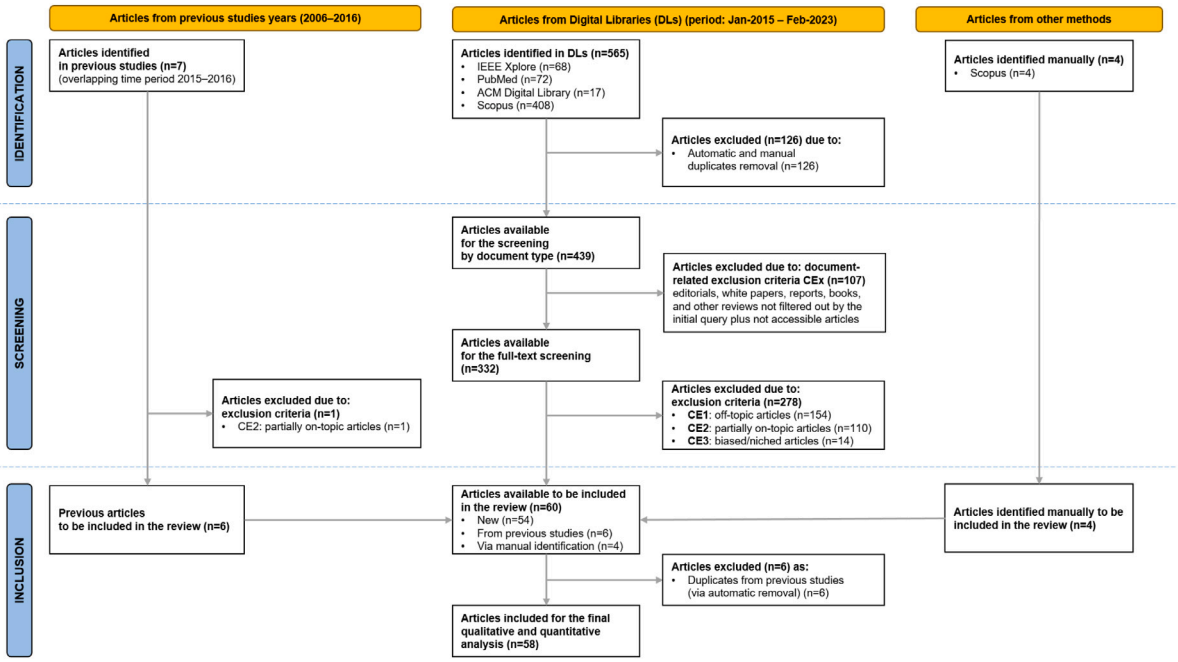

Fig. 1. PRISMA 2020 flowchart instantiated for this methodological literature review (MLR).

3.2. MLR protocol planning, resources, and queries

According to Carrera-Rivera et al. (2022), the first step in a computer science review is to adopt an analysis protocol. To that purpose, we grounded our MLR on the *Preferred Reporting Items for Systematic Reviews and Meta-Analyses* (PRISMA) protocol, in its 2020 updated version (Page et al., 2021a,b). PRISMA was originally defined as a checklist for standardising reviews in the biomedical domain (especially, clinical trials) and for achieving a shared interpretation of their outcomes. However, the same approach can be adapted to reviews addressing different domains with a slight tuning. Essentially, we replaced the original clinical study variables with a pool of evaluation parameters (to guide the full-text analysis described in Section 3.6) and we kept a subset of the original PRISMA 2020 stages, namely *1. identification*, *2. screening*, and *3. inclusion* (Fig. 1).

3.3. Tools and materials

Several tools were used during this work. The initial design phase of the MLR involved the online tool Parsifal (Parsifal, 2017), which supports researchers to perform literature reviews, especially in Software Engineering and Computer Science. More specifically, our MLR benefited from Parsifal when planning and defining the objectives, PICOC, research questions, query string, inclusion and exclusion criteria. The reference management system (RMS) Mendeley (Elsevier, 2022a) was then leveraged to store all the articles retrieved from the queried DLs and automatically remove the duplicates at the end of the first PRISMA 2020 stage. The open-source framework Pentaho Data Integration community edition (Hitachi Vantara, 2023), which allows creating Extract-Transform-Load (ETL) pipelines to extract data from various sources, to transform them into a desired format/structure, and to load them into a target system, was used to perform all the pre-processing steps supporting the results analysis stage. Data visualisation and charts, presented in Section 4, were realised in Flourish (Canva UK Operations Ltd, 2022) and Datawrapper (Datawrapper GmbH, 2022).

The full dataset obtained at the conclusion of the first PRISMA 2020 stage is available as a shared Google Spreadsheet[8], where each article is listed in terms of authors, title, keywords, abstract, publication target, and bibliographical indexes (e.g., DOI, Scopus, ISBN, etc.) if present. In the same file, the result of the full-text screening stage is reported as well. Similarly, another shared Google Spreadsheet[9] is dedicated to gather all the articles included for the final PRISMA 2020 stage, along with the outcomes of the multi-criteria analysis performed on them.

[8] Dataset for PRISMA 2020 Stage 2: full-text screening: https://docs.google.com/spreadsheets/d/1sjN-I1LRJxa_MM0RNyNbPTax0hOA2z4g6DtbYKFeD4U/edit?usp=sharing.

[9] Dataset for PRISMA 2020 Stage 3: multi-criteria analysis of included files: https://docs.google.com/spreadsheets/d/1QuGgmvAvepDpUaczPdUSH5JEBe3E6bOdf5Yiz-Ebl4A/edit?usp=sharing.

3.4. PRISMA stage 1: Identification

As anticipated, we performed several changes on the literature review structure applied in Dew et al. (2018), in terms of selected DLs, query structure, and criteria for the full-text analysis stage, in order to consider the most recent research trends regarding MT in clinical contexts (Khoong and Rodriguez, 2022).

According to the PRISMA protocol, the *identification* requires to define the research query and then to identify the DLs to which that query should be submitted. Therefore, we began by composing the following research query:

```
("Machine Translation" OR "Automatic Translation" OR "Automated Translation" OR "Autonomous Translation")
AND ("Healthcare" OR "Health" OR "Clinical" OR "Medical" OR "Medicine" OR "Patients" OR "Hospital")
```

The query (which is also reported in Table 1, last row) is the intersection of two sets of keywords that entailed the MT and the (bio)medical/healthcare domain, respectively. According to the guidelines presented in Carrera-Rivera et al. (2022), during the MLR planning stage, proper synonyms have to be identified for the query keywords to widen its scope. Therefore, each one of those two sets presented multiple ways to define the same concept (e.g. *"machine translation"*, *"automated translation"*[10]) and also included terms that partially overlap (e.g., *health*, *healthcare*) in order to gather the largest set of available articles. The query was then submitted to four cross-disciplinary DLs[11] (Table 1): IEEE Xplore (Institute of Electrical and Electronics Engineers (IEEE), 2022), ACM Digital Library (Association for Computing Machinery, 2022), PubMed (National Library of Medicine (NLM) - National Center for Biotechnology Information (NCBI), 2022), and Scopus (Elsevier, 2022b).

Those DLs were selected after a preliminary analysis that suggested to discard the Machine Translation Archive as it has not been updated since 2017 (Library of Congress, 2017). Similarly, the Association for Computational Linguistics (ACL) anthology database was not considered as a viable DL since many of its records were already included in the Scopus query output and since the provided search page does not allow to perform searches within certain fields, such as author names or keywords (ACL, 2023).

Google scholar was not included as a searchable source for a two-fold reason: first, it does not allow building complex filtered queries (and this would have prevented narrowing the search to a specific time period as well as to a given subset of publication targets) and, second, because of the long-debated question of how much *grey literature* (undesired in a literature review) a Google Scholar search might intercept (Haddaway et al., 2015).

Overall, we collected 565 new articles, whose largest subset ($n = 408$) was from Scopus and smallest ($n = 17$) from the ACM Digital Library.

Selected DLs were searched for peer-reviewed articles published in scientific journals and conference proceedings, solely written in English, and spanning across the period January 2015–February 2023. Therefore, as it will be also detailed in Section 3.5.4, we excluded non-peer-reviewed papers (e.g., those from the arXiv repository Cornell University, 2022), as well as technical reports, white papers, and editorials (Table 2). As for the time period, we decided not to look for articles published before 2015 as that time range was already investigated by previous reviews (Costa-jussà et al., 2012; Dew et al., 2018; Taylor et al., 2015) and because before 2015 the number of articles dealing with NMT was very low, thus making less meaningful a comparison spanning so many years. The overlapping 2-year range (i.e., 2015, 2016) between our MLR and the one in Dew et al. (2018), gave us a suitable subset ($n = 7$) of previous articles, which were passed on to the next stage.

We also identified manually four articles, since this is allowed by the PRISMA 2020 protocol, as they were not retrieved by the query but considered relevant to the MLR purpose. More specifically, those articles are all annual findings of the WMT shared tasks of biomedical translation.

The RMS Mendeley was used to store the articles and for automatic duplicates removal. Nevertheless, human reading of article abstracts was required to further check the Mendeley group for additional duplicates. In some cases, the same paper was wrongly reported in two different records because of partially diverging metadata: for instance, Melero Nogués (2018) appeared with an English-only title (and DOI) in one DL and with an English-Spanish title (but without DOI) in another DL, thus amounting to two separate records in our raw initial collection. Overall, 126 duplicates were removed automatically.

3.5. PRISMA stage 2: Screening and exclusion criteria

After the PRISMA *Inclusion* stage (i.e., query submission to the DLs and duplicates removal), the remaining papers ($n = 439$) were submitted to the full-text screening stage. We adopted a set of criteria of exclusion (CE for brevity) in order to check: (1) whether they addressed MT applications on clinical communication and/or public health, (2) whether the discussed MT solution was the core element of analysis, and (3) whether the retrieved document type was compliant with the initial assumptions of this MLR.

As it will be thoroughly explained in the subsections from 3.5.1 to 3.5.4, we firstly removed 107 articles depending on their document typology and further 278 articles due to content-related exclusion, thus bringing the initial dataset of DL-retrieved items to 54 elements supplied to the final stage.

Then, we decided to apply the same full-text screening also to the 7 articles inherited from Dew et al. (2018). One item among them, Seligman and Dillinger (2015), even if published in 2015, was actually presented by its own authors as a paper *"of historical*

[10] Required collocations of terms were specified as exact query substrings.

[11] The DLs were queried in terms of ''title-abstract-keyword'' metadata (or their largest subset) via each DL's ''Advanced search'' function, by adapting the initial DL-agnostic query format to the specific query syntax of each DL.

Table 1
Identified Digital Libraries (DLs) and syntax-agnostic search query.

DL	Scope	Metadata[a]	Results
ACM Digital Library	Computer science and information technology	T-A-K	17
IEEE Xplore	Computer science, electrical engineering, and electronics	T-A-K	68
PuBMed	Life sciences and biomedicine	T-A	72
Scopus	Life sciences, social sciences, physical sciences, and health sciences	T-A-K	408

Search Query:
`("Machine Translation" OR "Automatic Translation" OR "Automated Translation" OR "Autonomous Translation")`
`AND ("Healthcare" OR "Health" OR "Clinical" OR "Medical" OR "Medicine" OR "Patients" OR "Hospital")`

[a] Targeted metadata in the search query: T = Title, A = Abstract, K = Keywords.

Table 2
Aspects examined in the MLR and corresponding inclusion/exclusion boundaries for DL queries and full-text article screening.

Aspect	In[a]	Inclusion boundary	Exclusion boundary and criteria (CE)
Publication year	DLQ	Jan. 2015–Feb. 2023	Before Jan. 2015, after Feb. 2023
Language	DLQ,FTS	English	Any language other than English (CEx)
Accessibility	DLQ,FTS	Fully accessible via the queried DL	Articles whose full content is not accessible (CEx)
Article type	DLQ,FTS	Written entirely in English, peer reviewed, presenting empirical studies or simulations, published either on journals or on conference proceedings	Technical reports, white papers, editorials, if not initially filtered out by the starting queries (CEx)
Domain/topic	FTS	The focus is on clinical communication, public health, healthcare; MT as written/verbal communication facilitator, with/without human intervention; comparison of MT effectiveness/accuracy; design/training/implementation of ad-hoc MT solutions	False positive/off-topic (CE1), partially on-topic (CE2), or biased/niched (CE3) articles
MT type	FTS	At least one MT type/algorithm is described	Only Computer-assisted translation is discussed (CE2), without a proper focus on MT

[a] Aspect considered in: DLQ = DL query, FTS = full-text screening.

interest" as the described pilot study for the proposed speech translation system took place in 2011. Therefore, we decided to discard it as too outdated (and we classified that as a CE2-determined exclusion). Eventually, the legacy subset of papers was composed of 6 items: Turner et al. (2015), Chen et al. (2016), Liu and Cai (2015), Taylor et al. (2015), Shin et al. (2015), and Muhaxov et al. (2016). All of them were already present in the subset of new articles obtained by screening the query outputs from the newly selected DLs and we removed them during the last stage.

The remaining four articles identified manually during the first stage were passed to the inclusion stage after the full-text screening confirmed they were relevant to the MLR.

Table 2 introduces all the inclusion/exclusion criteria adopted in this MLR, while the breakdown of articles removed from the dataset depending on the different CE is reported in Table 3.

Furthermore, the following 3 subsections (especially Section 3.5.3) also provide useful insights into other applications of MT in the medical field that have not been included in this MLR.

3.5.1. CE1: off-topic articles

This criterion of exclusion refers to articles not dealing with language-to-language medical MT ($n = 154$). These papers address different research domains even if some query elements misleadingly appear in their title/abstract/keywords and represent the largest subset of criteria-excluded items (Fig. 1). Various types of *false positives* are comprised in this subset.

- **Neither MT source data nor MT target data are languages**: ($n = 6$) articles dealing with automated translation procedures in healthcare that are applied to translate from and into other data types, such as genotypical data (Henriques et al., 2021), *omics*[12] data (Zhang and Guo, 2022) or semantic data (i.e., from HL7 to RDF) (Martinez-Costa and Schulz, 2017).
- **Medical MT is only mentioned**: ($n = 32$) articles presenting broader studies on MT where healthcare is just sampled/mentioned as a case study in the abstract but not described thoroughly. This typology is significantly variegated, as it comprises studies dealing with: advances in natural language processing in general (Hirschberg and Manning, 2015); sentiment analysis (Wołk, 2021); general learning processes in neural networks (Cui et al., 2020); medical text matching (Yu et al., 2021); various strategies for reducing LEP barriers in healthcare (Davis et al., 2019); translator training approaches (Torres-Hostench, 2020); POS tagging (Rajasekar and Udhayakumar, 2020).

[12] E.g., genomics, epigenomics, transcriptomics, proteomics, or metabolomics.

Table 3
Number of articles excluded during the screening stage (breakdown by exclusion criteria and sub-criteria).

Criterion	Sub-criterion	Articles	Sub-total
CE1	Medical MT is only mentioned	32	
	MT not applied to healthcare	30	
	Neither MT source data nor MT target data are languages	6	
	Off-topic	86	154
CE2	Limited amount of details available about medical MT	11	
	Main focus is on Computer-Assisted Translation	2	
	Main focus is on document corpora	11	
	Main focus is on MT pre-/post-editing	4	
	Medical MT source data or target data are not languages	26	
	MT is not primarily applied to the healthcare domain	18	
	MT used for other purposes in healthcare	33	
	Preliminary/partial version of another article	5	110
CE3	Novel solutions/metrics for evaluating MT effectiveness/quality	4	
	Novel solutions/metrics for improving MT performances	9	
	Specific unconventional uses of medical MT	1	14
CEx	Article not accessible	4	
	Document language excluded	3	
	Document type excluded (book/book section)	17	
	Document type excluded (editorial)	66	
	Document type excluded (other)	2	
	Document type excluded (project fact sheet)	1	
	Document type excluded (review/survey)	14	107

- **MT not applied to healthcare**: ($n = 30$): articles where MT is applied to domains that differ from (bio)medical one, as in He et al. (2020, 2021), Zhao et al. (2021), or Handsel et al. (2021).
- **Fully off-topic**: ($n = 86$) articles whose abstract entirely matches the query but actually deals with other topics. For instance, in Xu and Wang (2022), a MT-enabled protein function prediction is presented for automatically translating the descriptive word sequence of such a function into the amino acid sequence of a protein. Similarly, *medicine* and *machine translation* both appear in Jiang et al. (2022) but the paper deals with the challenges of Chinese Pinyin input methods. Another example is given by Oprea et al. (2016), where a platform for monitoring public cloud services is presented and where "personalised *medicine*, real-time speech recognition and *machine translation*" are just mentioned in the abstract to exemplify computing-intensive applications. This group also includes papers describing different types of health (e.g. system health data Leong, 2017).

3.5.2. CE2: partially on-topic articles

This criterion of exclusion identifies articles that are deemed as not relevant enough for this study ($n = 110$), because language-to-language medical MT is only partially covered. A paper not addressing the healthcare domain at all cannot be included in this subset. More specifically, this group encompasses works where:

- **MT is not primarily/uniquely applied to the healthcare domain**: ($n = 18$) articles where other applications of MT are also discussed, as in Tavosanis (2019), Sen et al. (2020), or Semmar and Laib (2018).
- **Medical MT source data or target data are not languages**: ($n = 26$) articles in which MT is applied to the healthcare domain but either its source or its target are not languages. Several combinations of such a kind have been identified during the screening: oral language to sign language[13] (e.g., Agrawal and Urolagin (2020), Veríssimo et al. (2019), Luqman and Mahmoud (2019), or Luqman and Mahmoud (2018)); sign-to-text; image-to-text (e.g., from surgical images to surgical textual instructions Zhang et al., 2021, or from medical image along with questions to text description Ambati and Dudyala, 2018); image-to-image (Amin et al., 2016); audio-to-text (Sadoughi et al., 2018); code-to-code (Hartensuer et al., 2015). Some additional translational mappings involving textual but not linguistic source/target data are reported in Table 4.
- **MT used for other purposes in healthcare**: ($n = 33$) articles whose main focus in healthcare is other than the translational one, such as text/term classification (Joo et al., 2021); document classification (García et al., 2018); cross-lingual medical information retrieval (Rahmani, 2017); semantic annotation (Lin et al., 2020); semantics identification in clinical reports (Mujjiga et al., 2019); semantic text similarity identification (Mutinda et al., 2021); monolingual error correction in clinical documents (Siklósi et al., 2016); binary classifiers training to detect cross-lingual translations of (bio)medical terms (Hakami and Bollegala, 2015) (which is just a preparatory step to MT in biomedicine); data mining (Meng et al., 2022); named entity

[13] It is worth to point out that an article dealing with oral-to-sign language in healthcare is included in CE2, while an article dealing with oral-to-sign language in other domains is included in CE1.

Table 4
Studies addressing specific translational mappings that do not involve MT in traditional language-to-language scenarios (where not otherwise specified, source and output are in English).

Source data	Target data	Reference
SNOMED CT codes	German language	Schulz et al. (2022)
(Layperson) English language	Human Phenotype Ontology (HPO Köhler et al., 2018) terms	Manzini et al. (2022)
Multilingual (French, Italian) death certificates	ICD-10 cause of death code	Almagro et al. (2019) and Falissard et al. (2022)
ICD-9 diagnosis code from the last hospital discharge	ICD-10 cause of death code	Zhu et al. (2022)
Clinical parameters	Categorical data	Dant et al. (2018)
Clinical trial eligibility criteria	Formal queries	Xu et al. (2019)
List of symptoms (Chinese)	List of herbal prescriptions (Chinese)	Wang et al. (2019)
EHR"s discrete variables	Chief complaint text	Lee (2018)

recognition (Schäfer et al., 2022); concept recognition (Afzal et al., 2015); monolingual text generation (Du et al., 2020); post-processing of speech-to-text tasks (Finley et al., 2018); text simplification via acronym and abbreviation resolution[14] (Kirchhoff and Turner, 2016).

- **Main focus on medical MT pre-/post-editing**: ($n = 4$) articles mainly addressing the stages preceding or following the actual MT, along with the corresponding evaluation, as in Liang and Han (2022) or Alvarez et al. (2020).
- **Main focus on document corpora**: ($n = 11$) articles primarily about the creation of corpora covering multiple technical domains and not only healthcare (i.e., partial coverage) (Heafield et al., 2022) or about the creation of medical document corpora without enough details about how the corpus performs in a medical MT task (i.e., partial processing pipeline) (e.g., Kocijan et al. (2020), Ma et al. (2020), or Roussis et al. (2022)). Similarly, articles focusing on terminologies (Ma et al., 2021) or thesauri (to be exploited in medical MT) also fall within this subset.
- **Main focus on CAT**: ($n = 2$) articles whose focus is mainly on computer-assisted translation (CAT) in healthcare instead of MT (de Velde et al., 2015; Guo and Chen, 2021).
- **Limited amount of details**: ($n = 11$) articles representing preliminary studies, extended abstracts, or brief communications (Hill et al., 2022; Nurminen and Koponen, 2020), or only proposing qualitative interview studies (Mehandru et al., 2022).
- **Preliminary/partial versions of other articles**: ($n = 5$) articles excluded because of any more recent or complete version by the same authors already present in the dataset. This is the case of Wołk and Marasek (2015b) and Wołk et al. (2015), excluded as partial versions of Wołk and Marasek (2015a) (eventually included in the final stage of our MLR). The same applies to Lankford et al. (2021), excluded as the most recent (Lankford et al., 2022) was considered in the MLR. When, instead, a significant evolution of an article was identified in another one from the same authors (as in the case of Chen et al. (2016, 2017)), they both were included.

3.5.3. CE3: biased or niche articles

This criterion of exclusion applies to articles where language-to-language medical MT is dealt with an excessive focus on niche aspects ($n = 14$), as propose overly specialised MT aspects in the healthcare sector and do not address the full medical MT pipeline. A paper addressing the healthcare domain only partially cannot be included in this subset. More specifically, to this group have been associated articles proposing:

- **Novel solutions (or metrics) for evaluating medical MT effectiveness**: ($n = 4$) articles addressing medical MT overall quality (i.e., Qin and Liang (2017) and Xie et al. (2021a,b)) or considering MT quality in specific medical translation processes/steps (Wołk et al., 2018).
- **Novel solutions and techniques for improving medical MT performances**: ($n = 9$) articles focusing on data augmentation (An and Long, 2022); fine-tuning of medical MT algorithms for low-resource languages (Yang et al., 2021); use of back-translation (Soto et al., 2022); bilingual word embeddings based dictionaries for translating medical out-of-vocabulary words (OOVs) (Huck et al., 2019); cross-lingual word embedding generation (Chauhan et al., 2021).
- **Specific, unconventional usages**: ($n = 1$) article about how to exploit MT in language rehabilitation of specific categories of patients (Smaili et al., 2022).

3.5.4. CEx: other criteria of exclusion

This final subset of criteria refers to the exclusion aspects that are not content-related and whose full applicability cannot be guaranteed by the query submitted to the selected DLs. For instance, articles not accessible ($n = 4$) that were returned by the

[14] This specific use case was excluded from the final stage of the MLR since it was considered as even more specific than the MT-supported e-prescription simplification approach described in Lester et al. (2021) and Li et al. (2020) as well as of the MT-enabled clinical language translation proposed in Weng et al. (2019) and van den Bercken et al. (2019) (all included in the MLR).

query or articles (n = 3) that passed the language-related filter in the query as they have an English abstract even if the rest of their content is written in a different language. Moreover, the advanced filtering options of the selected DLs not always allowed to exclude the unwanted document types, so that we had to remove manually the following item types: books and book sections (n = 17), conference editorials[15] (n = 66), reviews (n = 14), project fact sheets (n = 1) and other unspecified document formats (n = 2). CEx criteria led to the exclusion of further 107 articles.

3.6. PRISMA stage 3: Inclusion for qualitative and quantitative analysis

The last stage of the PRISMA protocol comprises the human-made full-text quantitative and qualitative analysis of the selected articles.

We reached this stage with a subset of 54 new articles, 4 additional articles selected manually and a further subset of 6 articles from a previous review (Dew et al., 2018), which were already included in the first subset and, therefore, removed via Mendeley as during the *screening* stage. Consequently, the final dataset included 58 articles.

In the following two subsections, the analysis criteria and the quality evaluation scoring are presented. The list of criteria act as the template structure for the data extraction form, while the evaluation scoring represents the quality assessment checklist, as requested in the guidelines for literature reviews in computer science (Carrera-Rivera et al., 2022).

3.6.1. Criteria of analysis for the data extraction phase

In order to fulfil our research goals (Section 1), we defined 32 criteria grouped in seven classes (listed below) to be used throughout this stage for examining the included articles. Our aim is twofold: (1) considering all the aspects a research paper about MT in healthcare should address, and (2) filling the analysis gaps identified in previous literature reviews on the same topic (Section 2).

Class-1. *General aspects on languages, approaches, and scenarios*

i. *Language pair(s)*: all the languages considered in the study.
ii. *Translation direction*: whether MT was unidirectional (e.g., Eng→Spa) or bidirectional (e.g., Eng↔Chi).[16]
iii. *Translation approach*: monolingual (i.e., source and target language coincide), bilingual (i.e., one source to one target language), or multilingual (i.e., one source language to many target languages) (Hutchins, 1995).
iv. *Translation timing*: either real-time translation or pre-translation (e.g., fixed-phrase, example-based, etc.).
v. *Translation type*: text-to-text, speech-to-text, text-to-speech, or speech-to-speech[17] translation.
vi. *Application scenario*: first-level categorisation that identifies to what field the study is applied (i.e., clinical communication or health education).
vii. *Application type*: second-level partitioning that specifies further the application scenario (e.g., different types of clinical communication).

Class-2. *Technological features of MT*

i. *MT approach*: the adopted MT solution (e.g., neural, statistical, etc.).
ii. *MT engine type*: whether the MTE is free, proprietary, or customised.
iii. *MT implementation*: the specification of the technological enabler used to implement the proposed MT solution (e.g., OpenNMT, Moses, etc.) or used in the MTE under examination/comparison.
iv. *Comparison of MT engines*: whether two or more MTEs are compared in the study.

Class-3. *MT training procedures*

i. *Description*: whether the adopted MT training process is described with an adequate level of details.
ii. *Vocabularies/Dictionaries*: any specific in-domain or general-domain dictionary/vocabulary used to train the MT solution(s) considered in the article.
iii. *Text corpora*: any specific in-domain or general-domain text corpus used for training purposes.

Class-4. *Study population and experiment settings*

[15] It is noteworthy that the majority of them came from the Scopus query.

[16] Henceforth, for the sake of brevity, languages are reported according to the ISO 639/2B international standard that defines three-letter English identifiers for all known human languages (Hutchins, 2017). Therefore, *Eng* stands for English, *Spa* for Spanish, *Chi* for Chinese, and so on.

[17] On a more formal standpoint, *text-to-text* automatic translation should be defined as *Machine Translation* (MT), while *speech-to-speech* automatic translation should be considered as *Machine Interpreting* (MI) (Vieira et al., 2021). However, MI is essentially a context where a core set of MT functionalities is enriched with speech recognition and speech synthesis capabilities. Therefore, we decided to refer to both the contexts as MT. For the same reason, we decided to include them both in our MLR.

i. *Target user type*: typologies of users considered in the study (e.g., patients, doctors, researchers, medicine students, nurses, etc.).
ii. *Language proficiency group*: any specific language proficiency group the target users belong to, such as LEP, non-English speaking, Spanish-only speaking, etc.
iii. *Wrong-translation risk level*: estimated level of potential safety risk caused by wrong MT outputs delivered to final users, depending on the source documents: *Low* (e.g., patient's basic communication needs), *Medium* (e.g., bedside interactions, general health guidelines, general public health information, etc.), *High* (e.g., consent requests for clinical procedures, medicine dose, etc.), or *Variable* (i.e., when the proposed MT solution is evaluated on multiple source document types).

Class-5. *Study typology and materials*

i. *Deployment period*: availability period of the MT solution.
ii. *Deployment stage*: maturity level of the MT solution (e.g., simulation, prototype, pilot study, stable implementation, etc.).
iii. *Source document type*: typology of documents against which the MT solution was tested/validated (e.g., public health materials, (bio)medical research articles, drug prescriptions, etc.).
iv. *Testing dataset size*: amount of source documents used in the testing/validation phase of the MT solution.
v. *Deployment geographical area*: regional/national context where the MT solution is tested/adopted.

Class-6. *Evaluation/validation procedures*

i. *Approach/Method*: the way the proposed MT solution is evaluated (e.g., qualitatively, quantitatively, or hybrid).
ii. *Description*: whether the evaluation process is described with an adequate level of detail.
iii. *Manual evaluation*: any manual evaluation procedure mentioned in the article, along with corresponding metrics (e.g., translation quality assessment, translation fluency, etc.).
iv. *Automatic evaluation*: any automatic evaluation procedure mentioned in the article, along with corresponding metrics (e.g., BLEU, TER, METEOR, etc.).
v. *Pre-editing*: whether pre-editing is considered in the study.
vi. *Post-editing*: whether post-editing is considered in the study.
vii. *Number/type of validators*: number and/or typology of human evaluators/validators involved in the study.

Class-7. *Findings reported in the study*

i. *On pre-editing*: any finding about pre-editing reported in the examined article.
ii. *On post-editing*: any finding about post-editing reported in the examined article.
iii. *Overall*: conclusive findings over the usefulness/effectiveness of the MT solution for the target scenarios/users, as reported by the authors of the examined article.

With so many criteria to manage, it is important to ascertain how they map on every available article, in order to identify what papers address only few analysis parameters and what parameters are considered in few papers only. Therefore, a categorical data heatmap matrix was realised (Fig. 2), having papers on rows and criteria on columns (coloured depending on the class every criterion belongs to). This chart typology is extremely effective to reveal any patterns in data (Evergreen, 2019): we used a two-grade matrix, where a coloured cell means a given parameter is discussed in a given article and a blank cell means it is not. At the bottom and rightmost sides of the matrix, coverage percentages are proposed per rows and columns, with a red-colour gradient (i.e., the most intense the red, the fewest cells on that row/column have a value). Row headers report the first author and publication year of each paper, along with a colour code (green for conference papers and amber for journal articles). The following section will examine thoroughly the full-text analysis results.

3.6.2. Quality assessment scoring

As introduced in Section 2, evaluating the translation quality of MT solutions is a fundamental step to ascertain the effectiveness reached by a given solution. However, as it will be shown in Section 4.6 and Table 11, a significant variety of approaches there exists, ranging from manual evaluations performed by in-domain human experts or professional translators to automatic metrics such as BLEU (Papineni et al., 2001). Such heterogeneity is accompanied by the absence of shared guidelines about how to compare different MT solutions in terms of a pre-agreed quality assessment score that would make the evaluation more rigorous and scientifically sound.

In the framework of this MLR, we decided to refer to the work by Marie et al. (2021), where a meta-evaluation scoring to assess the scientific credibility of MT solutions is proposed. The scoring requires to consider the following four aspects. Each aspect is addressed via a corresponding question whose score is 1 if affirmative (or not applicable), 0 if negative.

1. Whether human evaluation is performed or an automatic metric other than BLEU is used to better correlate with human judgement.

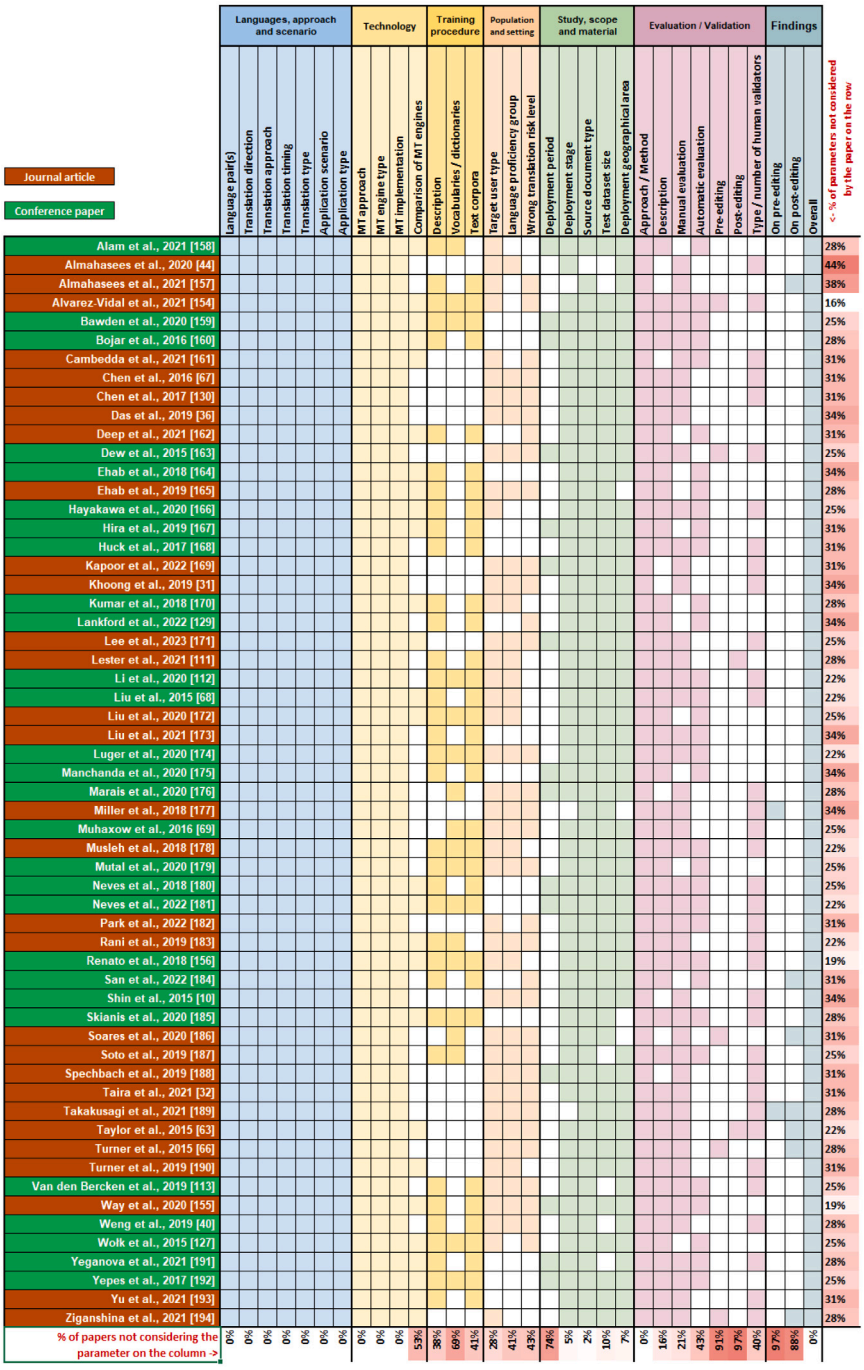

Fig. 2. Heatmap of articles and parameters of analysis.

2. Whether any statistical significance test is performed, so that coincidental comparison outcomes are avoided.
3. Whether automatic metrics are always computed or, if copied, their comparability is guaranteed by SacreBLEU-like metrics, in order ensure that the score is computed in the same way.
4. Whether the same datasets are used when different MT systems are compared, because training, validation, and testing phases in MT are dataset-dependent.

As it can be seen, aspects from 1 to 3 refer to MT quality, while aspect 4 is specific of works where MT solutions are compared.

Table 5
Quality assessment scoring and corresponding questions (each question can be given a score from 0 to 1).

Category	QA question
Domain-agnostic (Marie et al., 2021) (MT Quality Assessment, MTQA)	**MTQA1**: Is a metric that better correlates with human judgement than BLEU used or is a human evaluation performed?
	MTQA2: Is statistical significance testing performed?
	MTQA3: Are the automatic metrics computed and not copied from other work(s)? If copied, are all computed through tools that guarantee comparability (e.g., SacreBLEU)?
	MTQA4: If different MT solutions are compared, are the same pre-processed data exploited for training, validating, and testing? (if not applicable, +1 point by default)
In-domain (Scope of MT in Healthcare, MTSH)	**MTSH1**: Are in-domain datasets, either corpora (+0.5 pts.) or vocabularies (+0.5 pts.) used for training purposes?
	MTSH2: Are the in-domain risks related to wrong translations properly (+1 pt.) or partially (+0.5 pts.) considered and discussed?
	MTSH3: Is a specific and relevant research question for the biomedical field explicitly (+1 pt.) or partially (+0.5 pts.) formulated and addressed?

In our MLR, every article included in the final PRISMA stage was therefore evaluated in terms of these 4 criteria. In addition, since the proposal by Marie et al. was domain-agnostic, we decided to introduce three further questions to determine the scope of MT in the healthcare domain for each article. The questions relate to (1) the availability of in-domain vocabularies and text corpora for training, (2) the presence of a proper assessment of the risks of wrong MT output in healthcare, and (3) the existence of an adequate research question specifically related to the healthcare domain. Also in this case, the attributed score is 1 if the answer is fully affirmative, 0.5 if the examined article satisfies only partially the question, or 0 otherwise. All the identified quality assessment questions are listed in Table 5.

Finally, as it will be clarified in Section 5, the selected rigorous quality assessment scoring can be compared against the findings reported in each article about the quality of the MT solution proposed or examined, thus highlighting any potential mismatch or inconsistency.

4. Results

4.1. Overall considerations

If we examine Fig. 2, we can immediately notice that every article[18] discussed all Class-1 parameters, three out of four Class-2 parameters, and almost all but one Class-5 parameters. More specifically, the following overall quantitative insights were also achieved:

- 53% articles did not propose any comparison between MTEs (Class-2).
- More than a third of articles did not discuss training procedures (Class-3), with a peak of 69% articles not considering dictionaries for training.
- In Class-4 parameters, the risk of wrong MT translations was considered only in 57% of papers.
- Study scope and materials (Class-5) were usually described in articles except for a quantification of the deployment period, which was present in 26% articles only.
- Evaluation procedures showed the most variegated situation, even if the adopted approach was always described. Only 21% articles did not present a manual evaluation while 43% articles did not rely on automatic evaluation metrics. Noteworthy, every article discussed at least one evaluation approach.
- Overall findings (Class-7) were always discussed.
- Pre-/Post-editing use (Class-6) and related findings (Class-7) were scarcely considered (for both cases, in less than 10% articles).
- On average, every article addressed nearly 70% parameters, with Álvarez Vidal et al. (2021), Way et al. (2020), and Renato et al. (2018) having less than 20% missing parameters, with Almahasees and Jaccomard (2020) and Almahasees et al. (2021), which are interestingly from the same authors, having 40% or more missing parameters.

Journal articles ($n = 30$) slightly exceeded conference papers ($n = 28$). As for the time-distribution of included articles (except for the year 2023 that is represented by just one journal article since two months only were considered in the query), the lowest number of included publications (i.e., two conference papers and one journal article) was reported in 2016 and 2017, while the highest number occurred in 2020 (i.e., eight conference papers and four journal articles). The year 2021 is the one having the highest number of journal articles included ($n = 9$), followed by 2019 ($n = 6$).

[18] All Figures and Tables presented in this section will report only the first author's surname and the publication year of every article, for the sake of visual clarity.

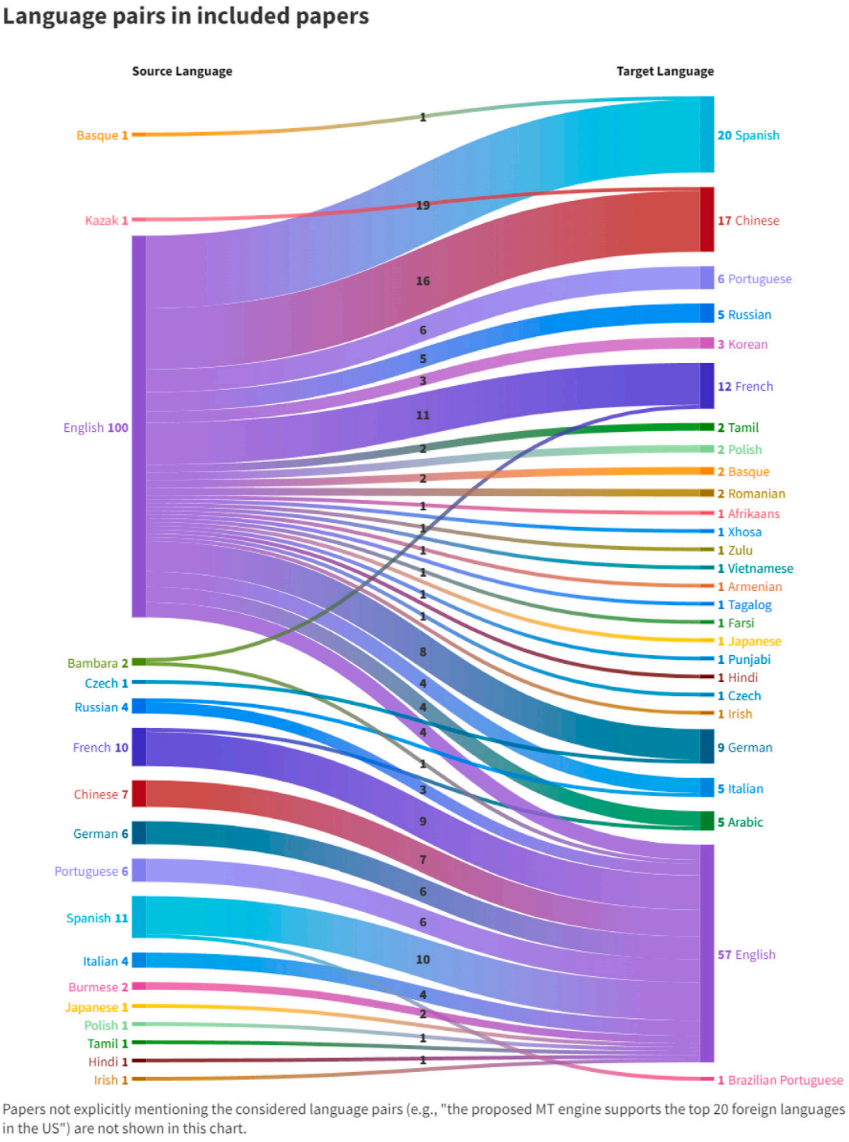

Fig. 3. Source/Target languages.

4.2. Languages, approaches, and scenarios

Class-1 parameters were addressed in all articles: they are summarised in Table 6.

The alluvial chart in Fig. 3, very useful to plot trends and magnitudes of categorical variables over specific process stages or time phases (Evergreen, 2019), depicts source and target languages of the examined articles as the starting and ending phase of every arc. The arc width represents the amount of articles associated with that specific combination of source and target language.

It is worth to point out that in the 58 included papers 48 different language pairs were addressed, across an overall amount of 160 language-to-language implementations (we are excluding from this count the implementations where the addressed language pairs were not explicitly mentioned). This is due to the fact that, in many articles, more than one MT solution was considered, sometimes with different language combinations. Consequently, the total amount of source (respectively, target) languages reported on the left (respectively, right) vertical axis of the alluvial chart in Fig. 3 is higher than the total amount of articles examined in the MLR. Therefore, if we focus on language-to-language combinations, we have that English was the source language in 100 implementations, thus representing the predominant scenario, while it was the target in 57 cases. Spanish was the second most considered target language ($n = 20$), immediately followed by Chinese ($n = 17$).[19]

[19] Very rarely the typology of Chinese language is specified in the articles. Therefore, we provided here a single value comprising all of them.

Table 6
Overview of selected papers: languages, translation parameters and application scenarios.

Article	Languages[a]	Translation				Application	
		Dir.[b]	Approach[c]	Timing[d]	Type[e]	Scenario	Type
Alam et al. (2021)	Eng→Chi,Fre,Rus,Kor; Cze→Ger	B	MuL	RT	TTT	Health educ.	(bio)medical termbase
Almahasees and Jaccomard (2020)	Eng→Ara	U	BL	RT	TTT	Health educ.	Public health
Almahasees et al. (2021)	Eng→Ara	U	BL	RT	TTT	Health educ.	Public health
Álvarez Vidal et al. (2021)	Eng→Spa	U	BL	RT	TTT	Health educ.	(bio)medical text
Bawden et al. (2020)	Eng↔Chi,Fre,Ger,Ita,Por, Rus,Spa; Eng→Baq	B	MuL	RT	TTT	Health educ.	(bio)medical text+term.
Bojar et al. (2016)	Eng↔Fre,Por,Spa	B	MuL	RT	TTT	Health educ.	(bio)medical text
Nunzio et al. (2021)	Rus→Ita	U	BL	RT	TTT	Health educ.	(bio)medical text
Chen et al. (2016)	Eng→Spa,Chi	U	MuL	RT	TTT	Health educ.	(bio)medical text
Chen et al. (2017)	Eng→Spa,Chi	U	MuL	RT	TTT,STT	Health educ.	(bio)medical text
Das et al. (2019)	Eng→Spa + other 19	U	MuL	RT	TTT	Clinical comm.	Patient guidance
Deep et al. (2021)	Eng→Pan	U	BL	RT	TTT	Health educ.	Public health
Dew et al. (2015)	39 as in MS Bing API	B	MuL	RT	TTT	Health educ.	Public health
Ehab et al. (2018)	Eng→Ara	U	BL	PT(eb)	TTT	Health educ.	(bio)medical text
Ehab et al. (2019)	Eng→Ara	U	MoL	RT	TTT	Health educ.	(bio)medical text
Hayakawa and Arase (2020)	Eng→Jap	U	BL	RT	TTT	Health educ.	Public health
Hira et al. (2019)	Eng↔Fre	B	BL	RT	TTT	Health educ.	(bio)medical text
Huck et al. (2017)	Eng→Ger	U	BL	RT	TTT	Health educ.	(bio)medical text
Kapoor et al. (2022)	Spa→Eng	U	BL	RT	TTT	Clinical comm.	Patient consultation
Khoong et al. (2019)	Eng→Spa,Chi	U	MuL	RT	TTT	Clinical comm.	Patient guidance
Kumar et al. (2018)	Eng↔Tam	B	BL	RT	TTT	Health educ.	(bio)medical text
Lankford et al. (2022)	Eng↔Gle	B	BL	RT	TTT	Health educ.	Public health
Lee et al. (2023)	Eng→Spa,Chi(cnm)	B	MuL	RT	STS	Clinical comm.	Patient consultation
Lester et al. (2021)	Eng→Eng	U	MoL	RT	TTT	Clinical comm.	e-prescriptions
Li et al. (2020)	Eng→Eng	U	MoL	RT	TTT	Clinical comm.	e-prescriptions
Liu and Cai (2015)	Eng→Spa	U	BL	RT	TTT	Clinical comm.	Patient history
Liu et al. (2020)	Eng→Chi	U	BL	RT	TTT	Health educ.	(bio)medical text
Liu and Huang (2021)	Eng↔Chi	B	BL	RT	TTT	Health educ.	(bio)medical text
Luger et al. (2020)	Bam→Fre,Eng	U	MuL	RT	TTT	Health educ.	Public health
Manchanda and Grunin (2020)	Eng→Spa	U	BL	RT	TTT	Clinical comm.	HC enterprise comm.
Marais et al. (2020)	Eng→AF,Xho,Zul	U	MuL	RT	TTT,STS	Clinical comm.	Patient consultation
Miller et al. (2018)	Eng↔Spa	B	BL	RT	TTT	Clinical comm.	Patient guidance
Muhaxov et al. (2016)	Kaz→Chi	U	BL	RT	TTT	Clinical comm.	Patient history
Musleh et al. (2018)	Hin↔Eng	B	BL	RT	TTT	Clinical comm.	Patient consultation
Mutal et al. (2020)	Fre→Eng	U	BL	RT,PT(fp)	STS	Clinical comm.	Patient consultation
Neves et al. (2018)	Eng↔Chi,Fre,Ger,Por,Spa; Eng→Rum	B	MuL	RT	TTT	Health educ.	(bio)medical text
Neves et al. (2022)	Eng↔Chi,Fre,Ger,Ita,Por, Rus,Spa	B	MuL	RT	TTT	Health educ.	(bio)medical text
Park et al. (2022)	Eng↔Ger	B	BL	RT	TTT	Health educ.	(bio)medical termbase
Rani et al. (2019)	Eng→Tam	U	BL	RT	TTT	Clinical comm.	Patient consultation
Renato et al. (2018)	Spa→Por	U	BL	RT	TTT	Health educ.	(bio)medical termbase
San et al. (2022)	Eng↔Bur	B	BL	RT	TTT	Clinical comm.	Patient consultation
Shin et al. (2015)	Eng→Kor	U	BL	RT	STS	Clinical comm.	Patient consultation
Skianis et al. (2020)	Eng→Fre	U	BL	RT	TTT	Health educ.	(bio)medical termbase
Soares et al. (2020)	Eng→Fre	U	BL	RT	TTT	Health educ.	(bio)medical text
Soto et al. (2019)	Baq→Spa	U	BL	RT	TTT	Clinical comm.	Patient history
Spechbach et al. (2019)	Fre→Ara	U	BL	PT(fp)	TTT,STT	Clinical comm.	Patient consultation
Taira et al. (2021)	Eng→Spa,Chi,Vie,Tgl,Kor, Arm,Ira	U	MuL	RT	TTT	Clinical comm.	Patient guidance
Takakusagi et al. (2021)	Jap→Eng	U	BL	RT	TTT	Health educ.	(bio)medical text
Taylor et al. (2015)	24 lang →Eng, →Spa,Chi	U	MuL	RT	TTT	Health educ.	(bio)medical text
Turner et al. (2015)	Eng→Chi	U	BL	RT	TTT	Health educ.	Public health

(*continued on next page*)

Table 6 (*continued*).

Turner et al. (2019)	Eng→Spa,Chi	U	MuL	RT,PT(fp)	TTT	Clinical comm.	Patient consultation
van den Bercken et al. (2019)	Eng→Eng	U	MoL	RT	TTT	Health educ.	(bio)medical text
Way et al. (2020)	Fre,Ger,Ita,Spa↔Eng	B	MuL	RT	TTT	Health educ.	Public health
Weng et al. (2019)	Eng→Eng	U	MoL	RT	TTT	Clinical comm.	Patient history
Wołk and Marasek (2015a)	Eng↔Pol	B	BL	RT	TTT	Health educ.	Public health
Yeganova et al. (2021)	Eng↔Chi,Fre,Ger,Ita,Por, Rus,Spa; Eng→Baq	B	MuL	RT	TTT	Health educ.	(bio)iomedical termbase and text
Jimeno Yepes et al. (2017)	Eng↔Fre,Por,Spa; Eng→Cze,Ger,Pol,Rum	B	MuL	RT	TTT	Health educ.	(bio)medical text
Yu and Zhu (2021)	Eng↔Chi	B	BL	RT	TTT	Health educ.	(bio)medical text
Ziganshina et al. (2021)	Eng→Rus	U	BL	RT	TTT	Health educ.	(bio)medical text

[a] Languages are reported according to the ISO 639/2B international standard that defines three-letter identifiers for all known human languages (Hutchins, 2017).

[b] Translation direction: U = Unidirectional, B = Bidirectional, Hutchins (1995).

[c] Translation approach: MoL = Monolingual, BL = Bilingual, MuL = Multilingual.

[d] Translation timing: RT = Real Time, PT = Pre-translation, fp = fixed-phrase, eb = example-based.

[e] Translation type: TTT = Text-to-Text, TTS = Text-to-Speech, STT = Speech-to-Text, STS = Speech-to-Speech.

From the alluvial chart it is also possible to retrieve directionality, which impacts differently on languages. For some language pairs, uneven directionality was identified: for instance, we had 19 articles for Eng→Spa and 10 for Spa→Eng; 16 for Eng→Chi and seven for Chi→Eng; five for Eng→Ara and none for Ara→Eng. Since this review only considers articles written in English, we can assume that such an inclusion/exclusion criterion reduces the chances to reach contributions on (bio)medical MT written in other languages and, plausibly, dealing with source languages other than English.

Directionality was instead more homogeneous for many European languages, as we retrieved the same number of implementations ($n = 4$) for Eng→Ita/Ita→Eng and a very similar amount of implementations for Eng→Fre/Fre→Eng, Eng→Ger/Ger→Eng, Eng→Por/Por→Eng.

Several low-resource languages were present as target only (e.g., Afrikaans, Armenian, Farsi, Polish, and Tagalog).

Overall, the LEP context for Spanish-/Chinese-speaking individuals was the most addressed in the literature. It is important to point out that in the typical situation of a LEP patient, having more MT solutions supporting the Eng→Spa translation and fewer supporting the reverse Spa→Eng direction means that the core focus is more on providing communications (either synchronous or asynchronous) to the patient and less on receiving communications from the patient. This suggests that an even greater attention should be given to the applied research on language inclusiveness in the healthcare sector. This is also confirmed by the other criteria described in this section.

In addition to languages, we can see (Fig. 4-A) that bilingual unidirectional translation was the most common scenario ($n = 24$), followed by unidirectional multilingual studies ($n = 9$). Bilingual and multilingual bidirectional translation were considered in 10 articles each. Interestingly, 5 articles presented monolingual translations (i.e., Ehab et al. (2019), Lester et al. (2021), Li et al. (2020), Weng et al. (2019), and van den Bercken et al. (2019)): these are peculiar studies where MT is used to achieve simplified versions of the source documents (i.e., e-prescriptions and (bio)medical or clinical texts), in the same language.

Similarly, 38 articles proposed real-time translation (Fig. 4-B), predominantly bilingual ($n = 31$), only two with pre-translation (Spechbach et al., 2019; Ehab et al., 2018) and two with both timings (Turner et al., 2019; Mutal et al., 2020).

In terms of application domain, we had 21 articles on clinical communication and 37 on health education. The data breakdown per translation approach (Fig. 5-A) revealed that bilingual translation of (bio)medical texts[20] ($n = 12$), public health documents[21] ($n = 7$), and patient consultations[22] ($n = 7$) were the most investigated areas. When considering translation type (Fig. 5-B), we had a similar predominance of text-to-text MT applied to health education ($n = 21$).

4.3. MT technologies

Several MT technologies were discussed in the articles, which did not only focus on the widely known NMT and SMT approaches but also proposed hybrid, example-based and rule-based solutions. A comparison among different MT solutions was presented in 25 articles, while the remaining 33 articles did not offer a direct comparison. The bubble plot (Evergreen, 2019) in Fig. 6-A depicts the outcomes of our MLR as per this class of parameters, by highlighting whether a comparison was involved (green bubbles) or not (yellow bubbles). The most common scenario was represented by custom NMT solutions without any comparison ($n = 13$), immediately followed by the quality assessment of free public MTEs (e.g., Google Translate, DeepL, etc.) based on NMT ($n = 9$). As for the articles presenting a comparison, we had three records with custom NMT vs. free NMT baselines (Way et al., 2020; Liu et al., 2020, and Hayakawa and Arase, 2020), five with custom NMT vs. custom SMT (Álvarez Vidal et al., 2021; Skianis et al.,

[20] Documents such as scientific articles and medical reports, which are primarily written for in-domain experts.

[21] Documents whose main audience is represented by the general public and/or patients.

[22] Intended as any communication/interview with a patient (Caldwell, 2019).

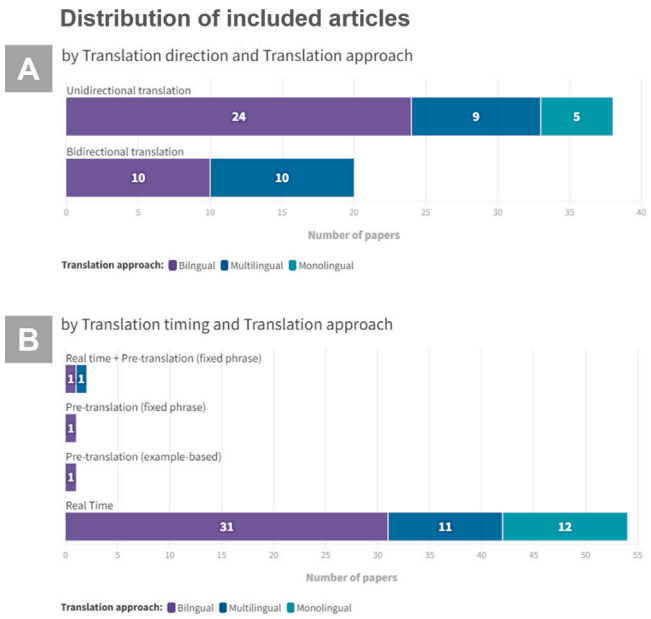

Fig. 4. Translation direction, timing, and approaches.

2020; Deep et al., 2021; Wołk and Marasek, 2015a, and Jimeno Yepes et al., 2017), five custom NMT vs. custom NMT (Ziganshina et al., 2021; Bawden et al., 2020; Hira et al., 2019; Yeganova et al., 2021, and Neves et al., 2022). All other combinations were present in smaller quantities.

As for the implementation solutions discussed in the included articles (Fig. 6-B), we can see that Google Translate (n = 22), OpenNMT (n = 12), and Moses (n = 11) were those occurring the highest number of times, since in several articles they were also considered as baselines for other customised MT solutions. Interestingly, in few cases the authors only mentioned the enabling Python library (e.g., TensorFlow, Groundhog and Theano, Tensor2Tensor), while in 22 cases not enough implementation details were provided.

Finally, when articles about WMT findings were considered, the corresponding values for class-2 parameters were given by the sum of all submissions evaluated in each WMT article.

The full analysis is summarised in Table 7, where the training aspect is also introduced, before being discussed more extensively in Section 4.4.1.

4.4. MT training

4.4.1. Vocabularies and dictionaries

As anticipated (Section 4.1, Fig. 2), MT training was described in 62% articles, where vocabularies and/or text corpora were used. More specifically, in-domain vocabularies (e.g., UMLS National Library of Medicine (NLM), 2022b, SNOMED-CT SNOMED International, 2022, MedDRA International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH), 2022, etc.), either public or custom, were exploited in 13 articles (Álvarez Vidal et al., 2021; Alam et al., 2021; Bawden et al., 2020; Li et al., 2020; Liu et al., 2020; Mutal et al., 2020; Neves et al., 2022; Rani et al., 2019; Renato et al., 2018; Skianis et al., 2020; Soares et al., 2020; Soto et al., 2019, and Yeganova et al., 2021), while general-domain vocabularies (e.g., Wikipedia, Wiktionary, etc.) appeared only in three articles (Luger et al., 2020; Wołk and Marasek, 2015a, and Musleh et al., 2018). It is important to notice that the amount of words/terms used for training was specified only rarely (n = 8) (Li et al., 2020; Musleh et al., 2018; Rani et al., 2019; Renato et al., 2018; Wołk and Marasek, 2015a; Bawden et al., 2020; Neves et al., 2022, and Yeganova et al., 2021), and that a few thousands was the typical order of magnitude in those occasions. All the vocabulary-related aspects are listed in Table 8.

4.4.2. Text corpora

Text corpora[23] were then examined, since they are one of the most effective data-driven training approaches for NMT/SMT and they were not considered in any previous literature review in this field. As already discussed for vocabularies (Table 8), the domain

[23] A *text corpus* is a very large textual dataset extracted from real-world sources, usually containing millions or billions of words and typically exploited to examine how words and languages are used.

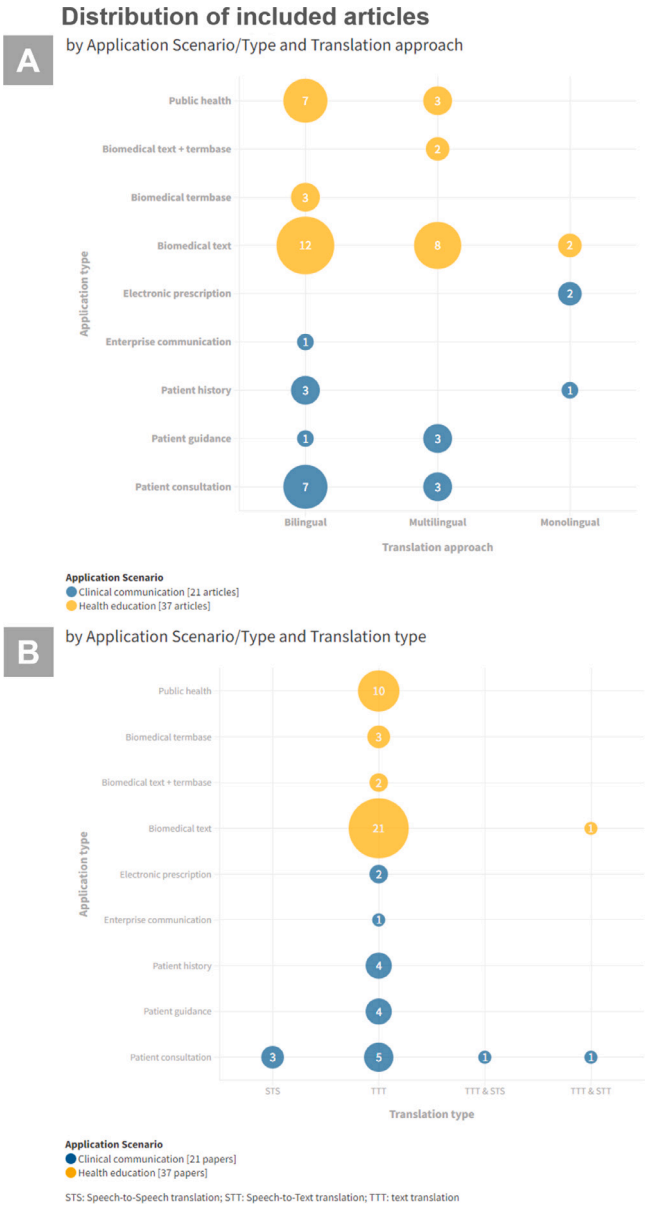

Fig. 5. Application scenarios, MT types/approaches.

and size were essential aspects to consider also for corpora, but other two key elements were introduced as well: phase of use (e.g., MT training or pre-training) and typology.

The typology of a text corpus depends on many features. In terms of language, *monolingual corpora* contain texts in a single language and are very often used to examine language patterns or word combinations but find rare application in MT (except for monolingual MT, as in Lester et al. (2021)); *parallel aligned corpora* present the matched combination (i.e., alignment) of two monolingual corpora, whose elements (i.e., segments[24]) are the translation of each other, and are the most widely used for MT training. They can be either bilingual or multilingual. As for time, corpora can be *diachronic*[25] or *synchronic*.[26] Finally, in terms of

[24] A *segment* in a corpus, as well as in a translation memory, normally coincides with a sentence (or a large chunk of it).

[25] The texts composing a *diachronic text corpus* were written in different time periods and are, therefore, typically used in language development studies. For that very reason, they should not be used in MT training.

[26] The texts composing a *synchronic text corpus* represent a snapshot of a given language in a given time window; normally, they are collected within no more than one-year time. Therefore, all text corpora used in MT should be synchronic.

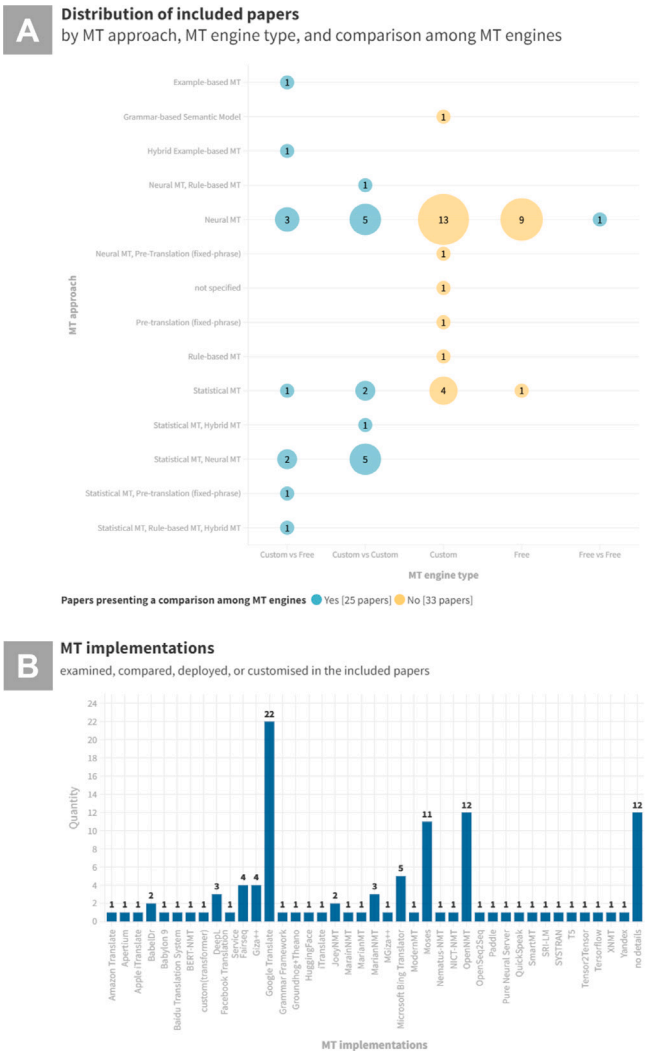

Fig. 6. MT types, approaches.

status, corpora can be either *static*[27] or *monitor*.[28] Typically, NMT/SMT engines are trained with synchronic parallel aligned corpora, either static or monitor.

In our MLR, 34 articles leveraged text corpora, as detailed in Table 9. In-domain corpora represented the most frequently used type, especially parallel aligned bilingual corpora used for the Eng→Spa training ($n = 4$: Álvarez Vidal et al. (2021), Liu et al. (2020), Manchanda and Grunin (2020), and Way et al. (2020)), but also monolingual English corpora were used ($n = 6$: Hayakawa and Arase (2020), Li et al. (2020), Mutal et al. (2020), Renato et al. (2018), San et al. (2022), and van den Bercken et al. (2019)). The use of multilingual parallel aligned corpora was explicitly declared only in Muhaxov et al. (2016). In only one case the corpus was made up of texts dealing with a single specific topic (i.e., COVID-19), as in Way et al. (2020). General-domain corpora were used in 10 articles, and always in combination with in-domain corpora (Álvarez Vidal et al., 2021; Li et al., 2020; Liu et al., 2020; Manchanda and Grunin, 2020; Musleh et al., 2018; Hira et al., 2019; Liu and Huang, 2021; Neves et al., 2022; San et al., 2022, and Weng et al., 2019). In three cases, text corpora were used for MT pre-training (Álvarez Vidal et al., 2021; Huck et al., 2017, and Li et al., 2020). It is interesting to notice that the corpora listed in Table 9 are coherent with the analysis of parallel corpora for the biomedical domain presented in the work by Névéol et al. (2018).

Remarkably, different units of measurement were adopted to indicate the size of the text corpora, but always according to a clear pattern. Indeed, articles proposing monolingual corpora used documents (e.g., Lester et al. (2021) and Li et al. (2020))

[27] These are also known as *reference text corpora*, as their content is not expected to change. Therefore, they are suitable for MT training.

[28] The texts in these corpora are regularly updated, thus making them suitable for MT training only if the procedure has to be repeated multiple times.

Table 7
Overview of selected papers in terms of MT approach, MT engine (MTE) implementation and type, training, comparison.

Article	MT approach[a]	MTE implementation[b]	MTE type[c]	MTE training described [Yes/No]	MTE comparison [Yes(Q.ty)/No]
Alam et al. (2021)	NMT,RBMT	OpenNMT, MarianMT, JoyeNMT, SmartMT	C	Y	Y(9)
Almahasees and Jaccomard (2020)	NMT	FTS	F	N	N
Almahasees et al. (2021)	NMT	GT	F	Y	N
Álvarez Vidal et al. (2021)	SMT,NMT	ModernMT, Apertium, GT	C	Y	Y(4)
Bawden et al. (2020)	NMT	MarianNMT, OpenNMT, Fairseq, Tersorflow, Tensor2Tensor, HuggingFace, BERT-NMT, Paddle, T5	C	Y	Y(20)
Bojar et al. (2016)	SMT	Moses	C	Y	Y(5)
Nunzio et al. (2021)	NMT	DeepL, Yandex	F	N	Y(2)
Chen et al. (2016)	SMT	GT	F	N	N
Chen et al. (2017)	–	iTranslate	C	N	N
Das et al. (2019)	NMT	GT	F	n.s	N
Deep et al. (2021)	SMT,NMT	Moses, Giza++, OpenNMT	C	Y	Y(3)
Dew et al. (2015)	NMT	PHAST (no details)	C	N	N
Ehab et al. (2018)	EBMT	GT, EBMT (*no details*)	C,F	Y	Y(2)
Ehab et al. (2019)	hEBMT	GT, EBMT (*no details*)	C,F	Y	Y(4)
Hayakawa and Arase (2020)	NMT	GT, NICT-NMT	C,F	Y	Y(2)
Hira et al. (2019)	NMT	OpenNMT	C	Y	N
Huck et al. (2017)	NMT	Nematus-NMT	C	Y	N
Kapoor et al. (2022)	NMT	GT	F	N	N
Khoong et al. (2019)	NMT	GT	F	N	N
Kumar et al. (2018)	SMT	*No details*	C	Y	Y
Lankford et al. (2022)	NMT	OpenNMT	C	Y	N
Lee et al. (2023)	NMT	GT, Apple iTranslate, MSBT	F	N	Y(3)
Lester et al. (2021)	NMT	OpenNMT	C	Y	N
Li et al. (2020)	NMT	OpenNMT	C	Y	N
Liu and Cai (2015)	SMT,HMT	Moses, GT, MSBT	C	Y	Y(5)
Liu et al. (2020)	NMT	GT, BTS, NMT (*no details*)	C,F	Y	Y(3)
Liu and Huang (2021)	NMT	OpenNMT	C	Y	N
Luger et al. (2020)	NMT	JoeyNMT	C	Y	N
Manchanda and Grunin (2020)	NMT	OpenSeq2Seq	C	Y	N
Marais et al. (2020)	GSM	Grammar Framework	C	N	N
Miller et al. (2018)	NMT	GT	F	N	N
Muhaxov et al. (2016)	SMT	Moses	C	n.s	N
Musleh et al. (2018)	SMT	Moses, Giza++	C	Y	N
Mutal et al. (2020)	NMT,PT(fp)	BabelDr	C	Y	N
Neves et al. (2018)	NMT, SMT	OpenNMT, Moses, Other (*no details*)	C	Y	Y(6)
Neves et al. (2022)	NMT	Fairseq, MarainNMT, SYSTRAN, Pure Neural Server	C	Y	Y(37)
Park et al. (2022)	NMT	GT	F	N	N
Rani et al. (2019)	SMT	SMT (*no details*), GT	C	Y	Y(3)
Renato et al. (2018)	SMT,NMT	Moses, GT, MSBT	C,F	Y	Y(3)
San et al. (2022)	NMT	XNMT	C	Y	N
Shin et al. (2015)	RBMT	*No details*	C	N	N
Skianis et al. (2020)	SMT,NMT	Moses, fairseq	C	Y	Y(2)
Soares et al. (2020)	NMT	OpenNMT	C	N	N
Soto et al. (2019)	NMT	*No details*	C	Y	N
Spechbach et al. (2019)	PT(fp)	BabelDr	C	N	N
Taira et al. (2021)	NMT	GT	F	N	N
Takakusagi et al. (2021)	NMT	DeepL	F	N	N
Taylor et al. (2015)	SMT	GT, Babylon 9	C,F	N	Y(2)
Turner et al. (2015)	SMT	GT	F	N	N
Turner et al. (2019)	SMT,PT(fp)	GT, QuickSpeak	C,F	N	Y(2)

(*continued on next page*)

Table 7 (*continued*).

van den Bercken et al. (2019)	NMT	OpenNMT	C	Y	N
Way et al. (2020)	NMT	MarianNMT, GT, AT, MSBT	C,F	Y	Y(11)
Way et al. (2020)	NMT	MarianNMT, GT, AT, MSBT	C,F	Y	Y(11)
Weng et al. (2019)	SMT	Moses	C	Y	N
Wołk and Marasek (2015a)	SMT,NMT	Moses, Giza++, MGiza++, Groundhog+Theano	C	Y	Y(3)
Yeganova et al. (2021)	NMT	Moses, SRI-LM, Giza++	C	Y	Y(5)
Jimeno Yepes et al. (2017)	SMT, NMT	MarianNMT, OpenNMT, Fairseq, custom(transformer)	C	Y	Y(7)
Yu and Zhu (2021)	SMT,RBMT,HMT	*No details*	C,F	Y	Y(5)
Ziganshina et al. (2021)	NMT	DeepL, GT, MSBT	F	N	Y(3)

[a] MT approach: SMT = Statistical MT; NMT = Neural MT; RBMT = Rule-based MT; EBMT = Example-based MT; hEBMT = hybrid EBMT; PT(fp) = Pre-Translation (fixed phrase); GSM = Grammar-based Semantic Model (Ranta, 2011).

[b] AT = Amazon Translate; BTs = Baidu Transl. Syst.; FTS = Facebook Transl. Service; GT = Google Transl.; MSBT = Microsoft Bing Transl.

[c] MT Engine type: F = Free to use; C = Custom/ad-hoc.

Table 8
Use of dictionaries and vocabularies supporting the MTE training phase.

Article	Dictionary/Vocabulary used [number, ISO 639/2B language codes, description, size]
Alam et al. (2021)	1 terminology about COVID-19 taken from TICO-19 project (for Eng → Chi, Fre, Rus, Kor language pairs) [600 terms]; 1 terminology about Healthcare concepts automatically extracted from Wikipedia (for Cze → Ger) (n.q.)
Álvarez Vidal et al. (2021)	Data glossaries and glossary-like DBs containing frequent terms and expressions from the medical domain (Eng→Spa glossary from MeSpEn Villegas et al., 2018, 10th rev of the Int. Statistical Classification of ICD and SnowMedCT SNOMED International, 2022) (n.q.)[a]
Bawden et al. (2020)	1 biomedical terminology (Baq) [2k terms]
Li et al. (2020)	1 clinical vocabulary to simplify clinical-domain word embeddings [~400k terms]
Liu et al. (2020)	1 custom MedDRA (International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH), 2022) dictionary (medical terminology only) (n.q.)
Luger et al. (2020)	1 general-domain multilingual vocabulary (Fre, Spa, Eng, Bam) (n.q.)
Musleh et al. (2018)	alignment of Wikipedia [~40.7k words], Wiktionary (Wikimedia Foundation, 2022) [~10.3k words], OmegaWiki [~3.4k words], and combination of BabelNet (Navigli and Ponzetto, 2012) + MeSH (National Library of Medicine (NLM), 2022a) [~200k words]
Mutal et al. (2020)	1 clinical-domain, internal to the system (n.q.)
Neves et al. (2022)	1 terminology extracted from biomedical literature (Spa → Eng) [7.k terms] + 1 terminology extracted from clinical ontology
Rani et al. (2019)	Lexicon of breast cancer terms and bilingual dictionary along with Eng→Tam termbase [~3.5k terms]
Renato et al. (2018)	6 clinical-domain dictionaries [overall size: ~190k terms]: DeCS Health Science Descriptors (~163k terms), Dicionario Medico (Pt-br) (~9k terms), Vocabulario de medicina (~6.7k terms), Wikipedia medicine (~21k terms), diccionario de termos medicos (~6k terms), dicionario medico (~9.5k terms) (Renato et al., 2018)
Skianis et al. (2020)	2 medical terminologies (MedDRA International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH), 2022, ORDO Vasant et al., 2014) + terms extracted from other in-domain training corpora(n.q.)
Soares et al. (2020)	1, based on UMLS (National Library of Medicine (NLM), 2022b), SNOMED CT (SNOMED International, 2022) (n.q.)
Soto et al. (2019)	1 medical bilingual dictionary based on SNOMED-CT (SNOMED International, 2022) (n.q.)
Wołk and Marasek (2015a)	1 bilingual vocabulary for Pol [~148k terms] and Eng [~109k terms]
Yeganova et al. (2021)	1 biomedical terminology from the Basque ICD-10-CM ed. (Baq) [2k terms]

[a] (n.q.) = size not quantified.

or words/terms (e.g., Li et al. (2020) or Renato et al. (2018)) as units, while those using parallel corpora adopted sentences or segments as units. From a numerical standpoint, a considerable variety emerged, as the size ranged from few hundreds (e.g., Ehab et al. (2018, 2019), or Liu and Cai (2015), Luger et al. (2020), or Musleh et al. (2018)) to several millions (e.g., Way et al. (2020), Manchanda and Grunin (2020), or Liu et al. (2020)) sentences, thus significantly hindering the chances of effective comparison between different studies.

Table 9
Use of text corpora in the MTE training phase.

Article	Domain	Source corpora	Type[a], ISO 639/2B langs.	UoM[b]	Size
Álvarez Vidal et al. (2021)	General(p-t)[c]	Corpora: Scielo, Europarl, GlobalVoices, News Commentary	BL-PA (Eng-Spa)	sg	–
	Healthcare	Corpora: EMEA, PubMed, Medline Plus, IBECS, MSDManuals, Portal Clinic, UFAL Medical,	BL-PA (Eng-Spa)	sg	2.8M
Bawden et al. (2020)	Healthcare	Biomedical abstracts (multiple sources)	BL-PA (8 langs.)	sn	Variable
Bojar et al. (2016)	Healthcare	SCIELO and Medline corpora	BL-PA (3 langs.)	sn	Variable
Deep et al. (2021)	Healthcare	TDIL	BL-PA (Eng-Pan)	sn	26k
Ehab et al. (2018)	Healthcare	Internal medical publications	BL-PA (Eng-Ara)	sn	259
Ehab et al. (2019)	Healthcare	Internal medical publications	BL-PA (Eng-Ara)	sn	259
	Healthcare	Worldwide Arabic Medical Translation Guide (Common Medical Terms)	BL-PA (Eng-Ara)	sn	509
Hayakawa and Arase (2020)	Healthcare	MSD Manual Consumer/Professional Version, New England Journal of Medicine, Journal of Clinical Oncology, ICH guidelines	MoL (Eng)	sn	2.5k
Hira et al. (2019)	General+Healthcare	In-domain (SCIELO, EMEA, UFAL, Medline) + out-domain	BL-PA (Eng-Fre)	sn	21M + 10M
Huck et al. (2017)	Healthcare(p-t)	Europarl, News Commentary, Common Crawl	BL-PA (Eng-Deu)	sn	1.7M
	Healthcare	In-domain UFAL Medical	BL-PA (Eng-Deu)	sg	2M
Kumar et al. (2018)	Healthcare	–	BL-PA (Eng-Tam)	sn	15k
Lankford et al. (2022)	Healthcare+COVID-19	Ad-hoc developed. Sources: health-related data and COVID-19 data from Irish Dept. of Health	BL-PA (Eng-Gle)	ln	16,2k
Lester et al. (2021)	Healthcare	Online outpatient mail-order pharmacy	MoL-P (Eng-Eng)	d	530k
Li et al. (2020)	General(p-t) Healthcare(p-t) Healthcare	Wikipedia, Gigaword MIMIC-III Online outpatient mail-order pharmacy	MoL-PA (Eng) MoL-PA (Eng) MoL-PA (Eng)	w w d	– – 530k
Liu and Cai (2015)	Healthcare Healthcare	MedlinePlus Electronic Health Records	BL-PA (Eng-Spa) BL-PA (Eng-Spa)	sn sn	144k 108
Liu et al. (2020)	General+Healthcare	Chinese Medical Journal Network, Science Foundation Shared Services Network, WMT (World MT), NLP group of Nanjing Univ.	BL-PA (Eng-Chi)	sn	5.4M
Liu and Huang (2021)	General+Healthcare	Ad-hoc developed. In-domain: sentences extracted from the New England Journal of Medicine archive. Out-domain: newswire data	BL-PA (Eng-Chi)	sn	97k + 24.8M
Luger et al. (2020)	Healthcare	SIL-Mali	2 BL-PA (Bam-Eng/Fre)	sn	2100+2100
Manchanda and Grunin (2020)	General	Paracrawl open	BL-PA (Eng-Spa)	sn	38M
	Healthcare	Internal correspondence letters and in-domain TM	BL-PA (Eng-Spa)	sn	492k+14k
Marais et al. (2020)	Healthcare	GF (Grammatical Framework) (Ranta, 2011)	–	sn	3.9k
Muhaxov et al. (2016)	Healthcare	Documents from hospital clinics and medical universities	MuL-PA (Chi-Uig-Kaz)	sn	240k
Musleh et al. (2018)	General+Healthcare	Internal sources and movie subtitles repository	BL-PA (Eng-Hin)	sn	1200
	General+Healthcare	Indic multi-parallel	BL-PA (Eng-Urd)	sn	87k
Mutal et al. (2020)	Healthcare	Internal sources	MoL-PA (Fre)	sn	130k

(*continued on next page*)

Furthermore, the majority of articles lacked details on the average segment/sentence length in words for the adopted text corpora. This makes impossible performing comparison among various parallel text corpora depending on sentence length and language pairs, as typically happens in the ridgeline plots (Evergreen, 2019) used in the MT market (Intento Inc., 2022).

Table 9 (*continued*).

Neves et al. (2018)	Healthcare	Medline and EDP corpora	BL-PA (6 langs.)	d	120
Neves et al. (2022)	General+Healthcare	Medline corpus (in-domain)	BL-PA (mult. langs.)	sn	Variable
Renato et al. (2018)	Healthcare	ICD-10, DeCS, EMEA	BL-PA (Spa-Por)	sn	10.9k+73.5k +8.5k
	Healthcare	SciELO, ICD-10, BIREME, Brasilian-Portuguese vademecum	MoL (Por)	t	88k
San et al. (2022)	General+Healthcare	Ad-hoc developed. In-domain: medical handbooks and clinical assessment docs. Out-domain: ASEAN corpus (tourism)	BL-PA (Eng-Bur)	sn	14,6k
Skianis et al. (2020)	Healthcare	ICD-11	–	sn	500k
van den Bercken et al. (2019)	Healthcare	Texts about diseases from Wikipedia, DBPedia (manually + automatically built)	MoL-A (Eng)	sn	8.5k
Way et al. (2020)	COVID-19	TAUS Corona Crisis, EMEA, SketchEngine Covid19, ParaCrawl, Wikipedia	4 BL-PA (Deu/Spa/Fre/Ita-Eng)	sn	29.4M+10.7M +11.3M+10.2M
Weng et al. (2019)	General+Healthcare	In-domain: MIMIC-III (professional section + consumer section); out-domain: WMT English News Crawl	MoL (Eng)	sn	600k + 38.2M
Wołk and Marasek (2015a)	Healthcare	Internal sources	BL-PA (Eng-Pol)	sn	1,04M
Yeganova et al. (2021)	Healthcare	Medline corpus	BL-PA (8 langs.)	sn	Variable
Jimeno Yepes et al. (2017)	Healthcare	Titles and abstracts from scientific publications (SCIELO, EDP); health-related docs. (Cochrane, NHS)	BL-PA (10 langs.)	sn	–
Yu and Zhu (2021)	Healthcare	Internal sources	BL-PA (Eng-Chi)	sn	250k

[a] Corpus type: MoL = monolingual, BL = bilingual, MuL = multilingual, P = parallel, A = aligned.

[b] Unit of measurement: d = documents, sn = sentences, sg = segments, w = words, t = terms.

[c] p-t: for pre-training purposes.

4.5. Population and experiment settings, study scope and materials

Healthcare professionals[29] were undoubtedly the most represented user category (Fig. 7-A): either as the only targeted group ($n = 14$) or in conjunction with patients ($n = 12$) or with the general public ($n = 2$). Even if fewer articles considered patients only, either as generic ($n = 4$) or from a specific category ($n = 3$), it is evident that the majority of MT applications focused on these two groups of users, considering the general public ($n = 5$) only marginally.

Similarly, it was possible to group the articles depending on how respective authors defined the source language proficiency of target users (Fig. 7-B): overall limited ($n = 11$), limited in the healthcare sector ($n = 14$), or no proficiency at all ($n = 9$). Sectorial limited proficiency accounts for those cases in which special categories of non-native speakers were involved (clinicians, students, etc.), while the other two categories suggest the discussed MT solutions were applicable to a much broader audience and, therefore, their effectiveness (if demonstrated) could be eventually perceived on a larger scale. Noteworthy, in 12 cases no considerations about language groups were provided by the authors. In two cases, translators (Alam et al., 2021) and researchers in computational linguistics, natural language processing, or machine translation (Lankford et al., 2022) were identified as the target user typology.

In terms of potential risks deriving from wrong MT outputs (Fig. 7-C), many articles ($n = 25$) did not discuss this aspect, while the others oscillated in a high-to-low scale.[30] It is important to notice that articles mentioning explicitly high ($n = 7$), medium-high ($n = 9$), and medium ($n = 8$) risks surpassed all the others, thus clearly confirming the importance of assessing the MT quality before delivering its output to the intended target users (especially when they are patients or professionals) and consequently impacting on their decision-making or evaluation activities.

All data are reported in Table 10.

As for the Class-5 criteria, only 15 articles quantified explicitly the deployment period of the proposed MT solution. In one case, it was a 3-month pilot study (Marais et al., 2020) and in three other articles it was a long-term clinical deployment (Spechbach et al., 2019; Dew et al., 2015, and Muhaxov et al., 2016). In all other cases ($n = 8$), the articles were those reporting the findings of the annual WMT shared tasks and, therefore, the time period coincided with the task duration established by the WMT organisers.

In terms of deployment stage, the majority of articles dealt with pilot studies or prototype tests. The WMT shared tasks represented the other subset of deployment types. Only in two cases GitHub repositories were provided (Manchanda and Grunin, 2020 and Way et al., 2020, which is not accessible anymore). One pilot study focused on mock emergency scenarios, thus proposing

[29] Under this definition we gathered doctors, physicians, clinicians, general practitioners, nurses, medical staff members, and medical researchers.

[30] The risk scale is defined by the authors of this MLR, according to what explained in Section 3.6, Class 4, criterion no.iii.

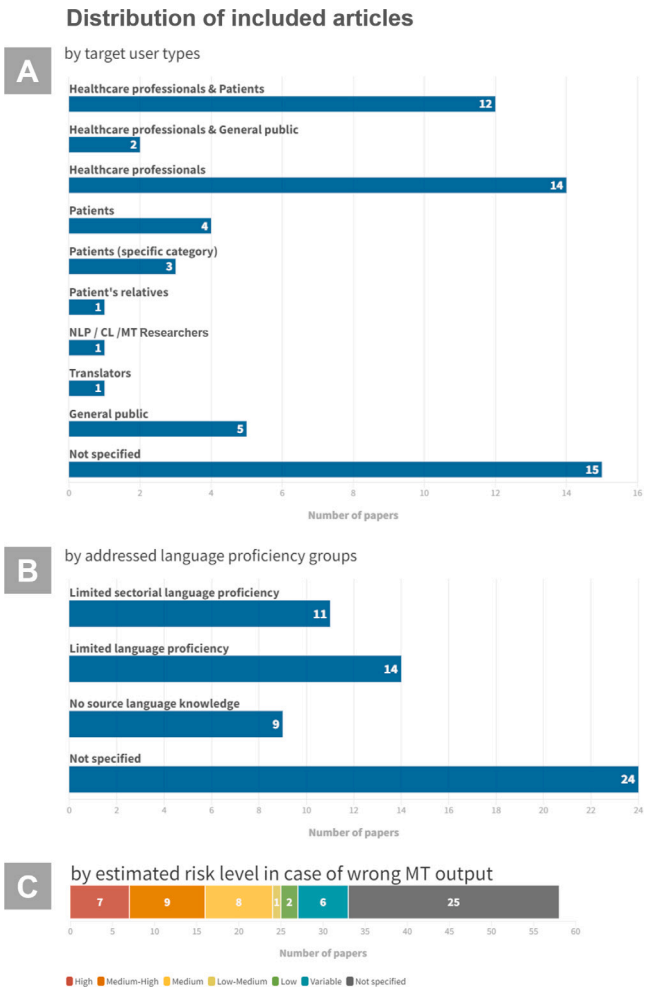

Fig. 7. Target users/language groups, potential risks.

scripted situations whose text contents required MT (Turner et al., 2019). Moreover, a prospective observational study (Taylor et al., 2015), a descriptive cross-sectional study (Kumar et al., 2018), and a randomised study (Wołk et al., 2015) were also presented. Even if the deployment location was specified explicitly in very few cases, we also observed a prevalence of studies conducted in the US ($n = 18$), followed by Europe ($n = 12$) and Asia ($n = 10$). This geographically-unbalanced distribution is plausibly motivated by the attention to the challenges faced by LEP people in the US. The choropleth map (Evergreen, 2019) depicted in Fig. 8 shows how the articles are distributed, according to a gradient colour intensity scale.

Finally, it is worth to point out that articles reporting on WMT findings do not explicitly address target user groups, language proficiency groups, and the risk estimation for wrong translation outputs. As for the deployment stage, duration, and country, the WMT findings where always classified in this MLR as "WMT task".

In Section 4.6, we will consider against what types and amounts of source documents these MT solutions were tested (as we did for their training stage in Sections 4.4.1 and 4.4.2), along with adopted validation procedures and findings reported by the authors of each included paper.

4.6. Evaluation procedures and findings reported in the articles

Evaluation was very often quantitative ($n = 23$) and less frequently qualitative ($n = 10$), otherwise it was a mix of the two approaches ($n = 23$). Only in two cases (i.e., Almahasees and Jaccomard (2020) and Kapoor et al. (2022)) a survey was relied on (see Table 11 for the breakdown by this criterion). Multiple times ($n = 9$), however, the adopted evaluation process was mentioned but not described in details (i.e., Almahasees and Jaccomard (2020), Almahasees et al. (2021), Nunzio et al. (2021), Kapoor et al. (2022), Khoong et al. (2019), Marais et al. (2020), Shin et al. (2015), Soares et al. (2020), and Spechbach et al. (2019)) and this has to be considered a relevant limitation, since it hampers any possible replication of the corresponding study.

Table 10
Breakdown of included articles by target users, language proficiency groups, and potential risks deriving from wrong MT.

Article	User type	Language proficiency group	Estimated risk
Alam et al. (2021)	Translators	n.s.	n.s.
Almahasees and Jaccomard (2020)	General public	Limited (sectorial)	n.s.[a]
Almahasees et al. (2021)	General public	n.s.	Medium-High
Álvarez Vidal et al. (2021)	Healthcare professionals	n.s.	Variable
Bawden et al. (2020)	Healthcare	n.s.	n.s.
Bojar et al. (2016)	n.s.	n.s.	n.s.
Nunzio et al. (2021)	Healthcare professionals	n.s.	Medium-High
Chen et al. (2016)	Patients (specific category)	Limited (overall)	Medium-High
Chen et al. (2017)	Healthcare professionals & Patients	Limited (overall)	Medium-High
Das et al. (2019)	Patient relatives	Limited (overall)	Medium
Deep et al. (2021)	n.s.	n.s.	Variable
Dew et al. (2015)	Healthcare professionals & Patients	Limited (overall)	Low
Ehab et al. (2018)	n.s.	n.s.	n.s.
Ehab et al. (2019)	Healthcare professionals	Limited (sectorial)	Variable
Hayakawa and Arase (2020)	Healthcare professionals & General public	n.s.	n.s.
Hira et al. (2019)	n.s.	n.s.	n.s.
Huck et al. (2017)	n.s.	n.s.	n.s.
Kapoor et al. (2022)	Patients	Limited (overall)	High
Khoong et al. (2019)	Healthcare professionals & Patients	Limited (overall)	High
Kumar et al. (2018)	Patients (specific category)	Limited (sectorial)	n.s.
Lankford et al. (2022)	NLP/CL/MT researchers	n.s.	Medium
Lee et al. (2023)	Patients	Limited (overall)	Low
Lester et al. (2021)	Healthcare professionals	Limited (sectorial)	n.s.
Li et al. (2020)	Healthcare professionals	Limited (sectorial)	n.s.
Liu and Cai (2015)	Healthcare professionals & Patients	Limited (sectorial)	n.s.
Liu et al. (2020)	Healthcare professionals	Limited (sectorial)	n.s.
Liu and Huang (2021)	n.s.	n.s.	n.s.
Luger et al. (2020)	Healthcare professionals & Patients	None	Medium-High
Manchanda and Grunin (2020)	n.s.	n.s.	n.s.
Marais et al. (2020)	Healthcare professionals & Patients	None	Medium
Miller et al. (2018)	Healthcare professionals	Limited (overall)	Medium
Muhaxov et al. (2016)	Healthcare professionals & Patients	Limited (overall)	Medium-High
Musleh et al. (2018)	Healthcare professionals & Patients	Limited (sectorial)	n.s.
Mutal et al. (2020)	Healthcare professionals & Patients	None	Variable
Neves et al. (2018)	n.s.	n.s.	n.s.
Neves et al. (2022)	n.s.	n.s.	n.s.
Park et al. (2022)	Healthcare professionals	n.s.	High
Rani et al. (2019)	Patients (specific category)	Limited (overall)	High
Renato et al. (2018)	Healthcare professionals	None	n.s.
San et al. (2022)	n.s.	n.s.	Low-Medium
Shin et al. (2015)	Healthcare professionals & Patients	None	Medium-High
Skianis et al. (2020)	n.s.	n.s.	n.s.
Soares et al. (2020)	Healthcare professionals	Limited (overall)	Variable
Soto et al. (2019)	Healthcare professionals	None	Medium-High
Spechbach et al. (2019)	Healthcare professionals & Patients	None	High
Taira et al. (2021)	Patients	Limited (overall)	High
Takakusagi et al. (2021)	Healthcare professionals	Limited (sectorial)	High
Taylor et al. (2015)	Patients	None	Medium
Turner et al. (2015)	General public	Limited (overall)	Variable
Turner et al. (2019)	Healthcare professionals & Patients	Limited (overall)	n.s.
van den Bercken et al. (2019)	General Public	Limited (sectorial)	Medium
Way et al. (2020)	Healthcare professionals & General public	None	Medium-High
Weng et al. (2019)	General public	Limited (sectorial)	Medium
Wołk and Marasek (2015a)	Healthcare professionals	n.s.	Medium
Yeganova et al. (2021)	n.s.	n.s.	n.s.
Jimeno Yepes et al. (2017)	n.s.	n.s.	n.s.
Yu and Zhu (2021)	n.s.	n.s.	n.s.
Ziganshina et al. (2021)	Healthcare professionals	n.s.	n.s.

[a] n.s. = not specified.

All the other criteria from Class-6 and Class-7 are reported in Table 14. More specifically, we can see that manual evaluation was adopted more frequently ($n = 46$) than automatic procedures ($n = 33$). In 21 articles both the processes were applied and all articles adopted at least one evaluation approach (this confirms indirectly the effectiveness of inclusion/exclusion criteria, as articles without any validation would have been less useful in this study if included).

Manual evaluations were mainly based on translation quality assessment (TQA), depending on different error classification taxonomies. The Multidimensional Quality Metrics (MQM) (Lommel et al., 2014) was often referred: it classifies MT errors in terms of accuracy, fluency, design, locale convention, style, terminology, and verity. In a certain number of articles, customised MQM

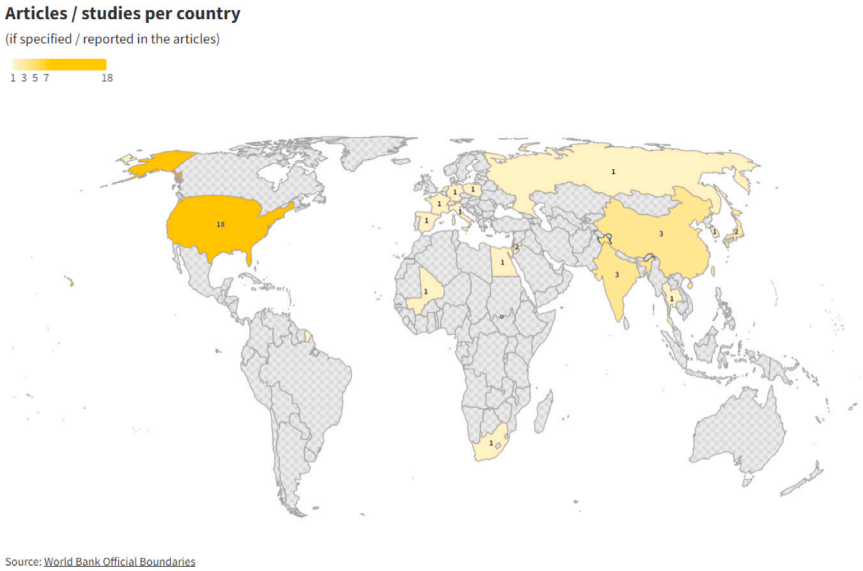

Fig. 8. Geographical locations of included studies/articles.

versions were also proposed, in which other error types such as addition, omission, and mistranslation were added (e.g., Lommel (2018), Vardaro et al. (2019) and Shi et al. (2019)). It is important to notice that, even in very simple case studies (e.g., two freely available MTEs evaluated against a validation text corpus) using heterogeneous taxonomies affects significantly the opportunities to compare the achieved results. Sometimes, a more generic "translation correctness" was used as a metric (e.g., Nunzio et al., 2021; Rani et al., 2019; Li et al., 2020; Liu and Cai, 2015, and Muhaxov et al., 2016). More rarely, other manual metrics were referenced, such as translation speed (Yu and Zhu, 2021), post-editing time (Turner et al., 2015), use time (Spechbach et al., 2019), and functionality/usability (as in Dew et al. (2015) or Spechbach et al. (2019)).

Automatic evaluation was based predominantly on classic metrics for SMT/NMT (Mauser et al., 2008; Radziszewski, 2013), such as BLEU,[31] METEOR,[32] TER,[33] or NIST.[34] In few articles, other metrics were also referenced, usually derivations or improvements of the main ones, such as BLEU-4,[35] SacreBLEU,[36] WER,[37] SER,[38] and chrF2.[39]

Similarly to what emerged for text-corpus-based training, the materials used in validation stages exhibited a great variety in terms of document types (always specified) and size, which ranged from less than 100 (as in Luger et al. (2020), Miller et al. (2018), Muhaxov et al. (2016), Yu and Zhu (2021), and Shin et al. (2015)) to thousands of sentences (e.g., Hayakawa and Arase (2020), Huck et al. (2017), Kumar et al. (2018), Liu et al. (2020), or Manchanda and Grunin (2020)). In some cases, only the number of used documents were reported, without details on the corresponding amount of sentences (e.g., Lester et al. (2021), Li et al. (2020), Liu and Cai (2015), Rani et al. (2019), or Yeganova et al. (2021)) while, in some other, no details about the size were provided at all (e.g., Dew et al. (2015), Ehab et al. (2018, 2019), van den Bercken et al. (2019), Weng et al. (2019), Jimeno Yepes et al. (2017), or Soto et al. (2019)), being this the most relevant limitation in our opinion.

An important aspect is represented by the breakdown per typologies of involved human evaluators or validators: as shown in Table 12, many categories of evaluators were involved in assessing the MT quality as well as to perform manual checks: healthcare professionals (possibly with bilingual expertise) are involved in a substantial number of cases, either alone ($n = 11$), or with

[31] Bilingual Evaluation Understudy (BLEU) is one of the most-widely used metrics for SMT and NMT. It is a language-independent metric quantifying the closeness between MT and human translation (HT), whose [0;1] score tends to 1 when MT overlaps HT; overlapping sequential words are given higher scores (Papineni et al., 2001).

[32] Metric for Evaluation of Translation with Explicit Ordering (METEOR) is calculated depending on the harmonic mean of unigram precision and recall, where recall has a higher weight than precision (Denkowski and Lavie, 2014).

[33] Translation Error Rate (TER) is given by the number of edits required to change the MT output to the reference HT (Snover et al., 2006).

[34] US National Institute of Standards & Technology (NIST) metric was proposed to improve BLEU: it is a metric ranged in [0;15], where higher values correspond to a better MT output. This metric is based on the arithmetic and geometric means of the n-gram matches between MT and HT so to give much emphasis to the correct translation of less frequent terms (Wołk and Marasek, 2015a).

[35] A modified version of BLEU, referring to unigrams, bigrams, trigrams and n-grams (Sutskever et al., 2014).

[36] It is a particular version of BLEU, which computes scores on detokenised outputs and refers to the specific tokenisation test set produced during the 2017 edition of the Conference on Machine Translation (Post, 2018).

[37] Word Error Rate (WER) is the edit (i.e., Levenshtein) distance at word level (Mauser et al., 2008).

[38] Sentence Error Rate (SER) is the percentage of differences in utterances between MT and HT core sentences (Mutal et al., 2020).

[39] chrF2 is a character-based metric that calculates matchings between character n-grams (instead of word n-grams as in BLEU), where $n = 1, \ldots, 6$ and recall is weighted twice than precision (Popović, 2016).

Table 11
Breakdown of included articles by evaluation approach.

Approach	Articles
Qualitative	Dew et al. (2015), Rani et al. (2019), Marais et al. (2020), Soares et al. (2020), Turner et al. (2019), Yu and Zhu (2021), Ziganshina et al. (2021), Muhaxov et al. (2016), Lee et al. (2023) and Weng et al. (2019)
Quantitative	Almahasees et al. (2021), Álvarez Vidal et al. (2021), Chen et al. (2016, 2017), Das et al. (2019), Deep et al. (2021), Ehab et al. (2018, 2019), Khoong et al. (2019), Kumar et al. (2018), Lester et al. (2021), Liu et al. (2020), Manchanda and Grunin (2020), Mutal et al. (2020), Takakusagi et al. (2021), Shin et al. (2015), Taylor et al. (2015), Taira et al. (2021), Wołk and Marasek (2015a), Lankford et al. (2022), San et al. (2022), Alam et al. (2021) and Hira et al. (2019)
Mixed	Nunzio et al. (2021), Hayakawa and Arase (2020), Huck et al. (2017), Li et al. (2020), Liu and Cai (2015), Luger et al. (2020), Miller et al. (2018), Musleh et al. (2018), Renato et al. (2018), Skianis et al. (2020), Soto et al. (2019), Spechbach et al. (2019), Turner et al. (2015), Way et al. (2020), Park et al. (2022), Liu and Huang (2021), Jimeno Yepes et al. (2017), Bawden et al. (2020), van den Bercken et al. (2019), Neves et al. (2018), Bojar et al. (2016), Yeganova et al. (2021) and Neves et al. (2022)
Survey	Almahasees and Jaccomard (2020) and Kapoor et al. (2022)

Table 12
Breakdown of included articles by human evaluator typology (articles only presenting automatic evaluations or surveys are not reported).

Evaluator typology	Articles
Healthcare professionals	Dew et al. (2015), Khoong et al. (2019), Li et al. (2020), Liu and Cai (2015), Marais et al. (2020), Miller et al. (2018), Musleh et al. (2018), Takakusagi et al. (2021), Skianis et al. (2020), Soto et al. (2019) and Park et al. (2022)
Healthcare professionals and laymen	Weng et al. (2019)
Healthcare professionals and patients	Spechbach et al. (2019) and Kapoor et al. (2022)
Healthcare professionals and translators/interpreters	Rani et al. (2019) and Lee et al. (2023)
Laymen	van den Bercken et al. (2019)
Professional translators	Álvarez Vidal et al. (2021), Chen et al. (2016, 2017), Hayakawa and Arase (2020), Renato et al. (2018), Yu and Zhu (2021), Ziganshina et al. (2021) and Alam et al. (2021)
Researchers (in NLP, CL, or MT)	Lankford et al. (2022)
Specific language-group members	Shin et al. (2015), Taira et al. (2021), Taylor et al. (2015) and Turner et al. (2019)
Students	Muhaxov et al. (2016)
Not specified	Almahasees et al. (2021), Nunzio et al. (2021), Das et al. (2019), Ehab et al. (2018), Huck et al. (2017), Lester et al. (2021), Soares et al. (2020), Turner et al. (2015), Way et al. (2020), San et al. (2022), Liu and Huang (2021), Jimeno Yepes et al. (2017), Bawden et al. (2020), Hira et al. (2019), Neves et al. (2018), Bojar et al. (2016), Yeganova et al. (2021) and Neves et al. (2022)

patients ($n = 2$), or with translators ($n = 2$). Professional translators and interpreters were also frequently involved ($n = 8$), while in fewer cases evaluators were volunteers from the target language groups ($n = 4$). A student-only evaluation was proposed in just one article (Muhaxov et al., 2016), while in one case only laymen (van den Bercken et al., 2019) or only MT/NLP/CL researchers (Lankford et al., 2022) were involved as evaluators. In nine cases the absence of details about the evaluators was simply because automatic-only evaluation only or survey-only approaches were adopted. Finally, in 18 cases, no details at all were provided.

Pre-/post-editing were rarely considered in the included articles: pre-editing explicitly appeared as a step of the evaluation process only five times (i.e., Dew et al. (2015), Turner et al. (2015), Soares et al. (2020), Álvarez Vidal et al. (2021), and Ziganshina et al. (2021)) while post-editing two times (i.e., Taylor et al. (2015) and Lester et al. (2021)). When directly used, a significant effectiveness in improving MT quality was attributed to pre-/post-editing. Moreover, also in some other articles examined in this MLR, pre-/post-editing interventions were recognised (even if not directly adopted) as useful (or fundamental) elements to improve MT quality when a language as sectorial as the (bio)medical one is entailed (e.g., Almahasees et al. (2021), or Takakusagi et al. (2021)).

Finally, overall findings (as detailed in Table 14) showed how MT is vastly assessed as still not completely capable of replacing translators and interpreters in the healthcare communication. The majority of articles included in the MLR pointed out that the MT solutions are a valid support/supplement (see for instance (Das et al., 2019; Chen et al., 2017), or Spechbach et al. (2019)) but several issues still exist in terms of fluency (Álvarez Vidal et al., 2021), accuracy (Ehab et al., 2018), unnatural translations (mentioned in Nunzio et al. (2021), and Shin et al. (2015)), and domain-adequacy (Deep et al., 2021). To overcome these issues, post-editing is suggested before delivering MT outputs to end users (as detailed in Almahasees et al. (2021), Liu et al. (2020), Soares et al. (2020), Taylor et al. (2015), Ziganshina et al. (2021), and Soares et al. (2020)). Interestingly, the findings reported in the WMT articles (i.e., Alam et al. (2021), Bawden et al. (2020), Bojar et al. (2016), Hira et al. (2019), Neves et al. (2018, 2022), Jimeno Yepes et al. (2017), and Yeganova et al. (2021)) show a progressive improvement in MT output quality. In many cases,

even when translation quality issues were considered as not severe and MT was judged as fit for use, safety risks were identified as a relevant cause of concern (e.g., Chen et al. (2017), Das et al. (2019), Khoong et al. (2019), Lester et al. (2021), Miller et al. (2018), Soares et al. (2020), Taira et al. (2021), Taylor et al. (2015), and Yu and Zhu (2021)). Nevertheless, a certain number of articles proposed definitely positive findings ($n = 10$: Rani et al. (2019), Li et al. (2020), Manchanda and Grunin (2020), Marais et al. (2020), Muhaxov et al. (2016), Mutal et al. (2020), Takakusagi et al. (2021), Renato et al. (2018), Skianis et al. (2020), and Soto et al. (2019)).

However, as anticipated in Section 3.6.2, the heterogeneity of the included studies hinders a true scientifically credible comparison of MT quality that is solely based on the findings reported in each article. Therefore, the entire included dataset was evaluated not only in terms of the respective authors' claims, but also according to the proposed QA scoring, as described in Section 5.1.

5. Discussion

The great enthusiasm surrounding the MT applicability to numerous translation tasks in our daily lives has invariably reached also the healthcare and biomedical sector. The considerable amount of studies selected for this MLR proves the efforts being made in the recent years by researchers to assess whether the MT is worthy of adoption also in this peculiar field. Moreover, by thoroughly examining the papers included in the MLR, according to the identified set of 32 analysis criteria, a more systematic depiction of the current scenario is now possible and a twofold pattern emerges. On the one hand, technological advancements have widened the number of solutions available to implement MT also in healthcare, by providing their potential users with new and more performing algorithms. On the other hand, however, a significant heterogeneity characterise these studies in terms of contents to translate, experiment settings, translation quality levels, target languages, involved stakeholders, and evaluation procedures. Furthermore, in many cases the maturity level of the proposed solutions and their suitability to be deployed in real contexts and on a large scale seem somewhat limited.

The combination of heterogeneous features and perceived lack of adequate validation could be a disincentive to a more organised incorporation of MT in healthcare: being the scenario so variegated, many end users (from clinical practitioners to patients) could be discouraged to rely on MT or they could make wrong choices (e.g., simply avoiding MT because of scepticism, just selecting the most famous MT, overlooking proper validation procedures, or skipping necessary risk-mitigation strategies). The aim of this section is, therefore, to assess the quality of MT in healthcare according to a scientifically-sound approach (Section 5.1) in order to support the different typologies of involved users in their decisions whether exploiting MT in healthcare. In addition, the insights gathered in the MLR will be used to propose a set of guidelines on how to design feasibility studies about MT in healthcare capable of truly supporting healthcare professionals (Section 5.2).

5.1. MT quality in healthcare

While Section 4 allowed us to sketch the current landscape of MT in healthcare, spanning on the January 2015–February 2023 period, this section is dedicated to assess the translation quality of the MT solutions proposed in the included articles. In addition, it will be possible to draw from this analysis the most relevant insights in terms of benefits and limitations. Similarly, outlooks about the envisaged future trends will be provided, too.

5.1.1. Analysis of self-reported findings on MT quality

We began by considering how the authors of the included papers self-reported the findings of their works, (as listed in Table 13). In order to achieve a preliminary qualitative evaluation, we normalised those findings to a basic 5-item scale (i.e., *Very Poor*, *Poor*, *Fair*, *Good*, *Excellent*). It is important to highlight that each article was placed on that scale by simply considering the conclusions provided by the corresponding authors, thus not adding any further evaluation element. This scale will be then compared to the quality assessment scoring described in Section 3.6.2.

At the two extremes of the qualitative scale, we have two articles that reported excellent results (i.e., Li et al. (2020) and Skianis et al. (2020)) and one article that reported very poor results (Taira et al., 2021). In between, we have 7 articles in which the authors declared to have achieved poor results, 19 articles with fair results and 30 articles in which the authors assessed a good output.

Some considerations can be made starting from these self-reported evidences. First, the worst reported result was given in a study investigating the effectiveness of DeepL when translating a 242-sentence medical document in the Jap→Eng language pair (Takakusagi et al., 2021) so it did not involve any MT solution specifically designed to cope with medical texts but only a general-domain free MTE. The other studies that reported poor MT performances were characterised by some noteworthy aspects: absence of details about the evaluator type (Das et al., 2019; Ehab et al., 2018) or the target user type (Ehab et al., 2018; Yu and Zhu, 2021), which do not allow to determine to what extent those findings are totally reliable; large multi-language translation approaches (with 20 languages examined in Das et al. (2019) and 25 languages in Taylor et al. (2015)), which introduce a significant performance variety depending on the language pair and hinders any general consideration about effectiveness.

From a complementary perspective, fair and positive findings came from multi-faceted studies and it would be simplistic to generally assert that MT is always suitable to be applied in this field without further examination. Consequently, two considerations assume relevance: first, the more a given language pair and a given translation approach are investigated, the more elements are available to support the quality assessment of a given MT technology; second, the less investigated a given case study is, the more interest to it should be devoted by researchers before providing any conclusive evaluation. According to the former consideration, this MLR highlighted a specific pattern as the most studied in the field: text-to-text bilingual unidirectional MT in the Eng→Spa and Eng→Chi language combinations.

Table 13
Breakdown of included articles by reported findings about achieved overall MT quality, normalised to a 5-item quality scale.

Findings	Articles
Excellent	Liu and Cai (2015) and Soares et al. (2020)
Good	Almahasees and Jaccomard (2020), Nunzio et al. (2021), Ehab et al. (2019), Huck et al. (2017), Lester et al. (2021), Luger et al. (2020), Taira et al. (2021), Marais et al. (2020), Miller et al. (2018), Musleh et al. (2018), Mutal et al. (2020), Rani et al. (2019), Renato et al. (2018), Shin et al. (2015), Spechbach et al. (2019), Khoong et al. (2019), Turner et al. (2019), Chen et al. (2016), Way et al. (2020), Wołk and Marasek (2015a), Ziganshina et al. (2021), Lankford et al. (2022), San et al. (2022), Liu and Huang (2021), Alam et al. (2021), Jimeno Yepes et al. (2017), Bawden et al. (2020), Neves et al. (2018), Yeganova et al. (2021) and Neves et al. (2022)
Fair	Almahasees et al. (2021), Liu et al. (2020), Turner et al. (2015), Deep et al. (2021), Dew et al. (2015), Hayakawa and Arase (2020), Álvarez Vidal et al. (2021), Kumar et al. (2018), Li et al. (2020), Manchanda and Grunin (2020), Chen et al. (2017), Muhaxov et al. (2016), Skianis et al. (2020), Soto et al. (2019), Park et al. (2022), Kapoor et al. (2022), Hira et al. (2019) and Weng et al. (2019)
Poor	Das et al. (2019), Ehab et al. (2018), Taylor et al. (2015), Yu and Zhu (2021), Lee et al. (2023), van den Bercken et al. (2019) and Bojar et al. (2016)
Very Poor	Takakusagi et al. (2021)

5.1.2. Analysis of MT quality scoring

At this point, it seems evident that relying on the self-reported findings is definitely not adequate to provide a guideline on the actual efficacy of MT in healthcare or to ascertain how rigorous was the evaluation protocol adopted by the authors. Moreover, the amount of details presented in the articles exhibits a significant variability, as it ranges from few specifications in some cases (e.g., Almahasees and Jaccomard (2020)) to a robust and thorough description as in the works presented at the WMT conference (to this purpose, see especially Neves et al., 2022).

Therefore, we decided to exploit the thorough multi-criteria analysis performed so far by applying the four domain-agnostic scoring criteria proposed in Marie et al. (2021) (indexed from MTQA1 to MTQA4), enriched with three more questions addressing the scope of MT in healthcare (indexed from MTSH1 to MTSH3), as detailed in Section 3.6.2. In Fig. 9, all the included articles are listed along with their scoring. MTSH scores are placed on the left, while MTQA and the overall score (given by the sum of MTQA and MTSH) are placed on the right side of the chart. Scores are presented as horizontal data bars (one per article) and the specific scoring questions are given a symbolic depiction so to make clear what questions contributed to the different scores and whether those questions were answered completely, partially, or not at all.

As for the MT quality assessment questions (MTQA), the majority of included articles scored a value greater or equal than 2 points, thus indicating a satisfactory overall quality, with four articles achieving the full score (i.e., Alam et al. (2021), Álvarez Vidal et al. (2021), Bawden et al. (2020), and Neves et al. (2022)) and eight papers scoring only 1 point (i.e., Almahasees and Jaccomard (2020), Almahasees et al. (2021), Hira et al. (2019), Kapoor et al. (2022), Kumar et al. (2018), Liu et al. (2020), Manchanda and Grunin (2020), and Mutal et al. (2020)). However, several differences must be highlighted. First, a different distribution of MTQA questions is noticeable. MTQA1 (i.e., manual evaluation of other automatic metrics better related to human rating than BLEU) was answered in the majority of included articles ($n = 49$), while statistical significance testing (i.e., MTQA2) was present only in 16 articles (27%) thus indicating that it is still not widely perceived as a fundamental element in the analysis of translation quality in MT. MTQA3 is instead the least considered element in the included articles, as only six papers (i.e., Alam et al. (2021), Álvarez Vidal et al. (2021), Bawden et al. (2020), Neves et al. (2022), Skianis et al. (2020), and Yeganova et al. (2021)) explicitly proposed either a direct computation of the automatic metrics imported from other works or adopted SacreBLEU-like metrics to guarantee comparability. Contrarily, MTQA4 was always given the full score of 1 point as in many cases it was not applicable (i.e., when a comparison among multiple MT solutions was not proposed in the article).

For what concerns the questions addressing the scope of MT in healthcare (MTSH), the first aspect to highlight is that the full score was never reached: only two articles achieved 2.5 over 3 points (i.e., Li et al. (2020) and van den Bercken et al. (2019)), and five 2 over 3 points (i.e., Das et al. (2019), Lester et al. (2021), Muhaxov et al. (2016), Musleh et al. (2018), and Weng et al. (2019)), while the majority of articles scored only 0.5 points each. As for the specific questions, MTSH1 was answered at least partially in 40 articles, thus demonstrating that in the majority of cases either in-domain vocabularies or in-domain text corpora are used for training the proposed MT solution. An appropriate and extensive analysis of the risks associated with a wrong output of a (bio)medical MT solution is performed in fewer articles and very often marginally, with only five articles proposing an adequate discussion about that aspect (i.e., Das et al. (2019), Khoong et al. (2019), Marais et al. (2020), van den Bercken et al. (2019), and Weng et al. (2019)) Noteworthy, if we exclude the classical research aim of assessing how effective a MT solution is in translating in-domain document, 25 articles did not propose any research question specifically tailored to the healthcare domain (e.g., developing a MT solution to improve the quality of life of a specific category of patients, or to address a specific linguistic or societal issue). This is mainly due to the fact that the classical approach followed in articles dealing with MT (i.e., building/selecting the dataset, implementing the MT solution and assessing its output quality) is essentially domain-independent and, therefore, very often the authors do not start from a real-life need of specific category of users. As a further confirmation, we can notice that some of the articles with high scores in MTQA have low scores in MTSH (this is the case of Álvarez Vidal et al. (2021), Bawden et al. (2020), and Neves et al. (2022)), thus indicating that good MT quality assessment were performed not necessarily in conjunction with adequate consideration of the healthcare domain.

Table 14
Evaluation and findings.

Article	Manual evaluation	Automatic evaluation	Validation dataset*	Findings
Alam et al. (2021)	–	BLEU, chrF, BERTscore, COMET, 1-TERm	600 COVID-19 terms	Terminology compliance in biomedical MT does not hamper general translation quality, as long as the terminology is of adequate standards.
Almahasees and Jaccomard (2020)	Adequacy, fluency	–	COVID-19 texts (n.q.)**	FTS committed fewer errors (adequacy/fluency). Further studies needed.
Almahasees et al. (2021)	TQA, error analysis	–	COVID-19 texts (n.q.)	Google Translate decent but with multiple errors. MTPE needed. Human translators not replaceable.
Álvarez Vidal et al. (2021)	Terminology, style, accuracy, fluency	BLEU, NIST, RIBES, WER	1d on oncological treatment (791w)	NMT more meaningful than SMT, both have fluency issues in healthcare.
Bawden et al. (2020)	TQA	BLEU, chrF	biomedical abstracts and terminologies (variable size)	Increased BLEU scoring compared to previous WMT edition.
Bojar et al. (2016)	TQA	BLEU	500 biomedical docs.	Translation quality is poor in comparison to the reference translations.
Nunzio et al. (2021)	Correctness	BLEU	3d (specialised) + 3d (popular science)	Google Translate better overall. Yandex more cultural-specific. Neither naturally sounding, accurate enough.
Chen et al. (2016)	Fluency, adequacy, meaning	–	1d on diabetes patient education (6sn)	Google Translate better for Eng→Spa than Eng→Chi (accuracy, fluency, errors)
Chen et al. (2017)	Fluency, adequacy, meaning	–	1d on health education (9sn)	Potentially supplementing professional translators, further evidence needed. Caution required in healthcare.
Das et al. (2019)	Accuracy	–	9 bulleted statements in AAP safety guidelines handouts	Google Translate not accurate. Not recommended but sometimes the only option for LEP patients.
Deep et al. (2021)	–	BLEU, TER	Medical texts (900sn)	Bidirectional LSTM-NMT better than NMT baseline. Error rate not negligible.
Dew et al. (2015)	Functionality, usability	–	Public health materials (n.q.)	Feasible to increase multilingual public health material. More quality-translated documents for training needed.
Ehab et al. (2018)	–	BLEU	Internal medicine publications (n.q.)	EBMT inadequate, lower accuracy than Google Translate.
Ehab et al. (2019)	–	BLEU	Medical texts (n.q.)	EBMT+TM performed well. Morphological adjustments to Arabic needed.
Hayakawa and Arase (2020)	Addition, omission, mistranslation, terminology, grammar	BLEU	Medical documents (2.5k sn)	More MT errors in professional documents, less in general-public ones.
Hira et al. (2019)	–	BLEU	biomedical documents (9.2k +10.9k sn)	Tokenisation is an important pre-processing to improve NMT performances

(*continued on next page*)

Cumulatively, we identify the best-performing articles as those achieving 5 over 7 points ($n = 4$: Álvarez Vidal et al. (2021), Bawden et al. (2020), Das et al. (2019), and Neves et al. (2022)) and 4.5 over 7 points ($n = 6$: Alam et al. (2021), Li et al. (2020), Renato et al. (2018), Skianis et al. (2020), van den Bercken et al. (2019), and Way et al. (2020)).

Table 14 (*continued*).

Huck et al. (2017)	TQA	BLEU	HimL set (3k sn)	Significant improvements by adding parallel and synthetic corpora.
Kapoor et al. (2022)	Users' satisfaction	–	Pre-formulated patient questions (n.q.)	High satisfaction in patients, good satisfaction in nurses. GT suitable adjunct tool.
Khoong et al. (2019)	Accuracy, readability, medical jargon, content, clinical significance, risk rating	–	100 free-texted ED discharge instructions (647sn)	Google Translate effective but still potentially risky. Guidelines/warnings beneficial.
Kumar et al. (2018)	–	BLEU, TER	Medical texts (3k sn)	Domain-specific parallel corpus improved performances of SMT baseline.
Lankford et al. (2022)	–	BLEU, TER, ChrF	Official health documents (strategy statements and annual reports) + COVID-related data (250 lines)	When translating health-related data for low-resource languages, in-domain datasets are beneficial.
Lee et al. (2023)	Accuracy, fluency, meaning, clinical risk, TQA	–	105 phrases (3 sn each) per language	MI quality inferior to professional medical interpreters (who should support with medical staffs).
Lester et al. (2021)	TQA	–	300 e-prescriptions directions	MT usable but with high-risk errors (e.g., dosage). Professional pre-editing required.
Li et al. (2020)	Correctness	BLEU-4, METEOR	36k e-prescriptions	95%-reliable, deployable for automatic e-prescriptions simplification.
Liu and Cai (2015)	Correctness	BLEU	3 EHRs	HMT worse than SMT but valid on EHR notes. Further research needed.
Liu et al. (2020)	–	BLEU, accuracy, perplexity	Training corpora (770k sn)	NLP, auxiliary dictionaries improved performances, not suitable without MTPE.
Liu and Huang (2021)	TQA	BLEU	96k medical sentences	In-domain corpora improves quality of baseline biomedical MT trained on out-of-domain data.
Luger et al. (2020)	TQA	BLEU	Medical documents (20sn)	With primarily-spoken languages, reference sets/standards needed to improve.
Manchanda and Grunin (2020)	–	BLEU	General-domain (14k sn) + in-domain (140k sn)	General-purpose NMT customisable to specialised-domain, enterprise-suitable scenarios.
Marais et al. (2020)	Phrase, Word Error Rate (PER,WER), time	–	Clinical documents (195 sn)	Idiomatic, domain-appropriate translations achieved. Positive user feedback, more target languages beneficial.
Miller et al. (2018)	Quality, text safety	–	100 patient care instructions (3sn each)	Google Translate relatively safe. More safety errors in medication instructions.
Muhaxov et al. (2016)	Correctness	–	28sn suitable for EHRs	Feasible for doctor↔patient dialogues and automatic HER creation.

(*continued on next page*)

A conclusive element of investigation is given by the potential mismatch that can exist between the reported findings and the overall quality scoring in a given article. Therefore, we compared MTQA+MTSH assessments against what was declared by the authors of every article (see the rightmost column in Fig. 9). We assumed that the works reaching the cumulative quality score of less than 3 over 7 points, and having at the same time a declared good or excellent MT output, present a significant mismatch in terms of either poor in-domain scope or unavailability of enough element to assess the scientific soundness in a rigorous and shareable way. Such a mismatch was identified in five papers (i.e., Almahasees et al. (2021), Nunzio et al. (2021), Huck et al. (2017), Mutal et al. (2020), and Yeganova et al. (2021)). In other words, we are not saying here that the outcomes reported in

Table 14 (*continued*).

Musleh et al. (2018)	Usefulness, error type	BLEU	Doctor↔Patient communications (636sn)	Improved significantly the baseline. Syntax/lexical errors remain.
Mutal et al. (2020)	–	Sentence error rate (SER)	Medical interviews (44k sn)	Custom model translated 88% elliptical utterances correctly.
Neves et al. (2018)	TQA	BLEU	50 biomedical abstracts per language	Increased BLEU scoring compared to previous WMT editions, especially in Ger, Spa.
Neves et al. (2022)	TQA	BLEU, SacreBLEU, COMET, METEOR, ROUGE	(bio)medical abstracts (50 per lang. pair) + 150 EHR-like COVID-19 clinical cases (Eng→Spa only)	Increased BLEU scoring compared to previous WMT editions; difficult comparison between different approaches of WMT teams; environmental impact of MT computation not considered.
Park et al. (2022)	TQA	String matching	RadLex terms (65k EN; 47k DE)	Useful for some radiological text multi-lingual translation. Translation direction affects accuracy.
Rani et al. (2019)	Syntactic, semantic correctness	–	150 pathology reports on breast cancer	Proposed MT augmented LEP-patient inclusiveness. Deployable in regional hospitals. Google Translate sometimes better.
Renato et al. (2018)	Error analysis	BLEU, METEOR, NIST	Medical publications (1.2k w)	SMT trained with domain-specific corpora significantly outperforms general-purpose MTEs.
San et al. (2022)	–	BLEU4, GLEU, WER, CER	500 medical sentences (patient-doctor conversations, disease names, etc.)	Data-augmented in-domain corpora are helpful in medical MT for low-resource languages.
Shin et al. (2015)	Translation success rate	–	10sn (simplified) on patient symptoms	Technically correct but unnatural expressions. Improvements required.
Skianis et al. (2020)	TQA	BLEU, SacreBLEU, METEOR, TER	ICF terms (25k w)	Customisable to other languages. Fast, effective medical terminology translation.
Soares et al. (2020)	TQA (critical mistranslation, vocabulary adequacy)	–	(bio)medical articles from PubMed (n.q.)	MT output usable but not fluent. Validation and MTPE needed to avoid harmful mistranslations.
Soto et al. (2019)	TQA, fluency, accuracy	BLEU	EHR (n.q.)	Suitable for EHRs, even without bilingual corpora.
Spechbach et al. (2019)	Usefulness, usability, use time	–	2 lists of symptoms	Good alternative if no ER interpreters available. Doctors prefer text, patients speech.
Taira et al. (2021)	Fluency, adequacy, meaning, severity	–	20 ED discharge instructions	Google Translate inconsistent between languages, not reliable for patient instructions.
Takakusagi et al. (2021)	Accuracy, mistranslation	–	1d on radiotherapy (242sn)	DeepL is accurate for Jap→Eng.
Taylor et al. (2015)	TQA	–	1q (25sn) + 1is (600w)	Not enough quality in healthcare without MTPE.

(*continued on next page*)

those papers cannot be relied on but we are highlighting how challenging can be their comparison against other works in the same field.

5.1.3. Current limitations of MT in healthcare

This review only considered English DLs and articles written in English (indeed, articles such as Trujillos-Yébenes and Muñoz-Miquel, 2022 were excluded depending on the language-related CEx): these initial assumptions surely influenced the unbalanced presence of English as the source language (and, to a lesser extent, as the target language) in the majority of the examined MT solutions evaluated in the included articles.

Table 14 (*continued*).

Turner et al. (2015)	Error analysis, MTPE time, quality	–	60d on health promotion	Eng→Spa MT not as effective as Eng→Chi.
Turner et al. (2019)	Functionality, usability	–	6 scripted emergency scenarios	QuickSpeak better than Google Translate. Better accuracy/usability needed before on-site use.
van den Bercken et al. (2019)	Grammar, meaning preservation, simplicity	BLEU, SARI	Biomedical texts about diseases (n.q.)	Conflicting and poor results demand further research. UMLS CUIs replacement lowered translation quality.
Way et al. (2020)	Lexical choices, fluency, omission	BLEU, chrF2	Training corpora (1k sn) + ad-hoc texts (100sn)	Proposed NMT achieved similar (sometimes better) performances of public MTEs.
Weng et al. (2019)	Correctness, readability	–	Free-text patient discharge notes (n.q.)	Increase in readability, correctness slightly lower than dictionary-based word replacement solutions. Over-simplified translation possibly harmful.
Wołk and Marasek (2015a)	–	BLEU, NIST, METEOR; TER	EMA leaflet (1k sn)	NMT promising. GPUs may improve computational feasibility.
Yeganova et al. (2021)	TQA	BLEU	50 biomedical docs. per lang.	Better quality than baseline, Eng → Chi still challenging.
Jimeno Yepes et al. (2017)	TQA	BLEU	biomedical documents (n.q.)	Increased BLEU scoring compared to previous WMT edition.
Yu and Zhu (2021)	Translation speed/update, accuracy	–	Medical texts (70sn)	NMT/SMT not widely applicable (syntactic and morphological features in healthcare).
Ziganshina et al. (2021)	TQA	–	90d (Cochrane PLS on health information)	Google Translate, DeepL better than MS Bing. MTPE required anyway.

Notes: *Unit of measurement: d = documents/articles, sn = sentences, sg = segments, w = words, t = terms, q = questionnaires, is = information sheets. **(n.q.) = size not quantified.

Nevertheless, the Eng→Spa and Eng→Chi language combinations are so largely studied also because of the presence of large LEP communities in English-speaking countries, while the same does apply neither to other language combinations, nor to the inverse direction (i.e., Spa→Eng and Chi→Eng). This for sure requires more research works and careful considerations before adopting MT also in different healthcare/clinical contexts.

As for what concerns the other inherent limitations of the included studies, their *maturity level* is the main challenge. With this concept, we refer to how preliminary the study is, in terms of its extension, scale, involved resources, and performed evaluations. A pattern common to the majority of articles, except those of the WMT conference, as they exhibit a significant level of completeness, reveals that the proposed MT solutions were analysed in pilot studies involving either relatively few source texts or few target users, lasting for short time periods only. This is definitely an aspect to tackle when designing similar studies so to help final users to decide whether a MT solution is worthy of adoption. Consequently, we suggest healthcare professionals willing to use MT to verify at first whether the intended language combination and use settings have been already addressed in the current literature (Section 4). If affirmative, it should be determined whether the number of involved evaluators in the study is adequate. We believe that without a certain amount of practitioners and translators involved in the evaluation processes, possibly during multiple evaluation rounds, the interested healthcare professionals should replicate the experiments with a larger number of evaluators before taking a final decision. Since this is not always possible for obvious reasons (limited funds, lack of skilled human personnel, etc.), the first step should be to engage target language domain experts and support them, whenever possible, with bilingual domain experts or professional translators. An alternative and even better approach would be to involve certified medical translators. In addition, healthcare professionals should prioritise those studies where both manual and automatic evaluation procedures were applied.

Another limitation highlighted by the insights coming from Section 4 refers to the need for human editing during the post-translation stage. When used, the authors found an improvement in the overall MT output quality if compared against the output of fully-automated MT. The typology of involved post-editors is another, closely related limitation: when relying on professional translators only, the risk of overlooking in-domain misconceptions is very often looming, as well as the risk of accepting unnatural translations if in-domain experts of the target language are involved as the only typology of evaluators.

These considerations bring to us another important question: when is it possible to benefit from human post-editing? This is true for sure in two scenarios. First, when human post editing is applied to the MT output before performing another round of training with the post-edited materials. Second, when the MT output has not to be used immediately: automatically translated

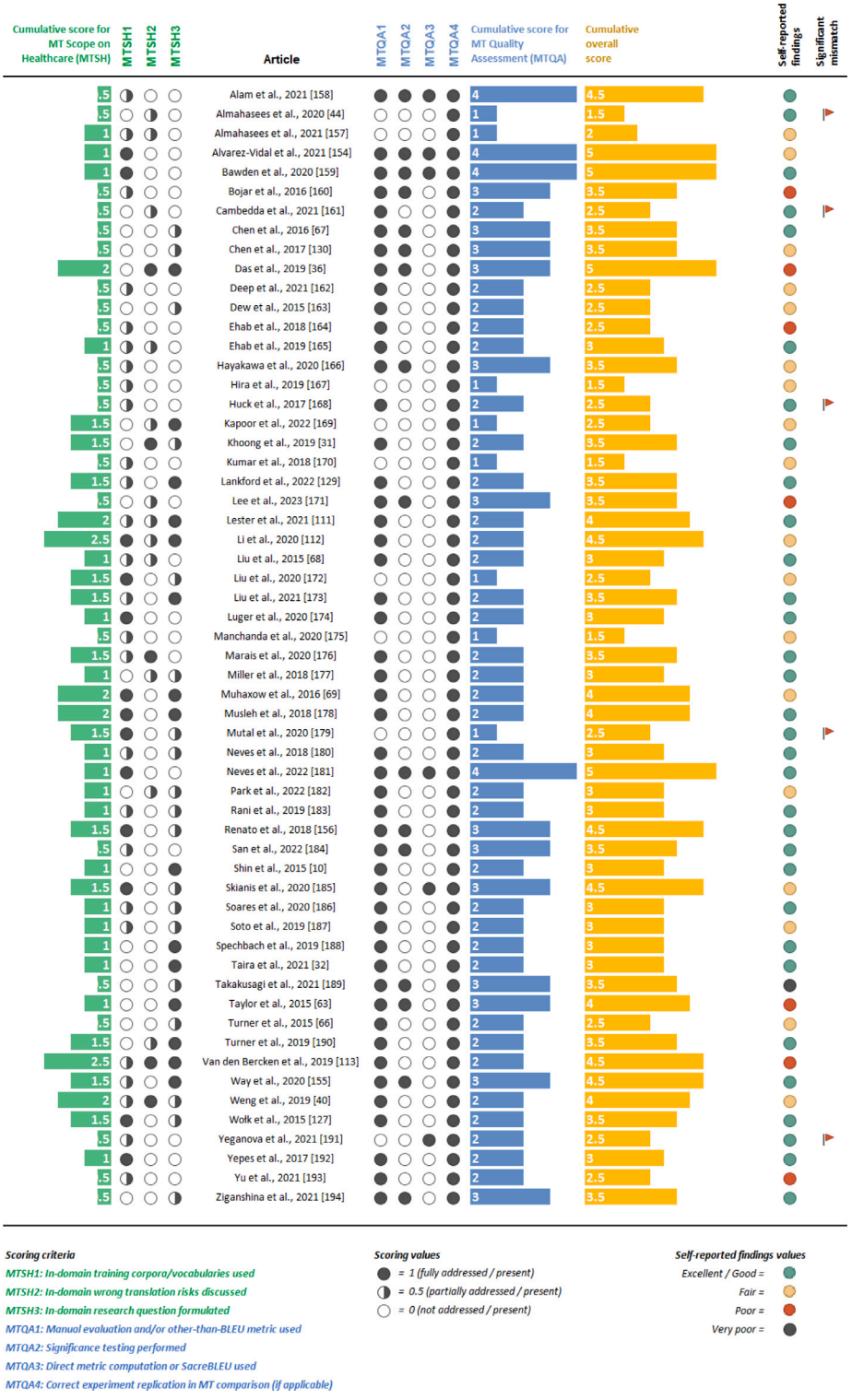

Fig. 9. Overall quality scoring.

patient-guidance texts or public health documents can be post-edited before their dissemination, as well as automatically translated biomedical scientific articles can be revised before being distributed to students. In some other cases, this is not possible. For instance, in real-time speech-to-speech communications among medical staff and patients, human post-editing cannot be relied on. In those

cases, the absence of proper MT training mechanisms can hamper the overall output quality and complementary solutions should be considered (see Section 5.1.4).

Finally, even if not strictly domain-related and as also noticed in Neves et al. (2022), the analysis of the environmental impact determined by MT solutions is not considered in any article even if it is expected to become a challenging aspect, especially when considerable computational resources are required for complex training procedures of NMT solutions.

5.1.4. Advisable risk-mitigation strategies and future trends

All the hindrances mentioned in Section 5.1.3 are intertwined with the risk of producing unsafe translations, which is the major element of concern when MT is used in healthcare and which is primarily associated with the typology of source materials to translate.

Among the so-called high-risk contents (either written or verbal), we can list informed consent/authorisations for medical treatments (Glaser et al., 2020; Ochieng et al., 2015), patient-handling procedures (Garzillo et al., 2020), and anamneses of patients belonging to specific categories (e.g., minor patients, patients with cognitive impairments, etc.) (Palkova and Semaka, 2016). Even if a properly-trained, customised NMT solution is theoretically capable of producing acceptable results for these source texts, most physicians would perceive the risk of causing significant harm with unreliable machine translations as too great and would therefore be unwilling to take it.

Therefore, we believe that the results in this MLR along with the current technological advancements encourage a different and a less error-prone use of MT in healthcare, which is given by MT applied to lower-risk texts. Daily interactions between medical staff members and patients, written communications between doctors and patient''s relatives, basic instructions for outpatients entering the hospital for medical examinations, and even simple patient dressing instructions can greatly benefit from MT as they do not exhibit only peculiar medical terminologies but they also features non-healthcare language style, thus making the MT contribution more effective (since this would make also general-domain training datasets more relevant).

The risk of getting potentially harming translations also varies depending on the timing they are required. Real-time speech-to-speech interactions are those showing the highest risk. First, advanced speech recognition able to cope not only with multiple language inflections but also with noisy environments such as hospitals or ambulances are needed. Second, speech synthesis functionalities are also necessary to provide end users with the translation outputs. Third, the MT solution cannot benefit from human post editing or from the presence of supporting medical translators to ensure a certain level of translation quality. Therefore, we believe that an adequate compromise to handle these situations, especially when first-responders are involved, is represented by phrase-based MT, since the majority of these interactions would revolve around typical communication patterns that can be supported by a set of pre-translated phrases. The phrase-based MT solution can be then incorporated in a mobile app offering speech-to-speech functionalities or, in an even more interesting configuration, directly into the first responder''s equipment (e.g., a protective face mask or helmet with a speech synthesiser).

Other risk-mitigation approaches can be envisaged by appropriately leveraging on technological advancements and computational linguistics in order to boost the adoption of MT in healthcare thanks to the improvement of its output quality. Among them, it is worth to list: performing morphological adjustments to specific language pairs (Ehab et al., 2019), increasing the size and quality of training datasets (Dew et al., 2015), training MT solutions with domain-specific parallel text corpora (Kumar et al., 2018) whenever possible or augmenting already existing corpora with synthetic datasets (Huck et al., 2017), adding auxiliary dictionaries (Liu et al., 2020), exploiting only reference/standardised training datasets (Luger et al., 2020), extending the feasibility study to more language pairs (Marais et al., 2020), using GPUs to improve computational efficiency (Wołk and Marasek, 2015a), introducing professional pre-editing at the training stage (Lester et al., 2021), supplying end users with guidelines that make them aware of the safety risks of MT solutions for healthcare (Khoong et al., 2019).

Furthermore, an aspect rarely discussed in the examined literature is the importance of user-training as a further way to reduce the risk. Depending on the context of use and on the user typology, a given MT solution could be proposed in different formats or mediated through different devices/applications and, consequently, end users should be properly trained in advance so to know when and how to using as well as when avoiding the MT depending on the context they are working in.

5.2. Designing studies on MT in healthcare

A relevant insight emerging from the results discussed so far, is the absence of a shared and standardised way of designing studies on MT in healthcare. Although some trends and some commonalities are identifiable in the included studies, their variety is still noteworthy and the definition of an analysis workflow is challenging. Therefore, a set of guidelines aimed at supporting researchers who want to address this field is now proposed. These guidelines are directly derived from the analyses performed in the MLR, used as the common ground along with the MT quality assessment scoring proposed in Marie et al. (2021).

The very first aspect to consider is the study type: it can be either a comparison of already existing MT solutions (either general-domain or in-domain), or the design of a novel MT solution, or a mix of these two typologies. Depending on the study type, some of the analysis parameters described in Section 3.6 become more important than others and vice versa. Some other parameters are, instead, always fundamental and could be defined as *application-agnostic parameters*.

Let us start from application-agnostic parameters, which should consist of all those reported in every article included in the MLR (see Fig. 2). Consequently, any new study on MT in healthcare should be necessarily described in terms of:

- translation context (i.e., language pair, translation direction and approach, translation timing and type, and application scenario),

- user context (i.e., target user type and corresponding language group),
- MT technologies involved,
- case study settings (i.e., deployment type and time period),
- source documents and associated risk of wrong translation (which varies depending on the document type),
- adopted validation and post-editing procedures,
- details about how the solution is deployed,
- achieved overall findings.

Now, let us focus on the case of comparison studies. In this scenario, comparison criteria assume a crucial role: they should be selected in order not only to comply with typical translation quality assessment in general-domain MT, but also to provide an effective guidance to domain-specific end users. It is widely known that, when MT suitability is investigated as a general-domain problem, the de-facto choice is given by BLEU-like and METEOR-like metrics, which are automatically calculated on the target texts. In the majority of articles discussed in Section 4, the same choice is applied. We believe that the applicability of MT to the medical field requires, instead, either the mandatory adoption of both automatic and manual evaluation, or at least the use of automatic metrics offering better human ranking behaviour than BLEU. Similarly, comparison studies should present statistical significance testing as a reliable way to ensure the achieved results are not coincidental, as suggested in Marie et al. (2021), and this assumes an even greater importance when the (bio)medical field is entailed.

Moreover, depending on the specific type of target users, the manual evaluation can be further enriched. If the MT is applied to clinical communications (e.g., doctor-to/from-patient, medical staff-to/from-patient, patient-to-patient), the manual translation quality assessment should be performed as a multi-level process, involving bilingual domain experts in conjunction with professional translators/interpreters (or certified medical translators) so that not only in-domain translation appropriateness but also cultural and context adequacy are checked. If the MT is applied to public health documents or to scientific (bio)medical publications, just the bilingual domain experts could be involved. In both cases, the evaluation should rely on unbiased and objective metrics, possibly gathered in an assessment rubric, based on shared translation error taxonomies, so that the evaluators'' outputs are actually comparable.

The second scenario refers to the proposal of novel MT solutions: additional parameters are needed to design properly those studies. First, an adequate level of detail about the adopted implementation has to be provided. Then, the MT training approach must be detailed, both qualitatively and quantitatively. In the MLR we highlighted that some articles proposing ad hoc MT solutions did not discuss training procedures properly, while they should be always detailed to make them understandable and reproducible. As for the training type, depending on whether the MT is required for terminology translation or sentence/document translation, only vocabularies or vocabularies plus parallel aligned text corpora are required, respectively. Adopted vocabularies should always incorporate at least (some subsets of) UMLS and SNOMED-CT. In both cases, in-domain datasets should be used to improve the MT effectiveness. As for what concerns the size of the training datasets, dictionaries and corpora having only few hundreds items should be avoided (Costa-jussà et al., 2012), as they expose the trained MT solution to overfitting. When the dataset size is very limited, appropriate data enrichment strategies should be adopted, so that the original number of data items can be improved with additional synthetic datasets. In any case, the dataset size (in words/terms for vocabularies and in sentences/segments for text corpora) should be always reported explicitly in the study along with the average segment/sentence length in words for text corpora. Moreover, only validated and consistent in-domain training datasets should be used, so to avoid any additional time-consuming preparatory pre-editing effort.

According to Section 4, many of the papers included in the MLR did not present any comparison, while we believe that validation should be always performed also in this typology of studies, by identifying a baseline against which the proposed MT approach can be compared. The baseline should be selected among already available in-domain alternative MT solutions. Very often, the baseline is selected exclusively among commercial/free MT general-domain engines: this should be allowed only if the selected MT engine has already proven its effectiveness in the healthcare field, otherwise this could result in setting an excessively low baseline. However, if these types of MT engines are the only feasible option, the general-domain baseline should be compared to multiple versions of the proposed in-domain MT solution, so to ascertain whether a progressive improvement in quality is achieved. Moreover, general-purpose free MT engines could be considered a suitable baseline also because they are typically used in daily life activities.

Again, any comparison involving datasets used to ascertain the quality of the identified MT alternatives should be conducted to ensure it is reproducible and scientifically credible, thus requiring to exploit the same datasets or at least the same computational steps (Marie et al., 2021).

The challenge of performing an effective comparison among different MT solutions in healthcare is perfectly expressed also in Neves et al. (2022), one of the best-scoring articles in this MLR, where the contributions of several competing groups are compared to illustrate the findings of the WMT 2022 shared tasks in biomedical translation:

In terms of validation, the same considerations reported above for studies on MT comparison apply also to this case, so that the MT quality assessment can be finally performed and the post-editing stage can follow. The workflow in Fig. 10 summarises the guidelines proposed in this subsection.

6. Conclusion

In this paper, a methodological literature review (MLR) on the use of machine translation in the healthcare and medical sector has been proposed, by complying with the guidelines about how to build a literature review in the computer science domain Carrera-Rivera et al. (2022). By querying four scientific online digital libraries (namely, IEEE Xplore, ACM Digital Library, Scopus, and

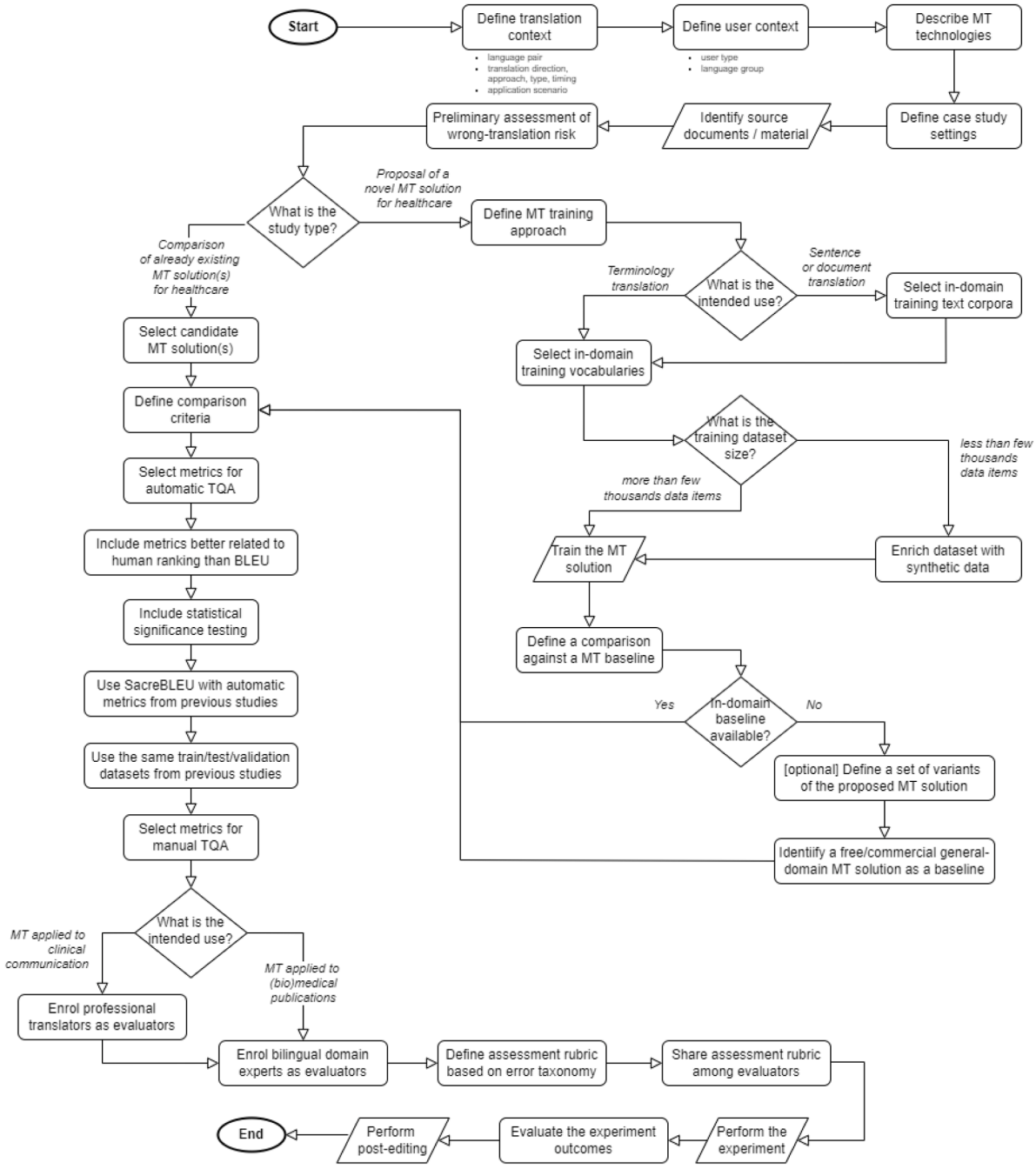

Fig. 10. Operational workflow supporting the design of a case study in MT for healthcare.

PubMed), and referring to the January 2015–February 2023 period, we initially collected 565 candidate articles, further extended with 6 more articles from a previously published literature review and 4 articles manually selected because of their importance. The articles were then managed and screened according to the *Preferred Reporting Items for Systematic reviews and Meta-Analyses* (PRISMA) methodology (2020 version) so that a final dataset of 58 items was gathered. Subsequently, seven classes of analysis criteria were defined to analyse the articles quantitatively and qualitatively.

The adopted criteria covered all the main research domains entailed by medical MT: languages, approaches, and scenarios; technologies; training procedures; target population and language groups; study deployment, scope, and material; evaluation processes; and overall findings. All the articles were mapped against these criteria, thus achieving a clear depiction of what articles

overlooked a given criteria and, similarly, of what aspects were rarely mentioned in the scientific literature, so to provide researchers with a list of gaps that new studies could try to address and close.

The analysis outcomes highlighted a substantial prevalence of English-to-Spanish (n = 19) and English-to-Chinese (n = 16) MT solutions for healthcare professionals alone (n = 14) or along with patients (n = 12), while fewer were dedicated to the general public (n = 5). An unbalanced distribution between MT applied to clinical communication (n = 21) and to health education (n = 37) was also ascertained. In terms of translational approaches, the unidirectional real-time bilingual MT was definitely the most frequent one (n = 24), implemented especially via custom NMT (n = 13) or free NMT (n = 9) solutions. Google Translate (n = 22), OpenNMT (n = 12), and Moses (n = 11) were the most frequently discussed implementation solutions also because they were adopted as baselines in several studies. Training evaluation processes exhibited the most significant heterogeneity in terms of typology and size of used resources. As for the evaluation, quantitative analysis (n = 23) or mixed quantitative–qualitative analysis (n = 23) was the preferred choice. Many studies (n = 46) proposed a manual evaluation approach to assess MT quality and accuracy, while automatic procedures were adopted fewer times (n = 33), mostly relying on the BLEU metric or its variants. Finally, a considerable generalised lack of details on deployment, pre-editing, and post-editing was also observed.

In order to assess their MT quality, the articles included in this MLR were also examined depending on the scientific credibility assessment score presented in Marie et al. (2021), enriched with an additional score aimed at quantifying the effective scope of those solutions in the healthcare domain. Nearly 10% of the articles reached an adequate overall ranking thus demonstrating their validity.

To sum up the achieved insights, our MLR highlighted clearly how MT (particularly NMT) is gaining strength as a helpful resource in the absence of professional translators/interpreters and how several studies pointed out its effectiveness. However, reliability concerns still exist when highly-sensitive texts (e.g., patient consent modules, therapeutic procedures, etc.) are involved. A relevant part of the studies included in the MLR acknowledged the importance of careful preparation and management of the training aspects for custom MT solutions, since a properly prepared training dataset guarantees a substantial performance improvement. Overall, the necessity of performing appropriate post-editing human intervention is perceived as a key element along with preparatory pre-editing, even if the role of pre-editing was considered to a lesser extent.

Consequently, even if the MT is still perceived as not completely capable of replacing translators and interpreters in handling (bio)medical contents, we believe that this limitation can be leveraged to introduce MT in the healthcare practice not as a replacement but as a complementary support, mostly depending on the communication type: when asynchronous communication is involved, the combination of MT and post editing ensures a decent translation quality, already assessed in the examined studies especially for unidirectional bilingual MT of Eng→Spa and Eng→Chi (bio)medical texts, while MT in synchronous communications should be used with lower-risk contents and more customised solutions. Similarly, less-explored language combinations should be also investigated, so to improve language inclusiveness in the healthcare sector.

From an operational perspective, accessing non-English DLs represents a field of investigation that can be explored in a future companion study of this MLR. Similarly, considering new analysis criteria such as the presence of back-translation tasks, the use of language models, or the adoption of fine-tuning procedure, all represent further improvements to this study.

List of abbreviations

BLEU: Bilingual Evaluation Understudy

CAGR: Compound Annual Growth Rate

CE: Criterion of Exclusion

CL: Computational Linguistics

DL: Digital Library

DLQ: Query to Digital Library

EBMT: Example-based Machine Translation

FTS: Full-text Screening

GRU: Gated Recurrent Unit network

HT: Human Translation

ICT: Information and Communication Technology

LEP: Limited English Proficiency

LSTM: Long Short Term Memory network

METEOR: Metric for Evaluation of Translation with Explicit Ordering

MLR: Methodological Literature Review

MQM: Multidimensional Quality Metrics

MT: Machine Translation

MTE: Machine Translation Engine

NIST: US National Institute of Standards & Technology

NLP: Natural Language Processing

NMT: Neural Machine Translation

PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses

RBMT: Rule-based Machine Translation

RMS: Reference Management System

RNN: Recurrent Neural Network

SER: Sentence Error Rate

SLR: Systematic Literature Review

SMT: Statistical Machine Translation

TER: Translation Error Rate

TQA: Translation Quality Assessment

WER: Word Error Rate

CRediT authorship contribution statement

Marco Zappatore: Conceptualization, Data curation, Formal analysis (lead), Methodology, Project administration, Resources (lead), Software (lead), Supervision, Validation (lead), Visualization, Writing – original draft (equal), Writing – review & editing. **Gilda Ruggieri:** Formal analysis (supporting), Resources (supporting), Software (supporting), Validation (supporting), Writing – original draft (equal).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The links to the datasets used in this systematic literature review are available directly in the paper, in the form of shared Google Spreadsheets.

References

ACL, 2023. Acl anthology. search results. how does the search work? https://aclanthology.org/search/?q. [Online; accessed 27-February-2023].

Afzal, Z., Akhondi, S.A., van Haagen, H., van Mulligen, E.M., Kors, J.A., 2015. Biomedical concept recognition in french text using automatic translation of english terms. In: Conference and Labs of the Evaluation Forum.

Agrawal, T., Urolagin, S., 2020. 2-way arabic sign language translator using CNNLSTM architecture and NLP. In: Proceedings of the 2020 2nd International Conference on Big Data Engineering and Technology. ACM, http://dx.doi.org/10.1145/3378904.3378915.

Alam, M.M.I., Kvapilíková, I., Anastasopoulos, A., Besacier, L., Dinu, G., Federico, M., Gallé, M., Jung, K., Koehn, P., Nikoulina, V., 2021. Findings of the WMT shared task on machine translation using terminologies. In: Proceedings of the Sixth Conference on Machine Translation. Association for Computational Linguistics, Online, pp. 652–663, URL: https://aclanthology.org/2021.wmt-1.69.

Almagro, M., Martínez, R., Montalvo, S., Fresno, V., 2019. A cross-lingual approach to automatic ICD-10 coding of death certificates by exploring machine translation. J. Biomed. Inform. 94, 103207.

Almahasees, Z., Jaccomard, H., 2020. Facebook translation service (fts) usage among jordanians during covid-19 lockdown. Adv. Sci. Technol. Eng. Syst. 5, 514–519.

Almahasees, Z., Meqdadi, S., Albudairi, Y., 2021. Evaluation of google translate in rendering english covid-19 texts into arabic. J. Lang. Linguist. Stud. 17, 2065–2080.

Alvarez, S., Oliver, A., Badia, T., 2020. Quantitative analysis of post-editing effort indicators for NMT. In: Proceedings of the 22nd Annual Conference of the European Association for Machine Translation. European Association for Machine Translation, Lisboa, Portugal, pp. 411–420, URL: https://aclanthology.org/2020.eamt-1.44.

Álvarez Vidal, S., Oliver, A., Badia, T., 2021. What do post-editors correct? A fine-grained analysis of smt and nmt errors. Rev. Tradumatica 131–147.

Ambati, R., Dudyala, C.R., 2018. A sequence-to-sequence model approach for ImageCLEF 2018 medical domain visual question answering. In: 2018 15th IEEE India Council International Conference (INDICON). IEEE, http://dx.doi.org/10.1109/indicon45594.2018.8987108.

Amin, M.F., Plis, S.M., Damaraju, E., Hjelm, D., Cho, K., Calhoun, V.D., 2016. Multimodal fusion of brain structural and functional imaging with a deep neural machine translation approach. In: 2016 IEEE Southwest Symposium on Image Analysis and Interpretation (SSIAI). IEEE, http://dx.doi.org/10.1109/ssiai.2016.7459160.

An, B., Long, C., 2022. Paraphrase based data augmentation for chinese-english medical machine translation. Dianzi Yu Xinxi Xuebao/J. Electron. Inf. Technol. 44, 118–126.

Association for Computing Machinery, 2022. Acm digital library (dl) - advanced search. https://dl.acm.org/search/advanced. [Online; accessed 27-February-2023].

Bawden, R., Di Nunzio, G.M., Grozea, C., Jauregi Unanue, I., Jimeno Yepes, A., Mah, N., Martinez, D., Névéol, A., Neves, M., Oronoz, M., Perez-de Viñaspre, O., Piccardi, M., Roller, R., Siu, A., Thomas, P., Vezzani, F., Vicente Navarro, M., Wiemann, D., Yeganova, L., 2020. Findings of the WMT 2020 biomedical translation shared task: Basque, Italian and Russian as new additional languages. In: Proceedings of the Fifth Conference on Machine Translation. Association for Computational Linguistics, Online, pp. 660–687, URL: https://aclanthology.org/2020.wmt-1.76.

Beh, T.H., Canty, D.J., 2015. English and Mandarin translation using Google Translate software for pre-anaesthetic consultation. Anaesth. Intensive Care 43, 792–793.

Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huck, M., Jimeno Yepes, A., Koehn, P., Logacheva, V., Monz, C., Negri, M., Névéol, A., Neves, M., Popel, M., Post, M., Rubino, R., Scarton, C., Specia, L., Turchi, M., Verspoor, K., Zampieri, M., 2016. Findings of the 2016 conference on machine translation. In: Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers. Association for Computational Linguistics, Berlin, Germany, pp. 131–198. http://dx.doi.org/10.18653/v1/W16-2301, URL: https://aclanthology.org/W16-2301.

Caldwell, G., 2019. The process of clinical consultation is crucial to patient outcomes and safety: 10 quality indicators. Clin. Med. J. R. College Phys. Lond. 19, 503–506.

Canva UK Operations Ltd, 2022. Flourish | data visualization & storytelling. https://flourish.studio/. [Online; accessed 27-February-2023].

Carrera-Rivera, A., Ochoa, W., Larrinaga, F., Lasa, G., 2022. How-to conduct a systematic literature review: A quick guide for computer science research. MethodsX 9, 101895.

Chauhan, S., Saxena, S., Daniel, P., 2021. Fully unsupervised word translation from cross-lingual word embeddings especially for healthcare professionals. Int. J. Syst. Assur. Eng. Manag. 13, 28–37.

Chen, X., Acosta, S., Barry, A.E., 2016. Evaluating the accuracy of google translate for diabetes education material. JMIR Diabetes 1, e3.

Chen, X., Acosta, S., Barry, A.E., 2017. Machine or human? evaluating the quality of a language translation mobile app for diabetes education material. JMIR Diabetes 2, e13.

Chen, S.F., Goodman, J., 1999. Empirical Study of Smoothing Techniques for Language Modeling. Technical Report 4, Harvard University, http://dx.doi.org/10.1006/csla.1999.0128, arXiv:9606011.

Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y., 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 1724–1734. http://dx.doi.org/10.3115/v1/D14-1179, URL: http://aclweb.org/anthology/D14-1179.

Cochrane Consumer Network, 2023. Cochrane and systematic reviews. https://consumers.cochrane.org/cochrane-and-systematic-reviews#systematic. [Online; accessed 27-February-2023].

Cochrane Consumers and Communication, 2023. Animated storyboard: What are systematic reviews? https://cccrg.cochrane.org/animated-storyboard-what-are-systematic-reviews. [Online; accessed 27-February-2023].

Cornell University, 2022. arxiv e-print archive, https://arxiv.org/. [Online; accessed 27-February-2023].

Costa-jussà, Marta R., Farrús, Mireia, S. P, J., 2012. Machine translation in medicine. In: ARSA - Proceedings in ARSA - Advanced Research in Scientific Areas. pp. 1995–1998.

Crego, J., Kim, J., Klein, G., Rebollo, A., Yang, K., Senellart, J., Akhanov, E., Brunelle, P., Coquard, A., Deng, Y., Enoue, S., Geiss, C., Johanson, J., Khalsa, A., Khiari, R., Ko, B., Kobus, C., Lorieux, J., Martins, L., Nguyen, D.-C., Priori, A., Riccardi, T., Segal, N., Servan, C., Tiquet, C., Wang, B., Yang, J., Zhang, D., Zhou, J., Zoldan, P., 2016. SYSTRAN's pure neural machine translation systems.

Cui, X.-M., Gim, D., Han, S.K., 2020. Graphical illustration of the learning process in simple neural networks. New Phys.: Sae Mulli 70, 885–895.

Dant, J.T., Reeves, G., Stricklin, D., 2018. Automated translation of clinical parameters in evaluating acute radiation injury: Results from a mass casualty exercise. Disaster Med. Public Health Prep. 12, 569–573.

Das, P., Kuznetsova, A., Zhu, M., Milanaik, R., 2019. Dangers of machine translation: The need for professionally translated anticipatory guidance resources for limited english proficiency caregivers. Clin. Pediatr. 58, 247–249.

Datawrapper GmbH, 2022. Datawrapper: Create charts, maps, and tables. https://www.datawrapper.de/. [Online; accessed 27-February-2023].

Davis, S.H., Rosenberg, J., Nguyen, J., Jimenez, M., Lion, K.C., Jenicek, G., Dallmann, H., Yun, K., 2019. Translating discharge instructions for limited english–proficient families: Strategies and barriers. Hosp. Pediatr. 9, 779–787.

de Velde, S.V., Macken, L., Vanneste, K., Goossens, M., Vanschoenbeek, J., Aertgeerts, B., Vanopstal, K., Stichele, R.V., Buysschaert, J., 2015. Technology for large-scale translation of clinical practice guidelines: A pilot study of the performance of a hybrid human and computer-assisted approach. JMIR Med. Inform. 3, e33.

Deep, K., Kumar, A., Goyal, V., 2021. SMT versus NMT: An experiment with punjabi–english. pp. 63–71. http://dx.doi.org/10.1007/978-981-15-8297-4{_}6, URL: http://link.springer.com/10.1007/978-981-15-8297-4{_}6.

Denkowski, M., Lavie, A., 2014. Meteor universal: Language specific translation evaluation for any target language. In: Proceedings of the Ninth Workshop on Statistical Machine Translation. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 376–380. http://dx.doi.org/10.3115/v1/W14-3348, URL: http://aclweb.org/anthology/W14-3348.

Dew, K.N., Turner, A.M., Choi, Y.K., Bosold, A., Kirchhoff, K., 2018. Development of machine translation technology for assisting health communication: A systematic review. J. Biomed. Inform. 85, 56–67.

Dew, K., Turner, A.M., Desai, L., Martin, N., Laurenzi, A., Kirchhoff, K., 2015. Phast: A collaborative machine translation and post-editing tool for public health.. In: AMIA... Annual Symposium Proceedings. AMIA Symposium, 2015. pp. 492–501.

Du, Z., Zhang, X., Xie, D., 2020. Research on medical intelligent consultation based on question generation technology. In: 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, http://dx.doi.org/10.5220/0008979901450155.

Ehab, R., Amer, E., Gadallah, M., 2018. Example-Based English to Arabic Machine Translation: Matching Stage using Internal Medicine Publications. ACM Press, pp. 131–135. http://dx.doi.org/10.1145/3220267.3220294.

Ehab, R., Gadallah, M., Amer, E., 2019. English-arabic hybrid machine translation system using ebmt and translation memory. Int. J. Adv. Comput. Sci. Appl. 10, 195–203.

Elsevier, 2022a. Mendeley - reference management system. https://www.mendeley.com/. [Online; accessed 27-February-2023].

Elsevier, 2022b. Scopus - advanced search. https://www.scopus.com/search/form.uri?display=advanced. [Online; accessed 27-February-2023].

Evergreen, S., 2019. Effective Data Visualization: The Right Chart for the Right Data, second ed. Sage Publications, Inc..

Falissard, L., Morgand, C., Ghosn, W., Imbaud, C., Bounebache, K., Rey, G., 2022. Neural translation and automated recognition of ICD-10 medical entities from natural language: Model development and performance assessment. JMIR Med. Inform. 10, e26353.

Federal Coordination and Compliance Section (FCS), Civil Rights Division - US Dept. of Justice, 2015. Language map app. https://www.lep.gov/maps/lma2015/Final. [Online; accessed 27-February-2023].

Finley, G.P., Salloum, W., Sadoughi, N., Edwards, E., Robinson, A., Miller, M., Suendermann-Oeft, D., Brenndoerfer, M., Axtmann, N., 2018. From Dictations To Clinical Reports using Machine Translation, Vol. 3. Association for Computational Linguistics (ACL), pp. 121–128, URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85083432327&partnerID=40&md5=0853eb444189f0bd30b00c1dc8ced912 cited by: 8.

García, M.A.M., Rodríguez, R.P., Rifón, L.A., 2018. Leveraging wikipedia knowledge to classify multilingual biomedical documents. Artif. Intell. Med. 88, 37–57.

Garzillo, E.M., Monaco, M.G.L., Corvino, A.R., D'Ancicco, F., Feola, D., Ventura, D.D., Miraglia, N., Lamberti, M., 2020. Healthcare workers and manual patient handling: A pilot study for interdisciplinary training. Int. J. Environ. Res. Public Health 17, 4971.

Glaser, J., Nouri, S., Fernandez, A., Sudore, R.L., Schillinger, D., Klein-Fedyshin, M., Schenker, Y., 2020. Interventions to improve patient comprehension in informed consent for medical and surgical procedures: An updated systematic review. Med. Decis. Mak. 40, 119–143.

Global Market Insights, 2022. Machine translation market size by technology (statistical machine translation (smt), rule-based machine translation (rbmt), neural machine translation (nmt), hybrid machine translation (hmt), example-based machine translation (ebmt)), by deployment model (on-premise, cloud), by application (automotive, bfsi, e-commerce, electronics, healthcare, it & telecommunications, military & defense), covid-19 impact analysis, regional outlook, growth potential, competitive market share & forecast, 2022–2030. https://www.gminsights.com/industry-analysis/machine-translation-market-size. [Online; accessed 27-February-2023].

Guo, Z., Chen, C., 2021. Practice and research of computer-aided medical translation based on big data. J. Phys. Conf. Ser. 2004, 012014.

Haddaway, N.R., Collins, A.M., Coughlin, D., Kirk, S., 2015. The role of google scholar in evidence reviews and its applicability to grey literature searching. PLOS ONE 10, e0138237.

Hakami, H., Bollegala, D., 2015. A classification approach for detecting cross-lingual biomedical term translations. Natural Lang. Eng. 23, 31–51.

Handsel, J., Matthews, B., Knight, N.J., Coles, S.J., 2021. Translating the InChI: adapting neural machine translation to predict IUPAC names from a chemical identifier. J. Cheminform. 13.

Hartensuer, R., Nikolov, B., Franz, D., Weimann, A., Raschke, M., Juhra, C., 2015. Vergleich von ICD-10 und AIS mit der entwicklung einer methode zur automatisierten umwandlung. Z. Orthop. Unfallchirurgie 153, 607–612.

Hayakawa, T., Arase, Y., 2020. Fine-grained error analysis on english-to-japanese machine translation in the medical domain. pp. 155–164, URL: https://aclanthology.org/2020.eamt-1.17.pdf.

He, P., Meister, C., Su, Z., 2020. Structure-invariant testing for machine translation. In: Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering. ACM, http://dx.doi.org/10.1145/3377811.3380339.

He, P., Meister, C., Su, Z., 2021. Testing machine translation via referential transparency. In: 2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE). IEEE, http://dx.doi.org/10.1109/icse43902.2021.00047.

Heafield, K., Farrow, E., van der Linde, J., Ramírez-Sánchez, G., Wiggins, D., 2022. The EuroPat corpus: A parallel corpus of European patent data. In: 2022 Language Resources and Evaluation Conference, LREC 2022. pp. 732–740.

Henriques, B.C., Buchner, A., Hu, X., Wang, Y., Yavorskyy, V., Wallace, K., Dong, R., Martens, K., Carr, M.S., Asl, B.B., Hague, J., Sivapalan, S., Maier, W., Dernovsek, M.Z., Henigsberg, N., Hauser, J., Souery, D., Cattaneo, A., Mors, O., Rietschel, M., Pfeffer, G., Hume, S., Aitchison, K.J., 2021. Methodology for clinical genotyping of CYP2d6 and CYP2c19. Transl. Psychiatry 11.

Hill, D.C., Gombay, C., Sanchez, O., Woappi, B., Vélez, A.S.R., Davidson, S., Richardson, E.Z.L., 2022. Lost in machine translation: The promises and pitfalls of machine translation for multilingual group work in global health education. Discov. Educ. 1.

Hira, N.-e., Abdul Rauf, S., Kiani, K., Zafar, A., Nawaz, R., 2019. Exploring transfer learning and domain data selection for the biomedical translation. In: Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2). Association for Computational Linguistics, Florence, Italy, pp. 156–163. http://dx.doi.org/10.18653/v1/W19-5419, URL: https://aclanthology.org/W19-5419.

Hirschberg, J., Manning, C.D., 2015. Advances in natural language processing. Science 349, 261–266.

Hitachi Vantara, 2023. Pentaho community edition. https://www.hitachivantara.com/en-us/products/dataops-software/data-integration-analytics/pentaho-community-edition.html. [Online; accessed 27-February-2023].

Huck, M., Braune, F., Fraser, A., 2017. Lmu Munich's Neural Machine Translation Systems for News Articles and Health Information Texts. Association for Computational Linguistics, pp. 315–322. http://dx.doi.org/10.18653/v1/W17-4730.

Huck, M., Hangya, V., Fraser, A., 2019. Better OOV translation with bilingual terminology mining. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Florence, Italy, pp. 5809–5815. http://dx.doi.org/10.18653/v1/P19-1581, URL: https://aclanthology.org/P19-1581.

Hutchins, W., 1995. Machine translation: a brief history. In: Koerner, E., Asher, R. (Eds.), Concise History of the Language Sciences: From the Sumerians to the Cognitivists. Pergamon Press, Oxford, pp. 431–445.

Hutchins, J.W., 2017. Iso 639-2 language code list - codes for the representation of names of languages. https://www.loc.gov/standards/iso639-2/php/code_list.php, [Online; accessed 27-February-2023].

Institute of Electrical and Electronics Engineers (IEEE), 2022. Ieee xplore - advanced search. https://ieeexplore.ieee.org/search/advanced. [Online; accessed 27-February-2023].

Intento Inc., 2022. The State of Machine Translation 2022. Technical Report, Berkeley, CA, USA.

International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH), 2022. Meddra - medical dictionary for regulatory activities. https://www.meddra.org/, [Online; accessed 27-February-2023].

Jiang, D., Cheng, X., Han, T., 2022. Adaptive chinese pinyin IME for most similar representation. IEEE Access 10, 119533–119545.

Jimeno Yepes, A., Névéol, A., Neves, M., Verspoor, K., Bojar, O., Boyer, A., Grozea, C., Haddow, B., Kittner, M., Lichtblau, Y., Pecina, P., Roller, R., Rosa, R., Siu, A., Thomas, P., Trescher, S., 2017. Findings of the WMT 2017 biomedical translation shared task. In: Proceedings of the Second Conference on Machine Translation. Association for Computational Linguistics, Copenhagen, Denmark, pp. 234–247. http://dx.doi.org/10.18653/v1/W17-4719, URL: https://aclanthology.org/W17-4719.

Johnson, M., Schuster, M., Le, Q., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., Macduff, H., Dean, J., 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. Trans. Assoc. Comput. Linguist. 5, 339–351.

Joo, H., Burns, M., Lakshmanan, S.S.K., Hu, Y., Vydiswaran, V.G.V., 2021. Neural machine translation–based automated current procedural terminology classification system using procedure text: Development and validation study. JMIR Form. Res. 5, e22461.

Kaji, H., 1988. An efficient execution method for rule-based machine translation. In: Coling Budapest 1988 Volume 2: International Conference on Computational Linguistics. pp. 824–829. http://dx.doi.org/10.3115/991719.991803.

Kaliyadan, F., Gopinathan Pillai, S., 2010. The use of Google language tools as an interpretation aid in cross-cultural doctor_patient interaction: a pilot study. J. Innov. Health Inform. 18, 141–143.

Kapoor, R., Corrales, G., Flores, M.P., Feng, L., Cata, J.P., 2022. Use of neural machine translation software for patients with limited english proficiency to assess postoperative pain and nausea. JAMA Netw. Open 5, e221485.
Kasperė, R., Horbačauskienė, J., Motiejūnienė, J., Liubinienė, V., Patašienė, I., Patašius, M., 2021. Towards sustainable use of machine translation: Usability and perceived quality from the end-user perspective. Sustainability 13, 13430.
Khoong, E.C., Rodriguez, J.A., 2022. A research agenda for using machine translation in clinical medicine. J. Gen. Intern. Med. 37, 1275–1277.
Khoong, E.C., Steinbrook, E., Brown, C., Fernandez, A., 2019. Assessing the use of google translate for spanish and chinese translations of emergency department discharge instructions. JAMA Intern. Med. 179, 580–582.
Kirchhoff, K., Turner, A.M., 2016. Unsupervised resolution of acronyms and abbreviations in nursing notes using document-level context models. In: Proceedings of the Seventh International Workshop on Health Text Mining and Information Analysis. Association for Computational Linguistics, Auxtin, TX, pp. 52–60. http://dx.doi.org/10.18653/v1/W16-6107, URL: https://aclanthology.org/W16-6107.
Kitchenham, B.A., Charters, S., 2007. Guidelines for Performing Systematic Literature Reviews in Software Engineering. Technical Report EBSE 2007-001, Keele University and Durham University Joint Report, URL: https://www.elsevier.com/_data/promis_misc/525444systematicreviewsguide.pdf.
Kocijan, K., Kurolt, S., Mijić, L., 2020. Building croatian medical dictionary from medical corpus. Rasprave Inst. Hrvatski Jezik Jezikoslovlje 46, 765–782.
Köhler, S., Carmody, L., Vasilevsky, N., Jacobsen, J.O.B., Danis, D., Gourdine, J.-P., Gargano, M., Harris, N.L., Matentzoglu, N., McMurry, J.A., Osumi-Sutherland, D., Cipriani, V., Balhoff, J.P., Conlin, T., Blau, H., Baynam, G., Palmer, R., Gratian, D., Dawkins, H., Segal, M., Jansen, A.C., Muaz, A., Chang, W.H., Bergerson, J., Laulederkind, S.J.F., Yüksel, Z., Beltran, S., Freeman, A.F., Sergouniotis, P.I., Durkin, D., Storm, A.L., Hanauer, M., Brudno, M., Bello, S.M., Sincan, M., Rageth, K., Wheeler, M.T., Oegema, R., Lourghi, H., Rocca, M.G.D., Thompson, R., Castellanos, F., Priest, J., Cunningham-Rundles, C., Hegde, A., Lovering, R.C., Hajek, C., Olry, A., Notarangelo, L., Similuk, M., Zhang, X.A., Gómez-Andrés, D., Lochmüller, H., Dollfus, H., Rosenzweig, S., Marwaha, S., Rath, A., Sullivan, K., Smith, C., Milner, J.D., Leroux, D., Boerkoel, C.F., Klion, A., Carter, M.C., Groza, T., Smedley, D., Haendel, M.A., Mungall, C., Robinson, P.N., 2018. Expansion of the human phenotype ontology (HPO) knowledge base and resources. Nucleic Acids Res. 47, D1018–D1027.
Kumar, N., Mrinalini, K., Vijayalakshmi, P., 2018. Improving the Performance of Low-Resource Smt using Neural-Inspired Sentence Generator. IEEE, pp. 1–4. http://dx.doi.org/10.1109/ICCCSP.2018.8452859, URL: https://ieeexplore.ieee.org/document/8452859/.
Lakew, S.M., Federico, M., Negri, M., Turchi, M., 2018. Multilingual neural machine translation for low-resource languages. Ital. J. Comput. Linguist. 4, 11–25.
Lankford, S., Afli, H., Ní Loinsigh, Ó., Way, A., 2022. gaHealth: An English–Irish bilingual corpus of health data. In: Proceedings of the Thirteenth Language Resources and Evaluation Conference. European Language Resources Association, Marseille, France, pp. 6753–6758, URL: https://aclanthology.org/2022.lrec-1.727.
Lankford, S., Afli, H., Way, A., 2021. Machine translation in the covid domain: an english-irish case study for loresMT 2021. In: Proceedings of the 4th Workshop on Technologies for MT of Low Resource Languages (LoResMT2021). Association for Machine Translation in the Americas, pp. 144–150, Virtual.
Lee, S.H., 2018. Natural language generation for electronic health records. NPJ Digit. Med. 1.
Lee, W., Khoong, E.C., Zeng, B., Rios-Fetchko, F., Ma, Y., Liu, K., Fernandez, A., 2023. Evaluation of commercially available machine interpretation applications for simple clinical communication. J. Gen. Intern. Med..
Leite, F.O., Cochat, C., Salgado, H., da Costa, M.P., Queirós, M., Campos, O., Carvalho, P., 2016. Using Google Translateˆ© in the hospital: A case report. Technol. Health Care 24, 965–968.
Leong, P., 2017. Keynote speech: FPGA-based machine learning for prognostics and system health management. In: 2017 Prognostics and System Health Management Conference (PHM-Harbin). IEEE, pp. 885–895. http://dx.doi.org/10.1109/phm.2017.8079103.
Lester, C.A., Ding, Y., Li, J., Jiang, Y., Rowell, B., Vydiswaran, V.V., 2021. Human versus machine editing of electronic prescription directions. J. Amer. Pharm. Assoc. 61, 484–491.e1.
Li, J., Lester, C., Zhao, X., Ding, Y., Jiang, Y., Vydiswaran, V., 2020. Pharmmt: A Neural Machine Translation Approach To Simplify Prescription Directions. Association for Computational Linguistics, pp. 2785–2796. http://dx.doi.org/10.18653/v1/2020.findings-emnlp.251, URL: https://www.aclweb.org/anthology/2020.findings-emnlp.251.
Liang, Yingping, Han, Weifeng, 2022. Source text pre-editing versus target text post-editing in using google translate to provide health services to culturally and linguistically diverse clients. In: Science, Engineering and Health Studies (SEHS). pp. 1–5.
Library of Congress, 2017. The john w. hutchins machine translation archive. https://mt-archive.net/. [Online; accessed 27-February-2023].
Lin, Y.-C., Christen, V., Groß, A., Kirsten, T., Cardoso, S., Pruski, C., Silveira, M.D., Rahm, E., 2020. Evaluating cross-lingual semantic annotation for medical forms. In: Proceedings of the 13th International Joint Conference on Biomedical Engineering Systems and Technologies. SCITEPRESS - Science and Technology Publications, pp. 145–155. http://dx.doi.org/10.5220/0008979901450155.
Liu, W., Cai, S., 2015. Translating Electronic Health Record Notes from English to Spanish: A Preliminary Study. Association for Computational Linguistics, pp. 134–140. http://dx.doi.org/10.18653/v1/W15-3816, URL: http://aclweb.org/anthology/W15-3816.
Liu, B., Huang, L., 2021. ParaMed: a parallel corpus for english–Chinese translation in the biomedical domain. BMC Med. Inform. Decis. Mak. 21.
Liu, H., Liang, Y., Wang, L., Feng, X., Guan, R., 2020. Bionmt: A biomedical neural machine translation system. Int. J. Comput. Commun. Control 15, 1–13.
Lommel, A., 2018. Metrics for translation quality assessment: A case for standardising error typologies. pp. 109–127. http://dx.doi.org/10.1007/978-3-319-91241-7{_}6, URL: http://link.springer.com/10.1007/978-3-319-91241-7{_}6.
Lommel, A., Uszkoreit, H., Burchardt, A., 2014. Multidimensional quality metrics (MQM): A framework for declaring and describing translation quality metrics. Tradumàtica: Tecnol. Trad. 12, 455.
Lopez, A., 2008. Statistical machine translation. ACM Comput. Surv. 40, 1–49.
Luger, S., Anto-Ocrah, M., Allahsera, T., Homan, C.M., Zampieri, M., Leventhal, M., 2020. Health Care Misinformation: An Artificial Intelligence Challenge for Low-Resource Languages, Vol. 2884. pp. 1–6.
Luqman, H., Mahmoud, S.A., 2018. Automatic translation of arabic text-to-arabic sign language. Univers. Access Inf. Soc. 18, 939–951.
Luqman, H., Mahmoud, S.A., 2019. A machine translation system from arabic sign language to arabic. Univers. Access Inf. Soc. 19, 891–904.
Ma, H., Shen, L., Sun, H., Xu, Z., Hou, L., Wu, S., Fang, A., Li, J., Qian, Q., 2021. COVID term: a bilingual terminology for COVID-19. BMC Med. Inform. Decis. Mak. 21.
Ma, H., Yang, F., Ren, J., Li, N., Dai, M., Wang, X., Fang, A., Li, J., Qian, Q., He, J., 2020. EccParaCorp: a cross-lingual parallel corpus towards cancer education, dissemination and application. BMC Med. Inform. Decis. Mak. 20.
Manchanda, S., Grunin, G., 2020. Domain informed neural machine translation: Developing translation services for healthcare enterprise. pp. 255–262, URL: https://aclanthology.org/2020.eamt-1.27.pdf.
Manome, N., Shinohara, S., Suzuki, K., Mitsuyoshi, S., 2020. Machine translation from Japanese to robot language for human-friendly communication. In: International Conference on Human-Computer Interaction. pp. 254–260. http://dx.doi.org/10.1007/978-3-030-50726-8_33, URL: http://link.springer.com/10.1007/978-3-030-50726-8_33.
Manzini, E., Garrido-Aguirre, J., Fonollosa, J., Perera-Lluna, A., 2022. Mapping layperson medical terminology into the human phenotype ontology using neural machine translation models. Expert Syst. Appl. 204, 117446.
Marais, L., Louw, J.A., Badenhorst, J., Calteaux, K., Wilken, I., van Niekerk, N., Stein, G., 2020. Awezamed: A Multilingual, Multimodal Speech-To-Speech Translation Application for Maternal Health Care. IEEE, pp. 1–8. http://dx.doi.org/10.23919/FUSION45008.2020.9190240.
Marie, B., Fujita, A., Rubino, R., 2021. Scientific credibility of machine translation research: A meta-evaluation of 769 papers. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics, pp. 7297–7306. http://dx.doi.org/10.18653/v1/2021.acl-long.566.

Martinez-Costa, C., Schulz, S., 2017. HL7 FHIR: Ontological reinterpretation of medication resources. Stud. Health Technol. Inform. 235, 451–455.
Mauser, A., Hasan, S., Ney, H., 2008. Automatic evaluation measures for statistical machine translation system optimization. In: Proceedings of the 6th International Conference on Language Resources and Evaluation, LREC 2008. pp. 3089–3092.
Mehandru, N., Robertson, S., Salehi, N., 2022. Reliable and safe use of machine translation in medical settings. In: 2022 ACM Conference on Fairness, Accountability, and Transparency. ACM, New York, NY, USA, pp. 2016–2025. http://dx.doi.org/10.1145/3531146.3533244, URL: https://dl.acm.org/doi/10.1145/3531146.3533244.
Melero Nogués, M., 2018. El futur de les llengües en lera digital: Oportunitats i bretxa lingüística. Rev. Lleng. Dret 152–165.
Meng, Z., Okhmatovskaia, A., Polleri, M., Shen, Y., Powell, G., Fu, Z., Ganser, I., Zhang, M., King, N.B., Buckeridge, D., Collier, N., 2022. BioCaster in 2021: automatic disease outbreaks detection from global news media. Bioinformatics 38, 4446–4448.
Miller, J.M., Harvey, E.M., Bedrick, S., Mohan, P., Calhoun, E., 2018. Simple patient care instructions translate best: Safety guidelines for physician use of google translate. J. Clin. Outcomes Manage. 25.
Muhaxov, H., Lou, Z., Tayila, S., Yedemucao, D., 2016. Multiple-language translation system focusing on long-distance medical and outpatient services. pp. 471–475. http://dx.doi.org/10.1109/BigMM.2016.55.
Mujjiga, S., Krishna, V., Chakravarthi, K., J, V., 2019. Identifying semantics in clinical reports using neural machine translation. In: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33. pp. 9552–9557.
Musleh, A., Durrani, N., Temnikova, I., Nakov, P., Vogel, S., Alsaad, O., 2018. Enabling medical translation for low-resource languages. In: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). In: LNCS, vol. 9624, pp. 3–16.
Mutal, J., Gerlach, J., Bouillon, P., Spechbach, H., 2020. Ellipsis translation for a medical speech to speech translation system. pp. 281–290, URL: https://aclanthology.org/2020.eamt-1.30.pdf.
Mutinda, F.W., Yada, S., Wakamiya, S., Aramaki, E., 2021. Semantic textual similarity in japanese clinical domain texts using BERT. Methods Inf. Med. 60, e56–e64.
National Library of Medicine (NLM), 2022a. Mesh: Medical subject headings thesaurus. https://www.nlm.nih.gov/mesh/meshhome.html. [Online; accessed 27-February-2023].
National Library of Medicine (NLM), 2022b. Unified medical language system (umls). https://www.nlm.nih.gov/research/umls/index.html. [Online; accessed 27-February-2023].
National Library of Medicine (NLM) - National Center for Biotechnology Information (NCBI), 2022. Pubmed - advanced search. https://pubmed.ncbi.nlm.nih.gov/advanced/. [Online; accessed 27-February-2023].
Navigli, R., Ponzetto, S.P., 2012. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artificial Intelligence 193, 217–250.
Névéol, A., Jimeno Yepes, A., Neves, M., Verspoor, K., 2018. Parallel corpora for the biomedical domain. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). European Language Resources Association (ELRA), Miyazaki, Japan, URL: https://aclanthology.org/L18-1043.
Neves, M., Jimeno Yepes, A., Névéol, A., Grozea, C., Siu, A., Kittner, M., Verspoor, K., 2018. Findings of the WMT 2018 biomedical translation shared task: Evaluation on medline test sets. In: Proceedings of the Third Conference on Machine Translation: Shared Task Papers. Association for Computational Linguistics, Belgium, Brussels, pp. 324–339. http://dx.doi.org/10.18653/v1/W18-6403, URL: https://aclanthology.org/W18-6403.
Neves, M., Jimeno Yepes, A., Siu, A., Roller, R., Thomas, P., Vicente Navarro, M., Yeganova, L., Wiemann, D., Di Nunzio, G.M., Vezzani, F., Gerardin, C., Bawden, R., Estrada, D.J., Lima-lopez, S., Farre-maduel, E., Krallinger, M., Grozea, C., Neveol, A., 2022. Findings of the WMT 2022 biomedical translation shared task: Monolingual clinical case reports. In: Proceedings of the Seventh Conference on Machine Translation (WMT). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid), pp. 694–723, URL: https://aclanthology.org/2022.wmt-1.69.
Nunzio, G.D., Nosilia, G., Cambedda, V., 2021. A study on automatic machine translation tools: A comparative error analysis between deepl and yandex for Russian–Italian medical translation. Umanistica Digit. 10, 139–163.
Nurminen, M., Koponen, M., 2020. Machine translation and fair access to information. Fair MT 9, 150–169.
Ochieng, J., Buwembo, W., Munabi, I., Ibingira, C., Kiryowa, H., Nzarubara, G., Mwaka, E., 2015. Informed consent in clinical practice: patients' experiences and perspectives following surgery. BMC Res. Not. 8.
Oprea, A., Turk, A., Nita-Rotaru, C., Krieger, O., 2016. MOSAIC: A platform for monitoring and security analytics in public clouds. In: 2016 IEEE Cybersecurity Development (SecDev). IEEE, http://dx.doi.org/10.1109/secdev.2016.025.
Page, M.J., McKenzie, J.E., Bossuyt, P.M., Boutron, I., Hoffmann, T.C., Mulrow, C.D., Shamseer, L., Tetzlaff, J.M., Akl, E.A., Brennan, S.E., Chou, R., Glanville, J., Grimshaw, J.M., Hróbjartsson, A., Lalu, M.M., Li, T., Loder, E.W., Mayo-Wilson, E., McDonald, S., McGuinness, L.A., Stewart, L.A., Thomas, J., Tricco, A.C., Welch, V.A., Whiting, P., Moher, D., 2021a. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ n71.
Page, M.J., Moher, D., Bossuyt, P.M., Boutron, I., Hoffmann, T.C., Mulrow, C.D., Shamseer, L., Tetzlaff, J.M., Akl, E.A., Brennan, S.E., Chou, R., Glanville, J., Grimshaw, J.M., Hróbjartsson, A., Lalu, M.M., Li, T., Loder, E.W., Mayo-Wilson, E., McDonald, S., McGuinness, L.A., Stewart, L.A., Thomas, J., Tricco, A.C., Welch, V.A., Whiting, P., McKenzie, J.E., 2021b. PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. BMJ n160.
Palkova, K., Semaka, S., 2016. Consent to treatment and anamnesis as problem of communication with minor patients in healthcare decision-making. Eur. J. Interdiscip. Stud. 2, 57.
Papineni, K., Roukos, S., Ward, T., Zhu, W.-J., 2001. BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02. Association for Computational Linguistics, Morristown, NJ, USA, pp. 311–318. http://dx.doi.org/10.3115/1073083.1073135, URL: http://portal.acm.org/citation.cfm?doid=1073083.1073135.
Park, C.J., Yi, P.H., Yousif, H.A., Wang, K.C., 2022. Machine vs. radiologist-based translations of RadLex: Implications for multi-language report interoperability. J. Digit. Imaging 35, 660–665.
Parsifal, 2017. Parsifal - perform systematic literature reviews, [Computer software]. Available from https://parsif.al. Version 1.0 - Online; accessed 27-February-2023.
Petrigna, L., Musumeci, G., 2022. The metaverse: A new challenge for the healthcare system: A scoping review. J. Funct. Morphol. Kinesiol. 7, 63.
Popović, M., 2016. Chrf deconstructed: beta parameters and n-gram weights. In: Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 499–504. http://dx.doi.org/10.18653/v1/W16-2341, URL: http://aclweb.org/anthology/W16-2341.
Post, M., 2018. A call for clarity in reporting BLEU scores. In: Proceedings of the Third Conference on Machine Translation: Research Papers. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 186–191. http://dx.doi.org/10.18653/v1/W18-6319, URL: http://aclweb.org/anthology/W18-6319.
Prieto Ramos, F., 2015. Quality assurance in legal translation: Evaluating process, competence and product in the pursuit of adequacy. Int. J. Semiot. Law - Rev. Int. Sémiot. Jurid. 28, 11–30.
Qin, Y., Liang, Y., 2017. Medical data machine translation evaluation based on dependency n-grams. In: Smart Health. Springer International Publishing, pp. 174–181. http://dx.doi.org/10.5220/0008979901450155.
Radziszewski, A., 2013. A tiered CRF tagger for polish. In: Studies in Computational Intelligence, Vol. 467. pp. 215–230. http://dx.doi.org/10.1007/978-3-642-35647-6{_}16.

Rahmani, A., 2017. Adapting google translate for english-persian cross-lingual information retrieval in medical domain. In: 2017 Artificial Intelligence and Signal Processing Conference (AISP). IEEE, pp. 43–46. http://dx.doi.org/10.1109/aisp.2017.8324104.

Rajasekar, M., Udhayakumar, A., 2020. POS tagging using naïe bayes algorithm for tamil. Int. J. Sci. Technol. Res. 9, 574–578.

Rani, G.J.J., Gladis, D., Mammen, J.J., 2019. Regional language support for patient-inclusive decision making in breast cancer pathology domain. Int. J. Recent Technol. Eng. 8, 8392–8399.

Ranta, A., 2011. Grammatical Framework: Programming with Multilingual Grammars. Stanford, CSLI Publications, ISBN-10: 1-57586-626-9 (Paper), 1-57586-627-7 (Cloth).

Renato, A., Castaño, J., Ávila, P., Berinsky, H., Gambarte, L., Park, H., Pérez, D., Otero, C., Luna, D., 2018. A Machine Translation Approach for Medical Terms, Vol. 5. SCITEPRESS - Science and Technology Publications, pp. 369–378. http://dx.doi.org/10.5220/0006555003690378.

Roussis, D., Papavassiliou, V., Sofianopoulos, S., Prokopidis, P., Piperidis, S., 2022. Constructing parallel corpora from COVID-19 news using MediSys metadata. In: 2022 Language Resources and Evaluation Conference, LREC 2022. pp. 1068–1072.

Sadoughi, N., Finley, G.P., Edwards, E., Robinson, A., Korenevsky, M., Brenndoerfer, M., Axtmann, N., Miller, M., Suendermann-Oeft, D., 2018. Detecting section boundaries in medical dictations: Toward real-time conversion of medical dictations to clinical reports. In: Speech and Computer. Springer International Publishing, pp. 563–573. http://dx.doi.org/10.1007/978-3-319-99579-3_58.

San, M.E., Thu, Y.K., Supnithi, T., Usanavasin, S., 2022. Improving neural machine translation for low-resource english-myanmar-thai language pairs with SwitchOut data augmentation algorithm. In: 2022 17th International Joint Symposium on Artificial Intelligence and Natural Language Processing (ISAI-NLP). IEEE, http://dx.doi.org/10.5220/0008979901450155.

Schäfer, H., Idrissi-Yaghir, A., Horn, P., Friedrich, C., 2022. Cross-language transfer of high-quality annotations: Combining neural machine translation with cross-linguistic span alignment to apply NER to clinical texts in a low-resource language. In: Proceedings of the 4th Clinical Natural Language Processing Workshop. Association for Computational Linguistics, http://dx.doi.org/10.5220/0008979901450155.

Schulz, S., Boeker, M., Prunotto, A., 2022. Validation of multiple path translation for SNOMED CT localisation. In: Studies in Health Technology and Informatics. IOS Press, http://dx.doi.org/10.5220/0008979901450155.

Seligman, M., Dillinger, M., 2015. Evaluation and revision of a speech translation system for healthcare. In: Proceedings of the International Workshop on Spoken Language Translation (IWSLT). pp. 209–216, URL: http://workshop2015.iwslt.org/downloads/IWSLT{_}2015{_}RP{_}4.pdf{%}0Ahttp://www.mt-archive.info/15/IWSLT-2015-seligman.pdf.

Semmar, N., Laib, M., 2018. Integrating specialized bilingual lexicons of multiword expressions for domain adaptation in statistical machine translation. In: Communications in Computer and Information Science. Springer, Singapore, pp. 101–114. http://dx.doi.org/10.1007/978-981-10-8438-6_9.

Sen, S., Hasanuzzaman, M., Ekbal, A., Bhattacharyya, P., Way, A., 2020. Neural machine translation of low-resource languages using SMT phrase pair injection. Nat. Lang. Eng. 27, 271–292.

Shi, Y., Shi, C., Zhou, Z., 2019. Error types of machine translation of popular science text. In: Proceedings of the 2019 International Conference on Artificial Intelligence and Advanced Manufacturing - AIAM 2019. ACM Press, New York, New York, USA, pp. 1–4. http://dx.doi.org/10.1145/3358331.3358366, URL: http://dl.acm.org/citation.cfm?doid=3358331.3358366.

Shin, S., Matson, E.T., Park, J., Yang, B., Lee, J., Jung, J.-W., 2015. Speech-to-speech translation humanoid robot in doctor's office. pp. 484–489. http://dx.doi.org/10.1109/ICARA.2015.7081196.

Siklósi, B., Novák, A., Prószéky, G., 2016. Context-aware correction of spelling errors in hungarian medical documents. Comput. Speech Lang. 35, 219–233.

Skianis, K., Briand, Y., Desgrippes, F., 2020. Evaluation of Machine Translation Methods Applied To Medical Terminologies. Association for Computational Linguistics, pp. 59–69. http://dx.doi.org/10.18653/v1/2020.louhi-1.7, URL: https://www.aclweb.org/anthology/2020.louhi-1.7.

Smaili, K., Langlois, D., Pribil, P., 2022. Language rehabilitation of people with BROCA aphasia using deep neural machine translation. In: Proceedings of the 5th International Conference on Computational Linguistics in Bulgaria (CLIB 2022). Department of Computational Linguistics, IBL – BAS, Sofia, Bulgaria, pp. 162–170, URL: https://aclanthology.org/2022.clib-1.19.

SNOMED International, 2022. Why snomed ct? https://www.snomed.org/snomed-ct/why-snomed-ct. [Online; accessed 27-February-2023].

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., Makhoul, J., 2006. A study of translation edit rate with targeted human annotation. In: AMTA 2006 - Proceedings of the 7th Conference of the Association for Machine Translation of the Americas: Visions for the Future of Machine Translation. pp. 223–231.

Soares, F., Rebechi, R., Stevenson, M., 2020. Scibabel: a system for crowd-sourced validation of automatic translations of scientific texts. Genom. Inform. 18, e21.

Soto, X., Perez-De-Viñaspre, O., Labaka, G., Oronoz, M., 2019. Neural machine translation of clinical texts between long distance languages. J. Amer. Med. Inform. Assoc. 26, 1478–1487.

Soto, X., Perez-De-Viñaspre, O., Labaka, G., Oronoz, M., 2022. Comparing and Combining Tagging with Different Decoding Algorithms for Back-Translation in Nmt: Learnings from a Low Resource Scenario. European Association for Machine Translation, pp. 31–40, URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85137731795&partnerID=40&md5=2222fe10cf6cf605c774dbfea7df13b6.

Spechbach, H., Gerlach, J., Karker, S.M., Tsourakis, N., Combescure, C., Bouillon, P., 2019. A speech-enabled fixed-phrase translator for emergency settings: Crossover study. JMIR Med. Inform. 7, e13167.

Sutskever, I., Vinyals, O., Le, Q.V., 2014. Sequence to sequence learning with neural networks. Adv. Neural Inf. Process. Syst. 4, 3104–3112.

Taira, B.R., Kreger, V., Orue, A., Diamond, L.C., 2021. A pragmatic assessment of google translate for emergency department instructions. J. Gen. Intern. Med. 36, 3361–3365.

Takakusagi, Y., Oike, T., Shirai, K., Sato, H., Kano, K., Shima, S., Tsuchida, K., Mizoguchi, N., Serizawa, I., Yoshida, D., Kamada, T., Katoh, H., 2021. Validation of the reliability of machine translation for a medical article from japanese to english using deepl translator. Cureus.

Tan, Z., Wang, S., Yang, Z., Chen, G., Huang, X., Sun, M., Liu, Y., 2020. Neural machine translation: A review of methods, resources, and tools. AI Open 1, 5–21.

Tavosanis, M., 2019. Human evaluation of google translator and deepl for translations of journalistic texts from english into italian. In: 6th Italian Conference on Computational Linguistics, CLiC-it 2019, Vol. 2481. CEUR Workshop Proceedings, Bari, Italy, p. 7, URL: https://ceur-ws.org/Vol-2481/paper70.pdf.

Taylor, R.M., Crichton, N., Moult, B., Gibson, F., 2015. A prospective observational study of machine translation software to overcome the challenge of including ethnic diversity in healthcare research. Nurs. Open 2, 14–23.

Technavio, 2022. Machine translation market by application and geography - forecast and analysis 2022–2026. https://www.technavio.com/report/machine-translation-market-industry-analysis. [Online; accessed 27-February-2023].

The Migration Observatory - University of Oxford, 2019. English languaage use and proficiency of migrants in the uk. https://migrationobservatory.ox.ac.uk/resources/briefings/english-language-use-and-proficiency-of-migrants-in-the-uk/#kp3. [Online; accessed 27-February-2023].

Torres-Hostench, O., 2020. Translator training outdoors. Transl. Spaces 9, 224–254.

Trujillos-Yébenes, L., Muñoz-Miquel, A., 2022. La traducción automática y la posedición en el ámbito médico, tradumàtica: tecnologies de la traducció.

Turner, A.M., Choi, Y.K., Dew, K., Tsai, M.-T., Bosold, A.L., Wu, S., Smith, D., Meischke, H., 2019. Evaluating the usefulness of translation technologies for emergency response communication: A scenario-based study. JMIR Public Health Surv. 5, e11171.

Turner, A.M., Dew, K.N., Desai, L., Martin, N., Kirchhoff, K., 2015. Machine translation of public health materials from english to chinese: A feasibility study. JMIR Public Health Surveill. 1, e17.

van den Bercken, L., Sips, R.-J., Lofi, C., 2019. Evaluating neural text simplification in the medical domain. In: The World Wide Web Conference, WWW '19. Association for Computing Machinery, New York, NY, USA, pp. 3286–3292. http://dx.doi.org/10.5220/0008979901450155.

Van Der Wees, M., Bisazza, A., Monz, C., 2019. Evaluation of machine translation performance across multiple genres and languages. In: LREC 2018-11th International Conference on Language Resources and Evaluation. pp. 3822–3827.

Vardaro, J., Schaeffer, M., Hansen-Schirra, S., 2019. Translation quality and error recognition in professional neural machine translation post-editing. Informatics 6, 41.

Vasant, D., Chanas, L., Malone, J., Hanauer, M., Olry, A., Jupp, S., Robinson, P., Parkinson, H., Rath, A., 2014. Ordo: An ontology connecting rare disease, epidemiology and genetic data. In: Proceedings of ISMB'14. pp. 1–4.

Veríssimo, V., Silva, C., Hanael, V., Moraes, C., Costa, R., Maritan, T., Aschoff, M., Gaudêncio, T., 2019. A study on the use of sequence-to-sequence neural networks for automatic translation of brazilian portuguese to LIBRAS. In: Proceedings of the 25th Brazillian Symposium on Multimedia and the Web. ACM, http://dx.doi.org/10.1145/3323503.3360292.

Vieira, L.N., O'Hagan, M., O'Sullivan, C., 2021. Understanding the societal impacts of machine translation: a critical review of the literature on medical and legal use cases. Inf. Commun. Soc. 24, 1515–1532.

Villegas, M., Intxaurrondo, A., Gonzalez-Agirre, A., Marimon, M., Krallinger, M., 2018. The MeSpEN resource for english-spanish medical machine translation and terminologies: census of parallel corpora, glossaries and term translations. In: Malero, M., Krallinger, M., Gonzalez-Agirre, A. (Eds.), LREC MultilingualBIO: Multilingual Biomedical Text Processing.

Wang, Z., Poon, J., Poon, S., 2019. TCM translator: A sequence generation approach for prescribing herbal medicines. In: 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, http://dx.doi.org/10.5220/0008979901450155.

Way, A., Haque, R., Xie, G., Gaspari, F., Popović, M., Poncelas, A., 2020. Rapid development of competitive translation engines for access to multilingual covid-19 information. Informatics 7.

Weng, W.-H., Chung, Y.-A., Szolovits, P., 2019. Unsupervised clinical language translation. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '19. Association for Computing Machinery, New York, NY, USA, pp. 3121–3131. http://dx.doi.org/10.1145/3292500.3330710.

Wiesmann, E., 2019. Machine translation in the field of law: A study of the translation of Italian legal texts into German. Comp. Legilinguistics 37, 117–153.

Wikimedia Foundation, 2022. Wiktionary, the free dictionary. https://www.wiktionary.org/. [Online; accessed 27-February-2023].

Wołk, K., 2021. Real-time sentiment analysis for polish dialog systems using MT as pivot. Electronics 10, 1813.

Wołk, K., Glinkowski, W., Żukowska, A., 2018. Enhancing the assessment of (polish) translation in PROMIS using statistical, semantic, and neural network metrics. In: Advances in Intelligent Systems and Computing. Springer International Publishing, pp. 351–366. http://dx.doi.org/10.5220/0008979901450155.

Wołk, K., Marasek, K., 2015a. Neural-based machine translation for medical text domain, based on european medicines agency leaflet texts. Procedia Comput. Sci. 64, 2–9.

Wołk, K., Marasek, K., 2015b. Polish-english statistical machine translation of medical texts. Adv. Intell. Syst. Comput. 314, 169–179.

Wołk, K., Marasek, K., Glinkowski, W., 2015. Telemedicine as a special case of machine translation. Comput. Med. Imaging Graph. 46, 249–256.

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, Ł., 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., Dean, J., 2016. Google's neural machine translation system: Bridging the gap between human and machine translation.

Xie, W., Ji, M., Zhao, M., Qian, X., Chow, C.-Y., Lam, K.-Y., Hao, T., 2021b. Supporting risk-aware use of online translation tools in delivering mental healthcare services among spanish-speaking populations. Comput. Intell. Neurosci. 2021, 1–13.

Xie, W., Ji, M., Zhao, M., Zhou, T., Yang, F., Qian, X., Chow, C.-Y., Lam, K.-Y., Hao, T., 2021a. Detecting symptom errors in neural machine translation of patient health information on depressive disorders: Developing interpretable bayesian machine learning classifiers. Front. Psychiatry 12.

Xu, C., Forkel, W., Borgwardt, S., Baader, F., Zhou, B., 2019. Automatic Translation of Clinical Trial Eligibility Criteria Into Formal Queries, Volume 2518. CEUR-WS, URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85077702402&partnerID=40&md5=c870a831278211e8e2db4c7f4b361d8f.

Xu, H., Wang, S., 2022. ProTranslator: Zero-shot protein function prediction using textual description. In: Lecture Notes in Computer Science. Springer International Publishing, pp. 279–294. http://dx.doi.org/10.1007/978-3-031-04749-7_17.

Yamashita, N., Ishida, T., 2006. Effects of machine translation on collaborative work. In: Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW. ACM, Banff, Alberta, Canada, pp. 515–524. http://dx.doi.org/10.1145/1180875.1180955.

Yang, T., Chen, H., Xue, X., Zhao, S., 2021. Low-resource machine translation based on fusion drop method. In: 2021 2nd International Conference on Artificial Intelligence and Computer Engineering (ICAICE). IEEE, http://dx.doi.org/10.5220/0008979901450155.

Yeganova, L., Wiemann, D., Neves, M., Vezzani, F., Siu, A., Jauregi Unanue, I., Oronoz, M., Mah, N., Névéol, A., Martinez, D., Bawden, R., Di Nunzio, G.M., Roller, R., Thomas, P., Grozea, C., Perez-de Viñaspre, O., Vicente Navarro, M., Jimeno Yepes, A., 2021. Findings of the WMT 2021 biomedical translation shared task: Summaries of animal experiments as new test set. In: Proceedings of the Sixth Conference on Machine Translation. Association for Computational Linguistics, Online, pp. 664–683, URL: https://aclanthology.org/2021.wmt-1.70.

Yu, X., Shen, Y., Ni, Y., Huang, X., Wang, X., Chen, Q., Tang, B., 2021. CapsTM: capsule network for chinese medical text matching. BMC Med. Inform. Decis. Mak. 21.

Yu, P., Zhu, Y., 2021. Model and verification of medical english machine translation based on optimized generalized likelihood ratio algorithm. J. Sensors 2021.

Zhang, X., Guo, Y., 2022. OmiTrans: Generative adversarial networks based omics-to-omics translation framework. In: 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, http://dx.doi.org/10.1109/bibm55620.2022.9995537.

Zhang, J., Nie, Y., Chang, J., Zhang, J.J., 2021. Surgical instruction generation with transformers. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. Springer International Publishing, pp. 290–299. http://dx.doi.org/10.1007/978-3-030-87202-1_28.

Zhao, L., Gao, W., Fang, J., 2021. High-performance english–chinese machine translation based on GPU-enabled deep neural networks with domain corpus. Appl. Sci. 11, 10915.

Zhu, Y., Sha, Y., Wu, H., Li, M., Hoffman, R.A., Wang, M.D., 2022. Proposing causal sequence of death by neural machine translation in public health informatics. IEEE J. Biomed. Health Inf. 26, 1422–1431.

Ziganshina, L.E., Yudina, E.V., Gabdrakhmanov, A.I., Ried, J., 2021. Assessing human post-editing efforts to compare the performance of three machine translation engines for english to russian translation of cochrane plain language health information: Results of a randomised comparison. Informatics 8.