KSU-JRLT

Journal of Research in Language & Translation
Special Issue (December, 2024)

جامعة الملك سعود
King Saud University

Constraints to Neural Machine Translation Quality, Human and Automated Evaluation, and Quality Improvement across Language Pairs: A Systematic Literature Review

Najia Abdulkareem AlGhamedi

King Saud University – College of Language Sciences
Riyadh, Saudi Arabia
nalghamedi@ksu.edu.sa

https://orcid.org/0009-0001-0453-9519

Received: 21/03/2024; Revised: 05/10/2024; Accepted: 06/10/2024

https://doi.org/10.33948/JRLT-KSU-S-1-4

الملخص

حتى الآن لا توجد مراجعة منهجية للأدبيات (SLR) لاستعراض ما توصلت له الأبحاث والدراسات حول جودة الترجمة الآلية العصبية .(NMT)

تتمثل أهداف هذه المراجعة المنهجية للأدبيات في استعراض مشاكل جودة الترجمة الالية والتعرف على نقاط القوة ونقاط الضعف وأوجه قصور الترجمة الآلية بالإضافة إلى التعرف على أداء تقييم جودة الترجمة الآلية بواسطة الإنسان وتلك التي تعتمد على الآلة، إلى جانب التعرف على المنهجيات التي يمكن استخدامها لتحسين جودة الترجمة الآلية العصبية. ولتحقيق هذه الأهداف اعتمدت الدراسة على منهجي (PRISMA) و (SALSA) لإجراء المراجعة للأدبيات في هذا الموضوع. واشتملت الأدبيات على المقالات الأكاديمية المحكّمة التي نشرت باللغة الإنجليزية في الفترة بين عامي 2018 و2024. واستخدمت في الدراسة المكتبة الرقمية السعودية وشبكة العلوم وشبكة سكوبس للبحث عن هذه المقالات. وتوصل البحث إلى 51 مقالة أكاديمية تغطي 89 زوجا لغويا والتي تحقق معايير البحث .

واستخلص البحث إلى أن من المعوقات الرئيسية التي تحد من جودة الترجمة الآلية (NMT) التنوع الصرفي لأزواج اللغات وجودة المدونات اللغوية وكمية النصوص التي تم جمعها، وهي تحديات تخص اللغات والمجالات ذات الموارد المنخفضة.

تُعتبر مصفوفة BLEU الأكثر انتشارًا في تقييم الترجمة، حيث حققت أعلى نتائجها في اللغات ذات الموارد الوفيرة والتنوع الصرفي الكبير، مثل الإنجليزية والعربية. أما في أزواج اللغات ذات الموارد الغنية والتشابه الصرفي، كاللغات الأوروبية وبعض اللغات الآسيوية مثل الصينية واليابانية والكورية، فقد سُجلت درجات BLEU متوسطة.

وقد اقترحت الدراسات أساليب تقييم جديدة تهدف إلى معالجة تحديات المواءمة بين المدونات اللغوية والتنوع الصرفي. وعلى الرغم من التقدم الملحوظ في أداء الترجمة الآلية العصبية (NMT) واقترابها من الأداء البشري على المستوى اللفظي،

إلا أن التقييم البشري كشف عن قصور في جوانب أخرى كالكفاية والطلاقة. وعليه يمكن القول إن الترجمة الآلية العصبية لم تصل بعد إلى مستوى الترجمة البشرية، مما يستدعي تحويل التركيز نحو أبعاد لغوية أخرى كالكفاية والطلاقة واللباقة والوعي بالسياق.

KSU-JRLT

Journal of Research in Language & Translation
Special Issue (December, 2024)

جامعة الملك سعود
King Saud University

Constraints to Neural Machine Translation Quality, Human and Automated Evaluation, and Quality Improvement across Language Pairs: A Systematic Literature Review

Najia Abdulkareem AlGhamedi

King Saud University – College of Language Sciences
Riyadh, Saudi Arabia
nalghamedi@ksu.edu.sa

https://orcid.org/0009-0001-0453-9519

Received: 21/03/2024; Revised: 05/10/2024; Accepted: 06/10/2024

https://doi.org/10.33948/JRLT-KSU-S-1-4

Abstract

There is no systematic literature review (SLR) that has attempted to synthesize current knowledge on Neural Machine Translation (NMT) quality. The objectives of this SLR are to investigate constraints to NMT quality; examine strengths, limitations, and performance of automated and human evaluation metrics; and identify approaches that can be used to improve NMT quality. The PRISMA and SALSA methodologies were adopted to carry out this SLR. Peer-reviewed articles published in English between 2018 and 2024 were searched on the Saudi Digital Library, Web of Science, and Scopus. Furthermore, references of included articles were searched. There were 51 articles spanning 89 language pairs that met the inclusion criteria and were included in this SLR. The major constraints to NMT quality are the morphological diversity of language pairs and low corpora quality and quantity, which are challenges specific to low-resource languages and domains. BLEU is the dominant automated metric, and it is highest in high-resource morphologically diverse languages such as English and Arabic. Moderate BLEU scores were observed in high resource morphologically similar pairs such as European languages and some Asian languages such as Chinese, Japanese, and Korean. Innovative approaches aimed at bridging corpora and morphological diversity have been proposed. Therefore, significant progress has been made in bridging human and NMT performance at the lexical dimension. However, human evaluation showed NMT performance was unsatisfactory in other dimensions, such as adequacy and fluency. NMT has not yet matched human translation, and the focus needs to shift to other language dimensions such as adequacy, fluency, politeness, and context awareness.

Keywords: *automated/machine translation; human translation, human evaluation, neural machine translation; quality translation evaluation*

Introduction

Machine translation (MT) has not yet matched human translation (HT). The significant challenges that have led to this situation are correctly resolving ambiguity in a source text, adequately providing meaning in the targeted language, and gender bias. The diversity of structure of words in source and target languages has made it difficult for MT systems to achieve human-level translation (Popel et al., 2020). Previous approaches to MT relied on rules or statistical machine translation (SMT), which could not yield satisfactory translation quality. Hand-made rules faced the difficulty of covering all language complexities. SMT faced the difficulty of "modeling long-distance dependencies between words" (Tan et al., 2020, p. 5). Deep learning neural networks, which have revolutionized other fields in artificial intelligence, have replaced rule-based and SMT methods resulting in neural machine translation (NMT) as the established approach in MT. These NMT models can access complete information anywhere in a sentence. It is this elimination of independence that has significantly improved translation quality and narrowed the gap between human and machine translation (Hassan et al., 2018; Wu et al., 2016).

In the modern globalized world, language barriers can challenge human interaction. Occasionally the demand for translation services surpasses available human translation capacity. MT tools are becoming popular as they can bridge this gap (Rivera-Trigueros, 2022). Several studies have reported the beneficial use of MT. Muftah (2022) compared human translations to Google Translate and Babylon Translate systems and found no difference. That study concluded a symbiotic relationship needs to exist between machines and MT. Lihua (2022) argues although HT and MT are similar, MT lacks the "faithfulness, expressiveness, and elegance" (p. 2) present in human translation. For minimal-requirement translations such as daily tourism and business translation, MT is adequate, but it cannot substitute for human translation. Hassan et al. (2018) found Microsoft translation system quality of news from Chinese to English was at par with professional human translation and was better than the quality of non-professional translations that were crowd-sourced. Zouhar et al. (2021) have reported two observations from English to Czech professional translators. First, better MT systems resulted in fewer sentence changes, but the relationship between system quality and the time required to edit MT output was unclear. Second, BLEU was not a stable system quality metric.

Although millions use MT daily, there are people who still doubt the value of MT in enhancing the productivity of human translators. A significant contributor to this situation is the absence of a unified quality standard, meaning quality is context- and time-specific (Way, 2018). Several studies have contributed to this argument by reporting the limitations of NMT. Vardaro et al. (2019) report major problematic NMT error categories are omissions and mistranslations. Hasibuan (2020) notes that when considering semantic meaning, the output of MT significantly differs from the truthful meaning to the extent that the translation can be regarded as a general translation. Yang et al. (2023) found that in the translation of news from English to Chinese, MT faced three challenges. MT fails to understand cultural and semantic details in the source language and provide a coherent translation.

Assessing the translation quality of MT is very challenging due to two factors. First, there is no universally accepted definition of a correct translation. Second translation quality is evaluated by comparing MT output to a human translation. The problem arises because human translations are never identical, although they convey the same meaning. Therefore, MT output can have a high match percentage to one human translation while having a low match percentage to another (Ulitkin et al., 2021). A few literature reviews have been carried out on MT translation quality. Rivera-Trigueros (2022) found while most studies either used human or automated evaluation, less than one-quarter of studies used human and automated evaluations. Chatzikoumi (2019) presents various "automated, semi-automated, and human metrics" (n.p.) for quality evaluation. Lee et al. (2023) present key contributions and limitations of automated evaluation metrics but exclude human evaluation methods and do not use a systematic literature review (SLR) methodology. Han (2018) surveys various manual and automated methods. Automated methods are categorized into lexical and syntactic, while human methods are divided into four categories. No SLR on NMT quality evaluation could be found. It is this gap in the literature that motivated this SLR.

The broad objective of this study is to exhaustively review the current literature on NMT quality. The specific research questions that will be investigated are:

i. What factors limit the quality of current NMT systems?
ii. How do automated and human NMT evaluations differ across language pairs?
iii. What are the limitations of current automated and human NMT quality evaluation metrics?
iv. What performance-enhancing measures can be used to improve NMT quality evaluation?

Definition of Abbreviations

BLEU – *Bilingual Evaluation Understudy*
NIST – *National Institute of Standards and Technology*
WER – *Word Error Rate*
TER – *Translation Error Rate*
GTM – *General Text Matcher*
METEOR – *Metric for Evaluation of Translation with Explicit Ordering*
CHRF – *Character n-gram F-Score*
BEER – *Better Evaluation as Ranking*
RUSE – *Regressor Using Sentence Embeddings*
NUBIA – *Neural Based Interchangeability Assessor*
COMET – *Cross-lingual Optimized Metric for Evaluation of Translation*
ESIM – *Enhanced Sequential Inference Model*
YiSi – '*Meaning'*
MQM – *Multi-dimensional Quality Metrics*
HTER – *Human-targeted Translation Error Rate*
DQF – *Dynamic Quality Framework*
MSA – *Modern Standard Arabic*
CNN – *Convolutional neural networks*
RNN – *Recurrent Neural Networks*
BRNN – Bidirectional Recurrent Neural Networks

Literature Review

MT Quality Evaluation

Developing an MT system is distinct from establishing the quality of the MT output. MT quality can be assessed using automated or human evaluation. Automated evaluation is the dominant approach, as human evaluation is usually “slow, expensive, and inconsistent” (Way, 2018). The critical elements in human evaluation are adequacy and fluency. Adequacy is concerned with assessing the correct transmission of information and requires comparing the original and translated text. Adequacy is concerned with examining syntactic quality and does not require comparing original and translation. Human evaluation can assess other elements such as acceptability, comprehension, and legibility (Castilho et al., 2018). Human evaluation uses Likert scales, error identification, and categorization (Chatzikoumi, 2019).

Various taxonomies have been proposed to assess the quality of MT output. Flanagan (1994) proposed a framework consisting of 21 major and minor errors observed from the output of English-French translation and advocated the need to develop bespoke categories for each language pair, as some error categories are only meaningful for specific language pairs. Vilar et al. (2006) proposed a five-category taxonomy observed from Chinese to English, Spanish to English, and English to Spanish pairs for classifying MT errors. These errors are missing words, word order, incorrect words, unknown words, and punctuation. Farrús et al. (2009) proposed a five-error scheme for SMT systems for bidirectional Spanish to Catalan translation. These error types are morphological, lexical, orthographic, syntactic, and semantic. Frederico et al. (2014) proposed a seven-category error taxonomy observed from English to Arabic and Chinese to Russian. The error categories are morphological, lexical choice, addition, omission, casing and punctuation, reordering, and too many errors. Kirchhoff et al. (2014) proposed a twelve-category error taxonomy observed from English to Spanish translations. These errors are missing words, extra words, word order, morphology, word sense errors, punctuation, spelling, capitalization, untranslated, pragmatics, diacritics, and others.

Popovic (2018) notes that within the last decade, projects aimed at standardizing and reducing inconsistencies in error typologies have emerged. Lommel (2018) identifies MQM and DQF frameworks. The MQM council (2024) has proposed a seven-category translation typology. These broad error categories are terminology, accuracy, linguistic conventions, style, locale conventions, audience appropriateness, and design and markup. Most of these error categories were proposed when SMT was the dominant approach. The DQF framework assesses quality using quantitative measures and qualitative categories of errors (Panic, 2020).

Compared to human evaluation, automated evaluation is cost-effective and can easily be compared across systems, but it does not provide quality comparable to human assessment. These metrics compare a reference against a hypothesis. Available NMT automated metrics can be categorized into lexical, which compares lexical characteristics such as words or phrases; embedding, which compares similarity in “embedding of language models,” and supervised metrics derived from a machine or deep learning model (Lee et al., 2023). Lexical metrics can be further categorized into word and character-based metrics. Word-based metrics include BLEU, NIST, WER, TER, GTM, and METEOR (Papineni et al., 2002; Doddington, 2002; Woodard & Nelson, 1982; Snover et al., 2006; Turian et al., 2003); Banerjee & Lavie, 2005).

BLEU is highly popular as it has demonstrated a decent correlation with human assessment (Castilho et al., 2018). CHRF is a character-based score (Popović & Arčan, 2015). Available embedding metrics are MEANT, YiSi, BERT, and BART (Lo & Wu, 2011; Lo, 2019; Zhang et al., 2020; Yuan et al., 2021). Supervised metrics are BEER, BLEND, RUSE, BERT for MTE, BLEURT, NUBIA, and COMET (Hirao et al., 2020; Ma et al., 2017; Rei et al., 2020; Stanojević & Sima'an, 2015).

The widespread adoption of MT in the translation profession has necessitated assessing post-editing efforts. The HTER metric combines TER and a human to estimate changes required to achieve a post-edited translation. A comparison between the translation and the post-edited version is made instead of a comparison to a reference (Maucec & Donaj, 2019). The AER metric quantifies the number of edit operations done by a translator. High HTER occurs together with low MT quality, but there is no correlation between AER and MT quality. This suggests MT quality is affected more by post-editing time than keyboard operations (Sanchez-Torron & Koehn, 2016).

Limitations of Human Evaluation and Automated Metrics

Various criticisms of automated metrics have been reported. Castilho et al. (2018) argue that automated metrics use a reference translation developed by humans, and the quality of these reference translations is not assessed, which can lead to variability. Han (2020) notes the lack of a universal correct translation limits the evaluation of "syntactic and semantic equivalence." Lee et al. (2023) note lexical metrics capture lexical similarity while ignoring "semantic, grammatical diversity, and sentence structure." BLEU has been observed to have unsatisfactory performance on semantically similar sentences with a wide variety of vocabulary and structure and has a weak correlation with human evaluation (Macháček & Bojar, 2014; Ma et al., 2018). Translations with a high BLEU score have been observed to have poor quality or are unintelligible (Smith et al., 2016). BLEU has been observed to lack interpretability and indication of content quality (Hamon & Mostefa, 2008; Reiter & Belz, 2009). Although neural metrics have been observed to overcome some limitations of BLEU, there is a lack of clarity on the extent of bias of neural metrics as they lack explainability (Freitag et al., 2021). TER has been found to lead to conflicting conclusions when comparing human and system translations, and generally, TER, BLEU, CHRF, ESIM, and YiSi-1 metrics have similar biases such that erroneous decisions using one metric will also happen in the other metrics (Mathur et al., 2020).

BLEU fails to reflect sentence information, and NIST was developed to overcome this limitation. Additionally, BLEU does not recognize synonyms and stems as the same words. TER emphasizes word-level matching while ignoring semantic similarities in reference and translation. Furthermore, TER ignores translation fluency (Lee et al., 2023). WER fails to compute word transformation, and the TER metric has been proposed to overcome this limitation (Snover et al., 2006). COMET and BLEURT have been found to lack adequate sensitivity in detecting errors related to the "translation of numbers and entities" (Amrhein & Sennrich, 2022). This results in a lack of trustworthiness and difficulty interpreting COMET and BLEURT. These limitations are not associated with lexical metrics like BLEU (Glushkova et al., 2023).

Although human evaluation is considered to have better reliability than human evaluation, it has the limitations of requiring considerable time and human resources, and it lacks reproducibility. Additionally, human evaluation involves training and assessment of agreement among evaluators (Han, 2016). Manual evaluation is financially demanding and slow, yet quick feedback is required in MT development (Huang & Papineni, 2007). Subjectivity in manual evaluation can arise due to evaluator bias, lack of clarity in the scoring scale, and evaluator fluency in the language under consideration (Vilar et al., 2006). Often human evaluators have limited knowledge leading to low agreement between evaluators. Furthermore, guidelines provided to evaluators are not clearly defined, leading to varying interpretations (Vidal & Oliver, 2023).

Assessing the quality of an MT system poses challenges, as no single translation can be presumed correct. However, objectively evaluating the quality of an MT system and how it affects the work process of professional translators is achievable. In quality assessment, human and automated evaluation, as well as assessing the post-editing effort required, are necessary. Furthermore, error classification is essential to understand the inherent subjectivity in human evaluation (Popovic, 2018; Rivera-Trigueros, 2022).

Method

A SLR comprehensively searches, synthesizes, and summarizes literature from a specific field in a transparent and reproducible way. An SLR can be distinguished from other literature reviews that do not use a transparent, objective, and systematic approach in selecting studies (Kraus et al., 2020). However, even when carrying out an SLR, bias can creep in when study inclusion and exclusion are not clearly defined (Nightingale, 2009). The "Protocol and Reporting result with Search, Appraisal, Synthesis, and Analysis" (PSALSAR) framework provides a transparent and reproducible approach for carrying out SLR. The PSALSAR framework combines SALSA and PRISMA methodologies widely used for SLR. The PSALSAR framework clearly prescribes six critical characteristics of an SLR: research questions, objectives, reproducible method, search strings, study quality appraisal, and data synthesis and reporting (Mengist et al., 2019). The PSALSAR framework was considered appropriate in this study as it excludes some PRISMA elements only relevant to randomized controlled trials. The PSALSAR framework requires six steps which are discussed in subsequent sections.

Protocol

The "Population, Intervention, Comparison, Outcome, and Context" (PICOC), which is part of the PSALSAR framework, provides guidelines for identifying the research scope and research questions. Application of this framework to the current study is illustrated in Table 1

Table 1

PICOC Framework Elements

Concept	Application
Population	Scientific research on human and automated evaluation of NMT quality

Intervention	Use of NMT quality evaluation metrics
Comparison	Strengths and limitations of various NMT quality evaluation metrics
Outcome	Knowledge of NMT quality, errors in NMT, strengths and limitations of NMT quality metrics, and variations in NMT metrics across language pairs.
Context	Current knowledge on NMT quality assessment

Search

Table 2 shows the keywords identified in the population of interest and used to search the Saudi Digital Library, SCOPUS, and Web of Science databases. These search terms were used in the title, abstract, and keywords. Articles that do not include relevant terms in the title and abstract may exist, but such articles are outside the scope of this SLR.

Table 2

Search Keywords

Database	**Search string**	**Number of articles**	**Date Acquired**
Saudi Digital Library	Neural machine translation AND quality AND metric OR error OR automated OR human OR evaluation	53	3/4/2024
Web of Science	Neural machine translation AND quality AND metric OR error OR automated OR human OR evaluation	48	3/4/2024
SCOPUS	Neural machine translation AND quality AND metric OR error OR automated OR human OR evaluation	281	3/4/2024

Appraisal

The appraisal phase aims to identify relevant articles. The first stage uses inclusion/exclusion criteria to identify relevant articles. The second stage evaluates the quality of selected articles.

Selection of Studies

The inclusion and exclusion criteria used to select relevant articles are shown in Table 3. The objective of these criteria is to include only recently published, peer-reviewed articles written in English, focusing on NMT quality evaluation and excluding grey literature. These criteria are applied to search results and papers identified from references. The process of selection of relevant papers is illustrated in Figure 1

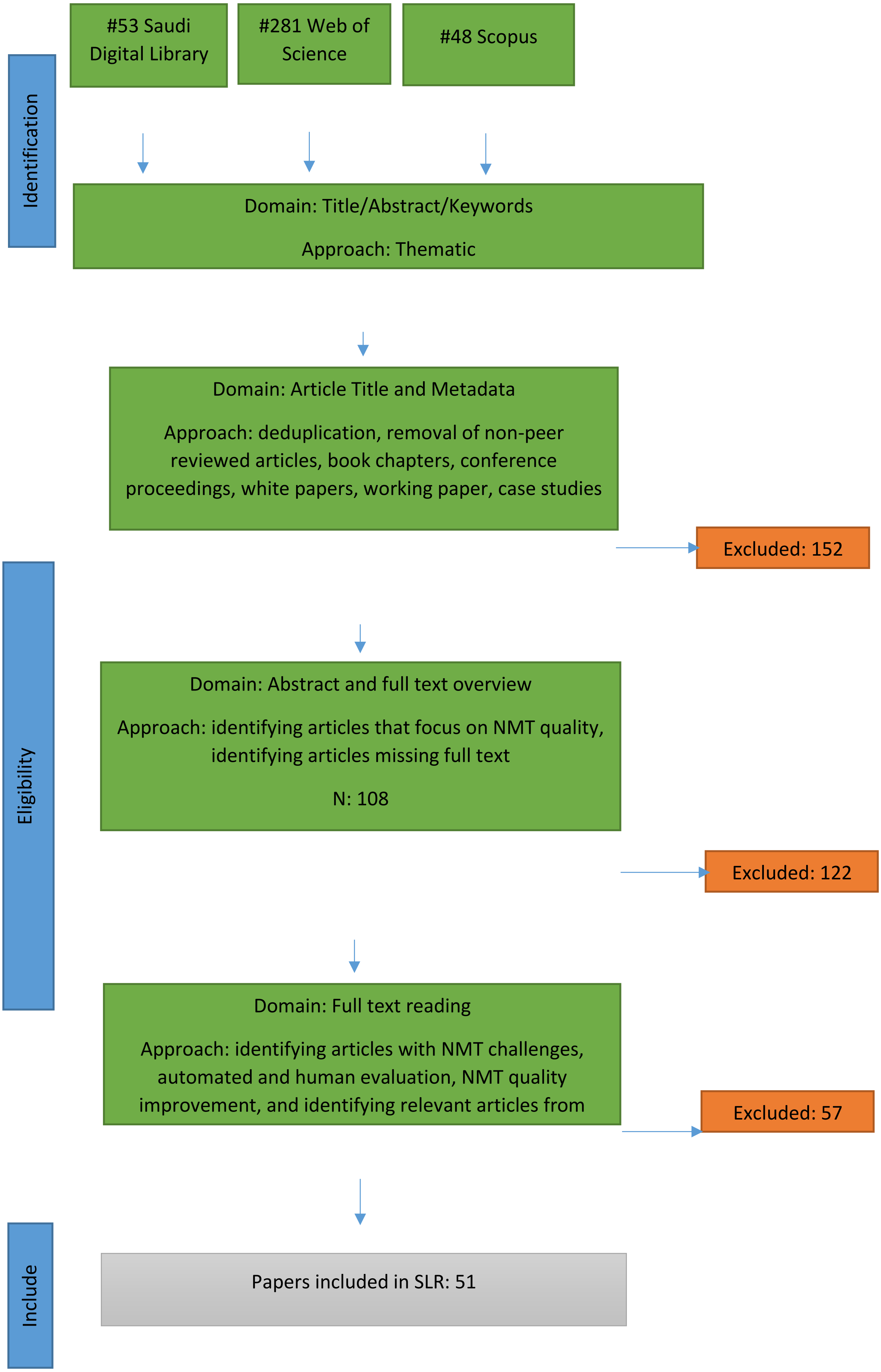
Table 3

Inclusion/Exclusion Criteria

Criteria	Decision
Search terms can be found in the abstract, title, or keywords	Include
The paper has been published in a reputable peer-reviewed journal	Include
The paper has been published in English	Include
Paper is original research or an SLR	Include
The paper has been cited in original research or SLR	Include
Paper is published before 2018	Exclude
The paper cannot be accessed or has been retracted	Exclude
The paper does not focus on NMT quality evaluation	Exclude
Grey literature such as white papers, working papers	Exclude

Figure 1

SLR Flowchart



Quality Assessment

The SLRs that met the inclusion criteria also had to meet the four other criteria listed below to be included in the SLRs.

i. The criteria used to include or exclude articles are clearly and adequately explained
ii. The search strategy is sufficient to provide all relevant articles
iii. The SLR is published in a reputable peer-reviewed journal
iv. The SLR adequately discusses NMT quality evaluation aspects

Synthesis

The synthesis step involves data extraction and categorization from articles that were considered relevant using the pre-determined inclusion/exclusion criteria.

Table 4

Extracted Data Items

Criteria	Justification
Publication year	To investigate the trend in the number of NMT-quality research papers
Journal name and publisher	To understand the distribution of NMT quality research across journals and publishers
Language pair	To understand dominant language pairs
Metrics	To understand the use of metrics in NMT quality
NMT quality constraints	To answer research question 1
Strengths and limitations of NMT quality evaluation metrics	To answer research question 2
Variation in NMT quality metrics across language pairs	To answer research question 3
NMT quality enhancements	To answer research question 4

Analysis

Tables and bar charts presented quantitative characteristics of studies, while thematic coding analyzed qualitative data.

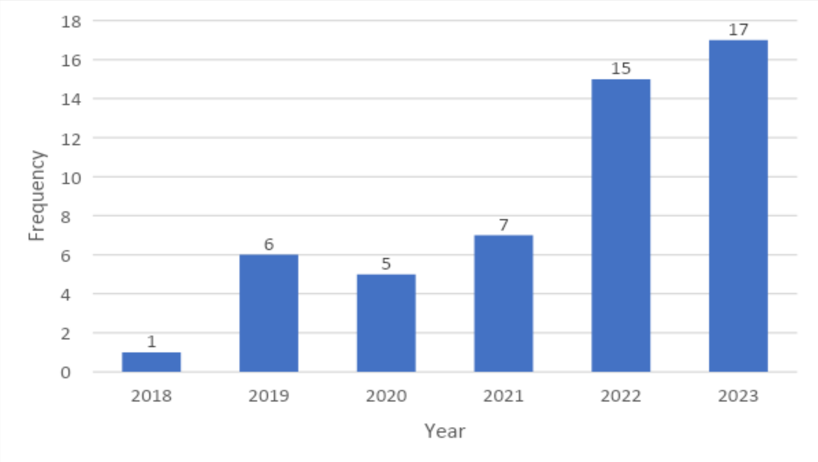
Results

Study Characteristics

The number of NMT-quality research papers has consistently grown from 2018 to 2023. Specifically, more studies were published in 2022 and 2023 than in the other years, suggesting interest in machine translation quality is increasing.

Figure 2

Number of Studies in Each Year



As shown in Table 5, there is comprehensive journal coverage. The 51 articles included in this SLR were published by 34 journals, and most contributed a single article.

Table 5

Number of Studies from Each Journal

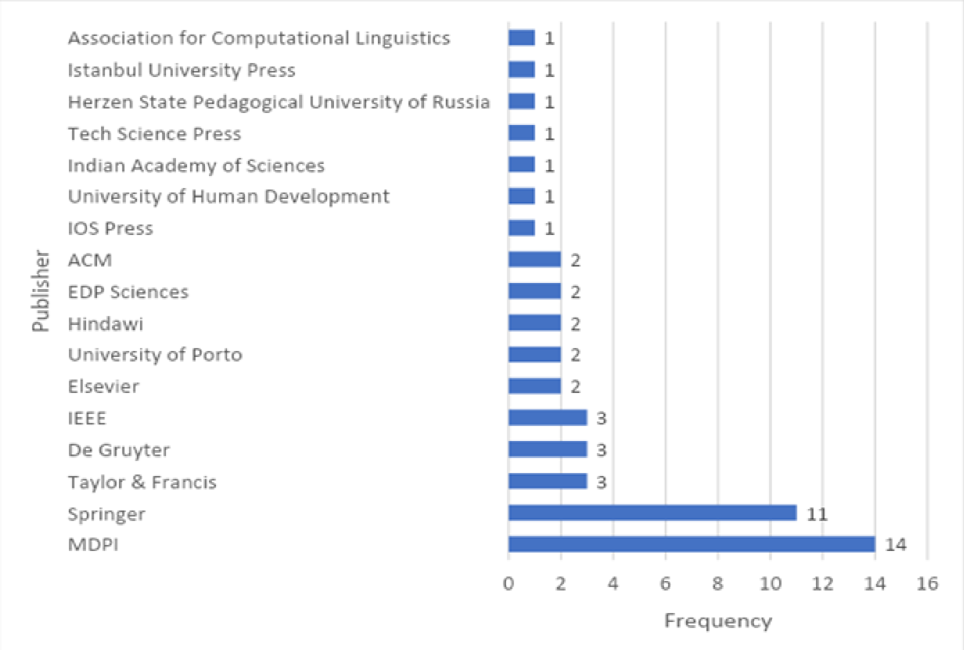
Journal	Frequency
Applied Sciences	4
Information	4
IEEE Access	3
Neural Processing Letters	3
Journal of Language and Law	2
Journal of Intelligent Systems	2
mathematics	2
Neural Computing and Applications	2
Electronics	2
International Journal of Information Technology	2
ACM Transactions Asian Low-Resource Languages	2
Mobile Information Systems	1
Machine Translation	1
PeerJ Computer Science	1

Arabian Journal for Science and Engineering	1
Computers, Materials, & Continua	1
Informatics	1
Applied Artificial Intelligence	1
Cogent Engineering	1
Sadhana	1
Complexity	1
MATEC Web of Conferences	1
UHD Journal of Science and Technology	1
MEDINFO	1
Journal of Applied Linguistics and Lexicography	1
E3S Web of Conferences	1
Computational Linguistics	1
Open Computer Science	1
Computer Science	1
Procesamiento del Lenguaje Natural, Revista	1
Journal of Social Studies	1
Future Internet	1
Machine Learning	1
Istanbul University Journal of Translation Studies	1

There were 17 publishers that contributed the 51 articles included in this SLR as illustrated in Figure 2. The dominant publishers were MDPI and Springer

Figure 2

Number of Studies from Each Publisher



The 51 articles included had 89 language pairs. English is the dominant source and target language, suggesting that most current NMT efforts translate other languages into English and English into different languages. Chinese is the second most important source language, while MSA and Chinese are the second most crucial target languages.

Figure 3

Common Source Languages

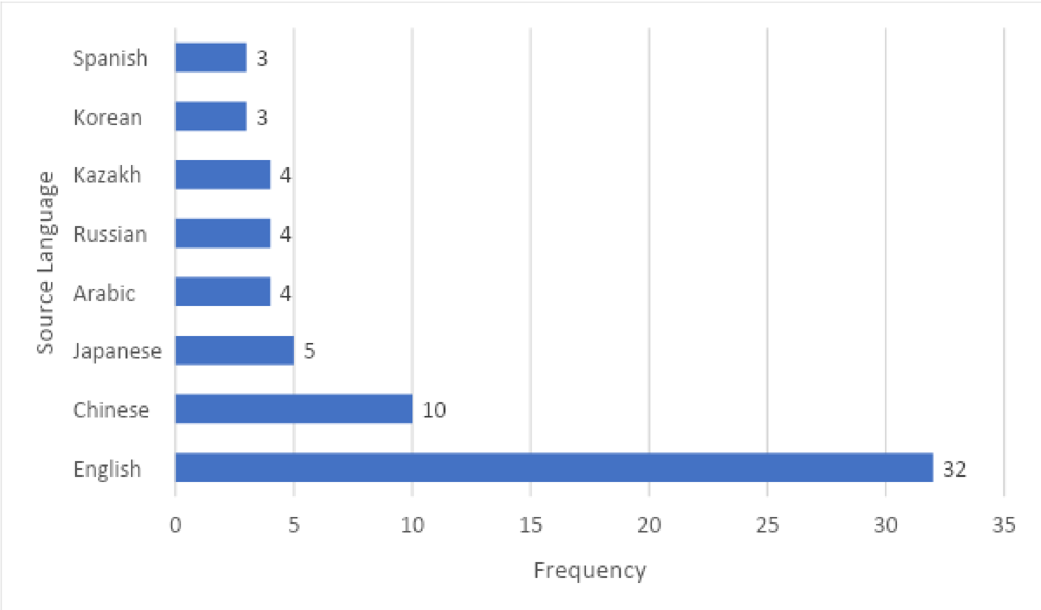
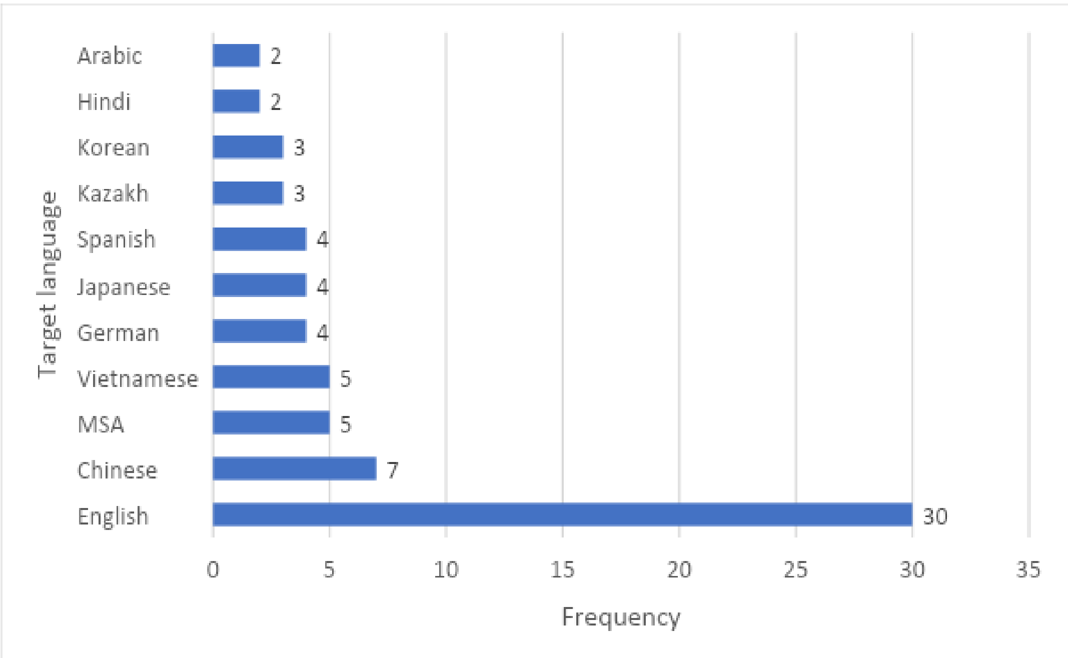


Figure 4

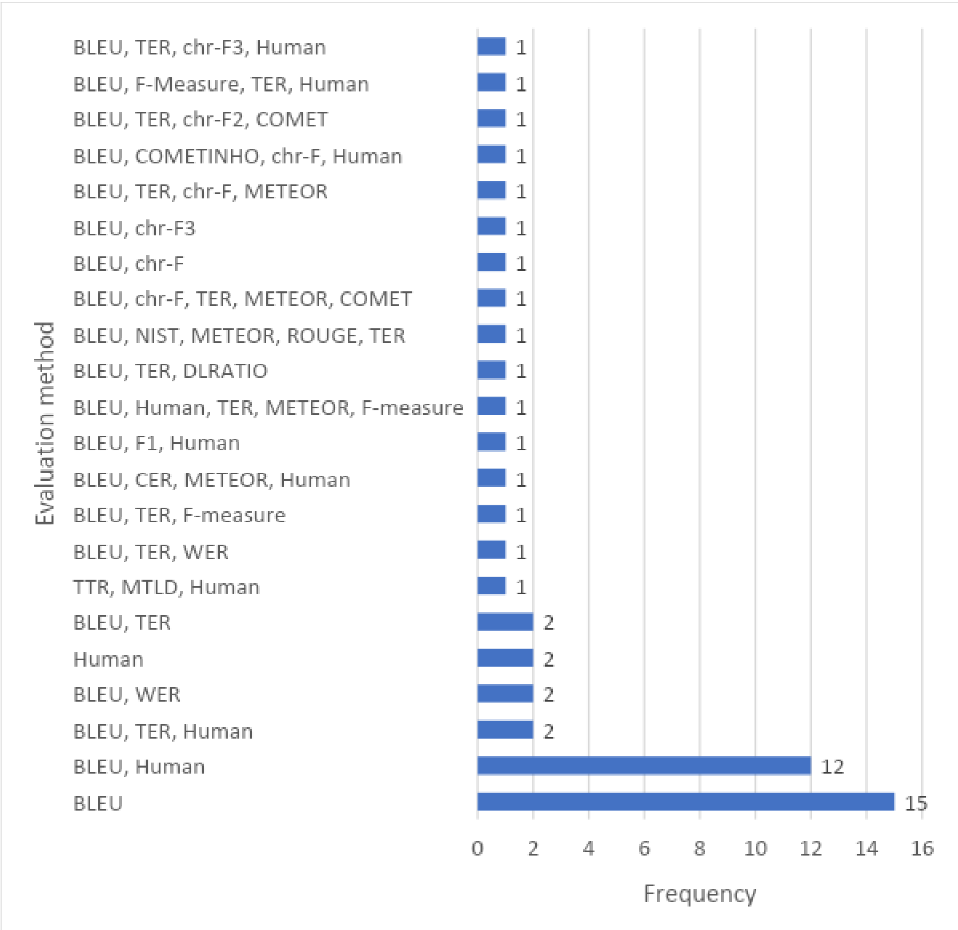
Common Target Languages



BLEU is the dominant automated evaluation metric in current NMT research. This acronym stands for (Bilingual Evaluation Understudy). All the studies except three studies used BLEU and another metric. Although NIST was developed to overcome some limitations of BLEU, it was used in only one study. Human evaluation is more frequently used to supplement automated metrics, and there were two studies that used human evaluation only.

Figure 5

Number of Studies Using Various Evaluation Metrics



Challenges to NMT Quality

The key challenges faced by NMT systems are highlighted below.

i. NMT translation quality is low on specific domains and low resource languages (Liu et al., 2023). Specifically, the quality is constrained by the low quality and quantity of available corpus. Constructing a large and high-quality corpus is complex and costly. This is the case especially for specific domains such as legal texts and low-resource languages such as Persian, Turkish, Nepali, and Sinhala (Ahmadnia & Dorr, 2019; Li et al., 2020; O'Shea et al., 2023; Pham et al., 2023; Tukeyev et al., 2019).

ii. Morphological diversity worsens the quality of NMT in low-resource situations such as translating to and from Kazakh, translating to and from Korean, as well as translating between Indian languages (Kumar et al., 2023; Nguyen et al., 2018;

Tukeyev et al., 2019). Languages that have a free-order grammatical structure, such as Arabic dialects, present a challenge to NMT (Baniata et al., 2022).

iii. Low vocabulary coverage between source and target languages leads to a high number of words missing in the NMT vocabulary. This is the case when translating between English and Arabic and translating to or from Korean (Berrichi & Mazroui, 2021; Nguyen et al., 2018).

iv. Although transfer learning has been observed to improve NMT quality in low-resource languages, this approach has limited success in logo-graphic languages like Japanese and Chinese (Ngo et al., 2022).

v. Unknown words and a large number of rare words in morphologically rich languages such as Arabic are a challenge as NMT has a fixed vocabulary (Aqlan et al., 2019; Wang, 2022).

vi. Translation of legal terms from Spanish to English is a challenge for general NMT systems as they lack contextual understanding of the translation objective (Vigier-Moreno & Macías, 2022). Document-level context ignored by NMT could significantly improve translation quality (Nayak et al., 2022). Due to a lack of contextual understanding, translation of literary texts such as novels lacks lexical richness and local context (Webster et al., 2020). Furthermore, NMT ignores essential aspects such as politeness (Uguet & Aranberri, 2023).

vii. NMT systems face challenges in translating specialized abbreviations, colloquialisms, and proper nouns such as names of people, geographical locations, and organizations. This is not challenging for a specialist human translator (Liu et al., 2023; Ulitkin et al., 2021; Xie et al., 2023). For example, in translating Arabic to English, NMT had challenges in translating Saba in its various forms. NMT translated the Sabaeans to 'Sabes' and the Sabaean era to 'Seventh Century' (Sismat, 2020).

viii. Long or short sentences are challenging to NMT resulting in mistranslation and over-translation (Berrichi & Mazroui, 2021; Wan et al., 2022).

ix. Current NMT models for translating natural text to sign language have low accuracy (Farooq et al., 2023).

x. Although automatic evaluation is the usual approach to NMT quality evaluation, they have been questioned as these metrics are just an approximation of quality (Alvarez-Vidal & Oliver, 2023).

Performance of Automated Metrics across Language Pairs

Comparison of BLEU

There are no clearly established guidelines for interpreting BLEU scores. Denkowski and Lavie (2010) suggest that BLEU scores higher than 0.3 indicate an understandable translation, and BLEU scores higher than 0.5 indicate that a translation is good and fluent. O'Shea et al. (2023) suggest BLEU scores higher than 50 indicate a translation requires minimal post-editing. Morphological similarity and resource availability are the key determinants of translation quality. Grouping of languages based on these two characteristics facilitates the interpretation of BLEU scores. Alimova (2021) notes that languages can be divided into four categories: “isolating, agglutinative, inflectional, and polysynthetic” (n.p.) languages. Classification of languages into high or low resources is not clearly established. Mirela (2024) defines 20 high-resource languages as languages many people speak and receive significant research and investment towards developing MT systems. English, Japanese, Arabic, and Spanish are high-resource languages. Greek, Urdu, French, and Dutch are medium resource languages. Norwegian, Telugu, Danish, and Pashto are low-resource languages (Zhang et al., 2022) BLEU scores of morphologically similar languages are shown in Table 6. The highest BLEU scores were obtained using NMT to translate between MSA and Arabic dialects in the general domain. These are Semitic languages. When using SMT to translate Tunisian to MSA, the BLEU score was notably lower than translating the other Arabic dialects to MSA. This suggests that NMT is superior to SMT when translating Arabic dialects to MSA.

The Indo-European languages had lower BLEU scores than Semitic languages. A comparison of Indo-European languages revealed the highest BLEU scores were obtained when translating to high-resource languages such as English and Spanish. Translation between Japanese and Korean, which are agglutinative languages, resulted in high BLEU scores comparable to those obtained when translating to English or Spanish. Furthermore, Japanese can be considered a high-resource language. This suggests morphological similarity and resource availability are essential to NMT quality. However, translating Russian and Hindi to English or Persian to Spanish resulted in notably lower BLEU scores.

Table 6

Morphologically Similar Languages

Language Pair	Domain/MT Type	BLEU Score	Study
Levantine-MSA	General - NMT	63.99	Baniata et al., 2022
Maghrebi-MSA	General - NMT	61.07	Baniata et al., 2022
Tunisian-MSA	General - NMT	60	KchaouSaméh et al., 2023
Iraqi-MSA	General - NMT	58.33	Baniata et al., 2022
English-Irish	General - NMT	52.7	Lankford et al., 2022
Gulf-MSA	General - NMT	47.21	Baniata et al., 2022
Nile-MSA	General - NMT	47.15	Baniata et al., 2022
Tunisian-MSA	General - SMT	32.25	KchaouSaméh et al., 2023
Slovenian-English	General - NMT	46.4	Dugonik et al., 2023
Kurdish-English	General - NMT	45	Badawi, 2023
Russian-English	Scientific - NMT	42.1	Ulitkin et al., 2021
English-Spanish	News - NMT	38.2	Alvarez-Vidal & Oliver, 2023
Spanish-English	General - NMT	36.19	Nayak et al., 2022
Spanish-English	General - NMT	35.71	Ahmadnia & Dorr, 2019
German-English	General - NMT	35.4	Xie et al., 2022

Castilian-Spanish	General - NMT	35.3	Uguet & Aranberri, 2023
English-Spanish	General - NMT	34.66	Ahmadnia & Dorr, 2019
Greek-English	General - NMT	32.59	O'Shea et al., 2023
German-English	General - NMT	32.01	Mahsuli et al., 2023
English-Slovenian	General - NMT	32	Dugonik et al., 2023

Table 6

Continued

Language Pair	Domain/MT Type	BLEU Score	Study
Hindi-English	General - NMT	31.78	Nayak et al., 2022
English-German	General - NMT	30.51	Wan et al., 2022
English-German	General - NMT	29.4	Xie et al., 2022
English-German	General - NMT	29.23	Yan, 2022
Persian-Spanish	General - NMT	30.12	Ahmadnia & Dorr, 2019
Spanish-Persian	General - NMT	28.02	Ahmadnia & Dorr, 2019
English-German	General - NMT	26.34	Peng et al., 2021
Hindi-English	General - NMT	22.39	Chauhan et al., 2022
English-Hindi	General - NMT	21.67	Chauhan et al., 2022
Russian-English	General - NMT	24.82	Shukshina, 2019
Japanese-Korean	General - NMT	34.22	Nguyen et al., 2018
Korean-Japanese	General - NMT	39.85	Nguyen et al., 2018

A comparison of BLEU scores among morphologically similar high-resource languages in Table 7 showed translation from Chinese to Japanese resulted in the highest score. These two languages are logographic, and NMT systems can take advantage of shared information resulting from similarity in sub-character units (Zhang & Komachi, 2018). However, Zhang et al. (2023) reported a very low BLEU score when translating Chinese to Japanese, but this score was significantly increased by improving corpus quality. This result emphasizes the importance of corpus quality, as similar results were obtained when translating Japanese to Chinese. When translating English to Chinese, BLEU scores were higher compared to translating Chinese to English. This can be explained by the use of varying corpus.

Table 7

Morphologically Dissimilar High Resource Languages

Language Pair	Domain/MT type	BLEU Score	Study
Chinese-Japanese	General - NMT	38.1	Zhang & Matsumoto, 2019
English-Chinese	Engineering - NMT	34.25	Liu et al., 2023

English-Chinese	General - NMT	34.1	Liu et al., 2023
English-Chinese	General - NMT	33.56	Yan, 2022
Japanese-English	Medical - NMT	27.3	Yagahara et al., 2024
English-Chinese	General - NMT	26.4	Xie et al., 2022
Chinese-English	General - NMT	24.9	Liu et al., 2023
Chinese-English	General - NMT	21.3	Xie et al., 2022
Chinese-English	General - NMT	19.49	Wan et al., 2022
Chinese-English	General - NMT	19.14	Peng et al., 2021
Chinese-English	General - NMT	15.6	Nayak et al., 2022
Chinese-Japanese	General - NMT	3.7-22.9	Zhang et al., 2023

BLEU scores higher than 30 were observed when translating Altaic languages (Kazakh, Turkish, Mongolian) to a high-resource language such as English or Chinese. This result suggests translating between these languages will result in an understandable translation. However, Tukeyev et al. (2019) reported a notably lower BLEU score when translating Kazakh to English. This result suggests there is uncertainty when using varying corpus. The high BLEU score obtained when translating Turkish to English in the cardiology domain is interesting. It compares favorably to the BLEU score obtained when translating in a general domain using NMT. Furthermore, NMT had a notably lower BLEU score than SMT in the cardiology domain. When translating the Bible from English to Mizo, which can be considered a domain-specific situation, NMT was not superior to SMT. Translation of English to Vietnamese resulted in a notably lower BLEU score in the legal domain compared to the general domain. These results suggest although NMT has become dominant, SMT can be useful in domain-specific situations where corpus availability is a challenge. However, SMT may be inferior to NMT in the general domain, as demonstrated by the lower BLEU score obtained when translating Turkish into English using SMT.

Table 8

Morphologically Dissimilar High/Low Resource Target/Source Languages

Language Pair	**Domain/MT Type**	**BLEU Score**	**Study**
Kazakh-English	General - NMT	45	Karyukin et al., 2023
Turkish-English	General - NMT	39	Dogru, 2022
Mongolian-Chinese	General - NMT	37.29	Qing-dao-er-ji et al., 2022
Turkish-English	Cardiology - SMT	36	Dogru, 2022
English-Vietnamese	General - NMT	28.3	Pham et al., 2023
Uyghur-Chinese	General - NMT	27.6	Pan et al., 2020
Turkish-English	General - NMT	25.95	Pan et al., 2020
Turkish-English	Cardiology - NMT	25	Dogru, 2022

Myanmar-Thai	General - NMT	24.92	Hlaing et al., 2022
English-Korean	General - NMT	23.49	Nguyen et al., 2018
English-Arabic	General - NMT	23.02	Aqlan et al., 2019
Thai-Myanmar	General - NMT	22.9	Hlaing et al., 2022
Turkish-English	General - SMT	22	Dogru, 2022
Korean-English	General - NMT	20.39	Nguyen et al., 2018
English-Vietnamese	Legal - NMT	19.83	Pham et al., 2023
Arabic-English	General - NMT	19.39	Aqlan et al., 2019
Arabic-English	General - NMT	18.77	Mahsuli et al., 2023
Korean-French	General - NMT	18.65	Nguyen et al., 2018
Chinese-Vietnamese	General - NMT	17.2	Ngo et al., 2022
Kazakh-English	General - NMT	16.4	Tukeyev et al., 2019
English-Mizo	Bible-NMT	15.82	Devi & Purkayastha, 2023
English-Mizo	Bible-SMT	15.82	Devi & Purkayastha, 2023
English-Kazakh	General – NMT	15.7	Tukeyev et al., 2019
Nyishi-English	General - NMT	15.43	Kakum et al., 2023
Russian-Kazakh	General – NMT	15.3	Tukeyev et al., 2019
Korean-Spanish	General - NMT	15.09	Nguyen et al., 2018

Table 8

Continued

Language Pair	**Domain/MT type**	**BLEU score**	**Study**
Kazakh-Russian	General – NMT	14.4	Tukeyev et al., 2019
Japanese-Vietnamese	General - NMT	14.1	Ngo et al., 2022
Spanish-Korean	General - NMT	13.44	Nguyen et al., 2018
French-Korean	General - NMT	12.94	Nguyen et al., 2018
English-Finnish	General - NMT	11.55	Peng et al., 2021
English-Nyishi	General - NMT	10.18	Kakum et al., 2023
Nepali-English	General - NMT	7.64	Li et al., 2020
Sinhala-English	General - NMT	6.68	Li et al., 2020
Russian-Vietnamese	General - NMT	13.84-14.84	Nguyen et al., 2021

Comparison of Other Metrics

Higher BLEU, NIST, and METEOR values indicate higher translation quality, while lower TER and WER metrics indicate higher quality (Cer et al., 2010). From Table 9 it can be observed language pairs such as English-Spanish, English-Irish, Spanish-English, Slovenian-English, and Japanese-Korean that have lower TER scores also had higher BLEU scores. The lower BLEU score observed in the translation of English-German and English-Slovenian corresponded to a higher TER score. However, higher BLEU scores do not always occur together with lower TER scores. The higher BLEU score observed in the translation of Russian-English did not correspond to a lower TER score. This finding suggests that BLEU and TER will often be consistent, but there could be exceptions. The higher METEOR scores observed in translation of Hindi-English, Slovenian-English, and Spanish-English correspond to higher BLEU scores. However, the low METEOR scores observed in the translation of Castilian-Spanish and German-English contrast with high BLEU scores. This finding suggests there could be inconsistencies between METEOR and BLEU. The high F-measures observed in the translation of Russian-English and English-Irish correspond to high BLEU scores. Lower F-measures observed in translating German-English and Hindi-English correspond to lower BLEU scores. However, the lower F-measure observed in the translation of Spanish-English is inconsistent with the higher BLEU score.

Table 9

Morphologically Similar Languages

Language Pair	TER	F-measure	NIST	WER	COMET	METEOR	Study
English-Spanish	46		7.98	0.49	0.47		Alvarez-Vidal & Oliver, 2023
Russian-English	54.43	72.6					Wan et al., 2022
English-German	54.17						Wan et al., 2022
English-German		53.08					Xie et al., 2022
German-English		63.34					Xie et al., 2022
Castilian-Spanish		56.7				0.19	Uguet & Aranberri, 2023
English-Irish	41	72					Lankford et al., 2022
Hindi-English	48.53	53.5				0.66	Nayak et al., 2022
Slovenian-English	40.1				83.3	0.705	Dugonik et al., 2023

English-Slovenian	54.4		80.7	0.553	Dugonik et al., 2023
Spanish-English	40.95	55.8		0.70	Nayak et al., 2022
German-English	72.68	51.11		0.10	Mahsuli et al., 2023
Korean-Japanese	45.43				Nguyen et al., 2018
Japanese-Korean	43.6				Nguyen et al., 2018

From Table 10, the translation of Kazakh-English had the lowest TER, which is consistent with the highest BLEU score among morphologically dissimilar languages. Translation between English and Nyishi had the highest TER scores, which is consistent with low BLEU scores. Translation of Chinese-English yielded conflicting results. Two studies reported TER scores of 65 and 67 (Nayak et al., 2022; Wan et al., 2022). However, Xi et al. (2022) reported a TER score of 48. This result suggests inconsistencies in BLEU scores where the same language pairs have high and low scores are also evident in TER. These results support earlier observations of inconsistency between BLEU and TER. However, the high TER scores observed in translation between Nyishi and English are consistent with low METEOR scores.

Table 10

Morphologically Dissimilar Languages

Language Pair	**TER**	**F-measure**	**CER**	**METEOR**	**WER**	**COMET**	**Study**
Chinese-English	65.71						Wan et al., 2022
Chinese-English	67.75	37.5		0.48			Nayak et al., 2022
English-Chinese	59.37						Wang, 2022
Chinese-Japanese	44.8						Zhang & Matsumoto, 2019
Korean-English	64.27						Nguyen et al., 2018
English-Korean	71.03						Nguyen et al., 2018

Korean-French	64.92						Nguyen et al., 2018
French-Korean	83.22						Nguyen et al., 2018
Korean-Spanish	69.86						Nguyen et al., 2018
Spanish-Korean	80.25						Nguyen et al., 2018
English-Chinese	42.52						Xie et al., 2022
Chinese-English	48.86						Xie et al., 2022

Table 10

Continued

Language Pair	**TER**	**F-measure**	**CER**	**METEOR**	**WER**	**COMET**	**Study**
English-Japanese			0.54	0.19			Yagahara et al., 2024
Thai-Myanmar		39.75					Hlaing et al., 2022
Kazakh-English	48				55		Karyukin et al., 2023
Myanmar-Thai		41.73					Hlaing et al., 2022
Turkish-English		48.6					Pan et al., 2020
Uyghur-Chinese		36.73					Pan et al., 2020
Arabic-English	72.68	34.55				-0.72	Mahsuli et al., 2023

Nyishi-English	83.4	42	0.19	Kakum et al., 2023
English-Nyishi	92.1	43	0.15	Kakum et al., 2023

Comparison between Automated Metrics and Human Evaluation

Human and automated metrics are compared in Table 11. Languages with a higher BLEU score also have a higher human rating score. Similarly, languages with a lower BLEU score also have a lower BLEU score. However, Liu et al. (2023) reported a low BLEU score and a high human rating score in translation of Chinese-English. These results suggest that BLEU and human rating scores are often consistent, but there could be exceptions. For studies that did not use rating scales, a comparison of human and automated evaluation is summarized below.

i. In translating English-Irish, human evaluation using the MQM framework identified three major error categories: omission, mistranslation, and grammar. Comparing evaluators revealed agreement in all error categories except mistranslation (Lankford et al., 2022).

ii. In translation between Russian-Kazakh and English-Kazakh, the human evaluation revealed the correct translation of the main parts, but the NMT system had challenges in translating pronouns and nouns (Tukeyev et al., 2019).

iii. In translation of Japanese to Chinese manual evaluation showed “relatively good” translation quality. (Zhang et al., 2023).

iv. In the translation of Turkish to English, SMT trained on cardiology domain corpus had a BLEU score of 36, while incorporating general domain corpus reduced SMT BLEU score to 22. NMT trained on cardiology domain corpus had a BLEU score of 25, and incorporating general domain corpus increased the BLEU score to 39. F-measure and TER also indicated that SMT in this particular domain was superior. However, a human evaluation indicated that NMT trained on general and domain corpus was superior (Dogru, 2022)

v. In the translation of property law from Greek to English, human evaluation provided mixed results. Human-translated text had higher accuracy errors, while post-edited texts had higher style errors (O’Shea et al., 2023).

vi. In the translation of Russian to Vietnamese, human evaluation revealed the general meaning was adequately translated. Still, there were problems with the translation of named entities and the accuracy of meanings (Nguyen et al., 2021).

vii. In translating Kurdish to English, human evaluation showed that the model faced challenges in aligning the pronominal (man) in the two languages (Badawi, 2023).

viii. Human evaluation revealed problems with missing words, parts of sentences, content, and filler words in the translation of Russian to English. Problems with incorrect words included mistranslation of proper nouns and incorrect sense (Shukshina, 2019).

ix. In the translation between English and Nyishi, human evaluation of adequacy and fluency found similar low scores of adequacy and high scores of fluency in both directions (Kakum et al., 2023).

These results illustrate the difficulty of comparing BLEU to human evaluations, which assess adequacy, fluency, and other error categories without rating scales.

Table 11

Comparison of Human and Automated Evaluation

Language Pair	**Automated Evaluation**		**Human Evaluation**		**Study**
	Metric	Value	Metric	Value	
Levantine-MSA	BLEU	63.99	Scale of 1-7	6.46	Baniata et al., 2022
Maghrebi-MSA	BLEU	61.07	Scale of 1-7	6.40	Baniata et al., 2022
Gulf-MSA	BLEU	47.21	Scale of 1-7	5.95	Baniata et al., 2022
Iraqi-MSA	BLEU	58.33	Scale of 1-7	5.90	Baniata et al., 2022
Nile-MSA	BLEU	47.15	Scale of 1-7	6.39	Baniata et al., 2022
English-Arabic	BLEU	97.22	Scale of 1-7	4.2	Nagi, 2023
Arabic-English	BLEU	88.72	Scale of 1-7	4.8	Nagi, 2023
Chinese-English	BLEU	24.9	Scale of 1-10	7.6	Liu et al., 2023
English-Irish	BLEU	52.7	MQM		Lankford et al., 2022
Russian-Kazakh	BLEU	15.3			Tukeyev et al., 2019
Kazakh-Russian	BLEU	14.4			Tukeyev et al., 2019
Chinese-Japanese	BLEU	22.9			Zhang et al., 2023
Turkish-English	BLEU	36-22			Dogru, 2022
Turkish-English	BLEU	25-29			Dogru, 2022
Greek-English	BLEU	32.59	Error categorization		O'Shea et al., 2023
Russian-Vietnamese	BLEU	14.84	Adequacy		Nguyen et al., 2021
Kurdish-English	BLEU	45	Adequacy		Badawi, 2023
Russian-English	BLEU	24.82	Error categorization		Shukshina, 2019
English-Nyishi	BLEU	10.18	Adequacy/fluency		Kakum et al., 2023

Nyishi-English	BLEU	15.43	Adequacy/ fluency		Kakum et al., 2023
English-Chinese	F1	42.52			Xie et al., 2022
Chinese-English	F1	48.86			Xie et al., 2022
English-German	F1	53.08			Xie et al., 2022
German-English	F1	63.34			Xie et al., 2022
English-Malayalam	BLEU	2.6	Scale 1-4	1.67	Pathak & Pakray, 2019
English-Tamil	BLEU	6.15	Scale 1-4	2.57	Pathak & Pakray, 2019
English-Hindi	BLEU	3.57	Scale 1-4	1.72	Pathak & Pakray, 2019
English-Punjabi	BLEU	11.38	Scale 1-4	2.71	Pathak & Pakray, 2019
Nyishi-English	TER	83.4			Kakum et al., 2023
Nyishi-English	METEOR	0.19			Kakum et al., 2023
Nyishi-English	F1	0.42			Kakum et al., 2023
Nyishi-English	TER	92.1			Kakum et al., 2023
Nyishi-English	METEOR	0.15			Kakum et al., 2023
Nyishi-English	F1	0.43			Kakum et al., 2023

Limitations of Automated Metrics

Limitations of automated metrics are summarized below

i. BLEU scores are high when translating in the general domain but drop significantly when translating in specific domains (Pham et al., 2023).

ii. BLEU disproportionately penalizes long and short sentences leading to lower BLEU scores in these situations (Berrichi & Mazroui, 2021; Hu et al., 2023; Peng et al., 2021; Wan et al., 2022). Similar degradation in WER, TER, chr-F, and COMET has been observed in short and long sentences (Mahanty et al., 2023; Mahsuli et al., 2023).

iii. BLEU scores are high in morphologically similar languages, but a high number of unknown words in morphologically dissimilar languages leads to lower BLEU scores (Pathak & Pakray, 2019). Similarly, in low resource situations, BLEU and chr-F scores are low (Berrichi & Mazroui, 2021; Lalrempui & Soni, 2023).

iv. Metrics such as BLEU are development tools that are inadequate indicators of NMT quality, and other metrics that factor in the post-editing effort should also be considered (Alvarez-Vidal & Oliver, 2023). Furthermore, automated metrics provide different perspectives on NMT quality. While F-measure shows similarity in the number of words, TER shows the amount of editing, and BLEU shows matching words in a line which can be confusing (Ulitkin et al., 2021). Additionally, BLEU does not show how each error influences quality (Wan et al., 2022). Also, BLEU can be negatively correlated with human evaluation as BLEU uses lexical precision in source and target texts. However, such lexical differences are insignificant to human evaluators (Pathak & Pakray, 2019).

v. Unknown words, noise, ambiguity, and case sensitivity reduce BLEU scores (Aqlan et al., 2019; Ulitkin et al., 2022; Wang, 2022).

vi. Quantitative lexical diversity metrics such as TTR and MTLD suggest NMT systems are more lexically diverse compared to humans. Still, human evaluation showed those metrics are not a reliable measure of lexical diversity in translating English to Slovenian (Brglez & Vintar, 2022).

NMT Quality Improvement

The approaches that were found to increase translation quality are highlighted below.

i. Back-translation improved the BLEU score and mitigated the problem of colloquial text. Back-translation has the advantages of not requiring changes in network architecture and adaptability to other language pairs (Bala Das et al., 2023; Liu et al., 2023; Pham et al., 2023; Zhang & Matsumoto, 2019).

ii. Data segmentation improved the BLEU score. Morphological segmentation and Romanization minimized the problem of unknown words and improved translation quality (Aqlan et al., 2022; Berrichi & Mazroui, 2021; Ngo et al., 2022; Zhang & Matsumoto, 2019).

iii. Adding contextual information and balancing data can mitigate translation problems associated with short sentences. Furthermore, incorporating source linguistic knowledge, syntax awareness, and word sense or entity disambiguation can improve the BLEU score (Nguyen et al., 2018; Pan et al., 2020; Peng et al., 2021; Qing-dao-er-ji et al., 2022; Wan et al., 2022; Xie et al., 2022; Yan, 2022). Although providing document-level context improved the translation of context-specific sentences, it had minimal or no effect on sentences that were not context-specific (Nayak et al., 2022).

iv. Byte-pair encoding, reverse positional encoding, and round-trip training improved automated metrics (Ahmadnia & Dorr, 2019; Baniata et al., 2022; Lankford et al., 2022). Specifically, using byte pair encoding alone significantly improved the BLEU score in the translation of Russian to English compared to either lowercase, tokenization, or both. Simultaneous use of the three approaches provided further gains (Shukshina, 2019). Furthermore, CSE segmentation was superior to byte-pair encoding in reducing vocabulary volume when translating Kazakh to English (Tukeyev et al., 2020).

v. Bidirectional data diversification, improving model structure, using synthetic corpora, corpora pre-processing, and using simplified corpus improved automated metrics in the translation of low-resource language pairs (Li et al., 2020; Mahanty et al., 2023; Mahata et al., 2022; Qing-dao-er-ji et al., 2022; Tukeyev et al., 2019).

vi. Using transformer architecture alternatives such as RNN and BRNN improved translation quality (Farooq et al., 2023; Karyukin et al., 2023).

vii. The domain adaptation approach of multi-register was found to improve automated metrics in translating Castilian to Spanish (Uguet & Aranberri, 2023).

viii. An intelligent algorithm and a transformer aimed at correcting the problem of unknown words have been observed to significantly improve BLEU scores when translating English to Chinese (Wang, 2022).

ix. Using CNN as a feature extraction layer improved BLEU scores better than part of speech tagging and entity recognition (Liu et al., 2023).

x. Incorporating SMT into NMT has been observed to significantly improve BLEU score in translating English to Slovenian, but there was only a marginal improvement in translating Slovenian to English (Dugonik et al., 2023).

xi. Modeling sentence length mitigated NMT limitation of quality degradation on unknown sentence length. In the translation of German to English and English to Arabic, BLEU score improvements of 9.82 and 6.28 were observed. Similar improvements in TER, chr-F2, and COMET were observed (Mahsuli et al., 2023).

xii. In bi-directional translation between English and 13 Indic languages, transliteration was found to minimize lexical gap and improve quality in all pairs (Lalrempuii & Soni, 2023).

Discussion

The first and second objectives of this SLR were to investigate challenges in NMT quality and performance of automated and human evaluation metrics across language pairs. The first significant challenge is the lack of a large and high-quality parallel corpus. This problem is specifically severe in low-resource languages and specific domains. This becomes clear when automated metrics are examined. In translating low-resource languages such as Sinhala to English, Nepali to English, and English to Nepali, BLEU scores of less than eight were observed, and data augmentation could not increase BLEU scores by more than two points. Bi-directional translation of English and Nyishi, Russian to Vietnamese, and translation of French to Korean yielded BLEU scores of less than 16. Lower NMT quality is clear when translating in specific domains.

When translating English to Vietnamese, which is not considered a low resource pair, there was a difference of 9 BLEU points between the general and legal domains. Translating the Bible from Mizo to English, a low resource and domain-specific situation, yielded BLEU scores of less than 16, and human evaluation suggested SMT had better translation than NMT. Singh and Hujon (2020) similarly found SMT had higher BLEU scores than NMT in low-resource and specific domains. The worse performance of NMT was attributed to the general limitation of NMT in low-resource situations and reliance on a single reference despite multiple possible translations. Other studies have similarly found NMT is inferior in low-resource situations (Ahmadnia & Dorr, 2020; Chu & Wang, 2020; Kri & Sambyo, 2024).

The challenge of corpus quantity and quality is further exemplified by looking at BLEU scores of high-resource languages. Bi-directional translation of English and Arabic yielded BLEU scores higher than 80. Domain-specific translation of Russian to English, Japanese to English, English to Chinese, Turkish to English, and Greek to English yielded BLEU scores higher than 27, suggesting corpora quality is the key to NMT translation quality. A case in point is an increase in BLEU score by 19.2 points when the corpus quality was improved in the translation of Chinese to Japanese. Banerjee et al. (2023) similarly observe parallel corpora is a critical prerequisite in machine translation. Although comparable corpora may be easy to find, their quality limits direct use in NMT or SMT. Pre-processing of the corpora is essential. Adjeisah et al. (2021) argue that "large-scale parallel corpora" are available only for Western languages. Translation between these languages was observed to yield higher BLEU scores. However, high BLEU scores were also observed when translating between Japanese and Korean, which may not be considered Western languages.

Inconsistencies in BLEU scores were evident, with some studies reporting high and low BLEU scores in the same language pair. This can be explained by the use of varying corpus. Inconsistencies between METEOR, BLEU, and TER were similarly observed. These differences can be attributed to the quality aspect measured by each metric. BLEU measures lexical similarity, WER measures edit distance, and METEOR measures semantic similarity (Lee et al., 2023). For example, a language may have a high lexical similarity but require more edit operations.

The second major challenge to NMT is morphological diversity. Languages such as Korean, Kazakh, Arabic, and Indian languages are morphologically diverse, which creates a high number of unknown words. This becomes clear when BLEU scores of individual pairs are examined. Bidirectional translation of Arabic and Chinese, Korean and English, Korean and Spanish yielded BLEU scores of less than 25. This is in contrast to higher BLEU scores observed in morphologically similar languages such as Arabic dialects and MSA, English and Spanish, Japanese and Chinese, Korean and Japanese, English and German, English and Irish, Castilian and Spanish, and Mongolian and Chinese. Nasir and Mchechesi (2022) note that transfer learning from morphologically similar languages is a viable strategy for improving low-resource translation. This strategy can also benefit morphologically dissimilar languages.

The third objective was to investigate the strengths and limitations of automated and human metrics. Current NMT automated quality evaluation is dominated by lexical-based metrics such as BLEU, TER, WER, chr-F, and METEOR. These metrics are often well correlated such that high BLEU scores occur together with low WER and TER scores, high F-measure, and high chr-F scores. Specifically, lower WER and TER values have been observed in the translation of English and Spanish, Japanese and Korean, and German and English, which are morphologically similar. In contrast, high TER scores have been observed in the translation of English and Nyishi, which are low-resource languages. This suggests lexical metrics measure a common dimension of NMT quality.

However, interpretation of these metrics is not straightforward as they do not provide end users with an accurate perspective of the quality to be expected from NMT systems. Specifically, these metrics do not give a clear indication of the post-editing effort required. BLEU scores higher than 0.5 indicate a good and fluent translation that requires minimal post-editing (Denkowski & Lavie, 2010; O'Shea et al., 2023). However, such scores were hardly achievable even in morphologically similar and high-resource languages. This suggests significant post-editing effort may be required, and in low-resource situations, NMT may not provide any productivity gains. However, Zouhar et al. (2021) argue there is an unclear relationship between "MT quality and post-editing time." Professional translators need to be aware higher automated metrics may not necessarily lead to shorter post-editing periods or better post-edited quality.

BLEU scores are worse in specific domains, on longer sentences, at higher grams, and when noise is present in the corpus. This is expected in other lexical metrics, but it may not be a specific limitation of lexical metrics but a general NMT limitation. Some studies showed BLEU was well correlated with human evaluation, but other studies indicated BLEU was poorly correlated with human evaluation. This poor correlation can be explained by the focus on lexical precision in language pairs when calculating BLEU. In contrast, such lexical differences are not important in human evaluation. Chauhan et al. (2021) note the poor correlation between BLEU and human evaluation can be worse in morphologically rich languages due to "strict matching of words" (n.p.) and propose AdaBLEU as an alternative. AdaBLEU incorporates lexical and syntactic characteristics into the BLEU score.

An important limitation of evaluation metrics examined in this SLR is the lack of consistency. Some studies used the MQM framework, other studies used scales between 0 and 5 or 0 and 10, while other studies used error classification. Besides methodological differences, the reproducibility of human evaluation is challenging (Han, 2016; Vidal & Oliver, 2023; Vilar et al., 2006). This makes human assessment comparison across studies difficult.

The fourth objective was to identify measures that can be used to improve NMT quality. High-resource and low-resource languages face different challenges; therefore, quality improvement measures will be different for these languages. For high-resource and morphologically diverse languages, back-translation, morphological segmentation, sentence segmentation, domain adaptation, and context awareness were found to be effective. Data augmentation was the major quality improvement observed in low-resource languages.

Implications for Research and Practice

i. Current NMT has made good progress in achieving and evaluating lexical precision between source and target languages. However, other language dimensions, such as fluency, adequacy, and style, are lacking. NMT research needs to shift focus to these other dimensions and specifically develop metrics that can be used to evaluate them. Furthermore, research is required to create robust post-editing effort metrics.

ii. Interpretability of current automated evaluation metrics is lacking. There is a need to develop benchmarks for specific language pairs to guide end users on the level of system performance expected at particular values of automated metrics.

iii. There is a lack of consistency in methodologies used for human evaluation. Therefore, there is a need to develop a harmonized framework for human evaluation.

iv. Although there has been a general shift from SMT to NMT, specifically the transformer architecture, more research is needed on the value of SMT and alternative NMT architectures in low-resource and domain-specific situations.

Conclusion

Although NMT has made important progress in bridging the gap with human translation, there is no SLR that has attempted to synthesize current knowledge on NMT quality. The objective of this SLR was to bridge this gap by specifically investigating NMT quality constraints, the performance of human and automated metrics across language pairs, and quality improvement. The key constraints to NMT that emerged from reviewed articles are corpus availability and morphological diversity. Examination of these characteristics alongside automated lexical metrics revealed five groupings of language pairs. The first grouping is high-resource languages that are morphologically different. A case in point is English and Arabic, which, despite being morphologically divergent, had very high BLEU scores. The second grouping is high resource morphologically similar languages, such as European languages, and some Asian languages, such as Chinese, Korean, and Japanese.

The third grouping is medium-resourced morphologically divergent languages such as Korean and French. The fourth grouping is low-resource languages such as Nyishi and English, which have a tiny corpus. The fifth group is domain-specific situations that can arise in any of the first four categories. There are wide-ranging disparities in quality in these categories. Therefore, it can be concluded that progress in NMT quality does not include all language pairs, but promising methods to mitigate corpus availability and morphological diversity have been proposed. Examination of evaluation methods revealed that lexical metrics are dominant in NMT quality evaluation and that they measure a common quality dimension. However, there was no consistency in human evaluation methods used.

Therefore, the conclusion made in some studies that automated metrics do not correlate well with human evaluation could not be made in this SLR. The lack of interpretability of lexical metrics and their inability to assess aspects such as fluency and adequacy show the need to change NMT focus to other language aspects. However, these results need to be interpreted with an understanding of the limitations of this SLR. Although the search was comprehensive, it is possible some relevant articles were not identified as they did not include search terms in the title, abstract, or keywords.

Bio

Dr. Najia AbdulKareem AlGhamedi is an Assistant Professor at the College of Language Sciences, Department of English, King Saud University in Riyadh. Her research interests include evaluation of translation, literary translation and the sociology of translation.

References

Adjeisah, M., Liu, G., Nyabuga, D. O., Nortey, R. N., & Song, J. (2021). Pseudotext injection and advance filtering of low-resource corpus for neural machine translation. *Computational Intelligence and Neuroscience, 2021*(1), Article 6682385. https://doi.org/10.1155/2021/6682385

Ahmadnia, B., & Dorr, B. J. (2019). Augmenting neural machine translation through round-trip training approach. *Open Computer Science*, *9*(1), 268–278. https://doi.org/10.1515/comp-2019-0019

Alimova, S. (2021). Morphological classification of languages. *International Journal of Multidisciplinary Research and Analysis.* https://doi.org/10.47191/ijmra/v4-i5-19

Alvarez-Vidal, S., & Oliver, A. (2023). Assessing MT with measures of PE effort. *Ampersand*, *11*, Article 100125. https://doi.org/10.1016/j.amper.2023.100125

Amrhein, C., & Sennrich, R. (2022). Identifying weaknesses in machine translation metrics through minimum Bayes risk decoding: A case study for COMET. *arXiv*. https://doi.org/10.48550/arXiv.2202.05148

Aqlan, F., Fan, X., Alqwbani, A., & Al-Mansoub, A. (2019). Arabic–Chinese neural machine translation: Romanized Arabic as subword unit for Arabic-sourced translation. *IEEE Access, 7,* 133122–133135. https://doi.org/10.1109/ACCESS.2019.2941161

Badawi, S. (2023). Transformer-based neural network machine translation model for the Kurdish Sorani dialect. *UHD Journal of Science and Technology, 7,* 15–21. https://doi.org/10.21928/uhdjst.v7n1y2023.pp15-21

Bala Das, S., Biradar, A., Kumar Mishra, T., & Kr. Patra, B. (2023). Improving multilingual neural machine translation system for Indic languages. *ACM Transactions on Asian and Low-Resource Language Information Processing, 22*(6), Article 169:1-169:24. https://doi.org/10.1145/3587932

Banerjee, A., Kumar, V., Shankar, A., Jhaveri, R. H., & Banik, D. (2023). Automatic resource augmentation for machine translation in low-resource language: EnIndic Corpus. *ACM Transactions on Asian and Low-Resource Language Information Processing.*

Banerjee, S., & Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In J. Goldstein, A. Lavie, C.-Y. Lin, & C. Voss (Eds.), *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization* (pp. 65–72). Association for Computational Linguistics. https://aclanthology.org/W05-0909

Baniata, L. H., Kang, S., & Ampomah, I. K. E. (2022). A reverse positional encoding multi-head attention-based neural machine translation model for Arabic dialects. *Mathematics, 10*(19), Article 19. https://doi.org/10.3390/math10193666

Benkova, L., Munkova, D., Benko, Ľ., & Munk, M. (2021). Evaluation of English–Slovak neural and statistical machine translation. *Applied Sciences, 11*(7), Article 7. https://doi.org/10.3390/app11072948

Berrichi, S., & Mazroui, A. (2021). Addressing limited vocabulary and long sentences constraints in English–Arabic neural machine translation. *Arabian Journal for Science and Engineering, 46*(9), 8245–8259. https://doi.org/10.1007/s13369-020-05328-2

Brglez, M., & Vintar, Š. (2022). Lexical diversity in statistical and neural machine translation. *Information, 13*(2), Article 2. https://doi.org/10.3390/info13020093

Castilho, S., Moorkens, J., Gaspari, F., Sennrich, R., Way, A., & Georgakopoulou, P. (2018). Evaluating MT for massive open online courses: A multifaceted comparison between PBSMT and NMT systems. *Machine Translation, 32*(3), 255–278.

Cer, D., Manning, C. D., & Jurafsky, D. (2010, June). The best lexical metric for phrase-based statistical MT system optimization. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 555–563).

Chatzikoumi, E. (2019). How to evaluate machine translation: A review of automated and human metrics. *Natural Language Engineering, 26,* 1–25. https://doi.org/10.1017/S1351324919000469

Chauhan, S., Daniel, P., Mishra, A., & Kumar, A. (2023). Adableu: A modified BLEU score for morphologically rich languages. *IETE Journal of Research, 69*(8), 5112–5123.

Chauhan, S., Saxena, S., & Daniel, P. (2022). Improved unsupervised neural machine translation with semantically weighted back translation for morphologically rich and low-resource languages. *Neural Processing Letters, 54*(3), 1707–1726. https://doi.org/10.1007/s11063-021-10702-8

Chu, C., & Wang, R. (2020). A survey of domain adaptation for machine translation. *Journal of Information Processing, 28,* 413–426.

Denkowski, M., & Lavie, A. (2010, June). Extending the METEOR machine translation evaluation metric to the phrase level. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 250–253).

Devi, C. S., & Purkayastha, B. S. (2023). An empirical analysis on statistical and neural machine translation system for English to Mizo language. *International Journal of Information Technology, 15*(8), 4021–4028. https://doi.org/10.1007/s41870-023-01488-0

Doddington, G. (2002). Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. *Proceedings of the Second International Conference on Human Language Technology Research,* 138–145.

Dogru, G. (2022). Translation quality regarding low-resource, custom machine translations: A fine-grained comparative study on Turkish-to-English statistical and neural machine translation systems. *95–115.* https://doi.org/10.26650/iujts.2022.1182687

Dugonik, J., Sepesy Maučec, M., Verber, D., & Brest, J. (2023). Reduction of neural machine translation failures by incorporating statistical machine translation. *Mathematics, 11*(11), Article 11. https://doi.org/10.3390/math11112484

Farooq, U., Rahim, M., & Abid, A. (2023). A multi-stack RNN-based neural machine translation model for English to Pakistan sign language translation. *Neural Computing and Applications, 35,* 1–14. https://doi.org/10.1007/s00521-023-08424-0

Farrús, M., Costa-jussa, M., Poch, M., Hernández, A., & Mariño, J. (2009). Improving a Catalan-Spanish statistical translation system using morphosyntactic knowledge.

Flanagan, M. (1994, October 5). Error classification for MT evaluation. *Proceedings of the First Conference of the Association for Machine Translation in the Americas. AMTA 1994*, Columbia, Maryland, USA. https://aclanthology.org/1994.amta-1.9

Freitag, M., Foster, G., Grangier, D., Ratnakar, V., Tan, Q., & Macherey, W. (2021). Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics, 9*, 1460–1474. https://doi.org/10.1162/tacl_a_00437

Glushkova, T., Zerva, C., & Martins, A. F. (2023). BLEU meets COMET: Combining lexical and neural metrics towards robust machine translation evaluation. *arXiv preprint* arXiv:2305.19144.

Han, L. (2018). Machine translation evaluation resources and methods: A survey. *arXiv*. https://doi.org/10.48550/arXiv.1605.04515

Han, C. (2020). Translation quality assessment: A critical methodological review. *The Translator*, *26*(3), 257-273.

Hasibuan, Z. (2020). A comparative study between human translation and machine translation as an interdisciplinary research. *Journal of English Teaching and Learning Issues, 3*(2), Article 2. https://doi.org/10.21043/jetli.v3i2.8545

Hassan, H., Aue, A., Chen, C., Chowdhary, V., Clark, J., Federmann, C., Huang, X., Junczys-Dowmunt, M., Lewis, W., Li, M., Liu, S., Liu, T.-Y., Luo, R., Menezes, A., Qin, T., Seide, F., Tan, X., Tian, F., Wu, L., … Zhou, M. (2018). Achieving human parity on automatic Chinese to English news translation. *arXiv*. https://doi.org/10.48550/arXiv.1803.05567

Hirao, R., Arai, M., Shimanaka, H., Katsumata, S., & Komachi, M. (2020). Automated essay scoring system for nonnative Japanese learners. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Twelfth Language Resources and Evaluation Conference* (pp. 1250–1257). European Language Resources Association. https://aclanthology.org/2020.lrec-1.157

Hlaing, Z. Z., Thu, Y. K., Supnithi, T., & Netisopakul, P. (2022). Improving neural machine translation with POS-tag features for low-resource language pairs. *Heliyon*, *8*(8), e10375. https://doi.org/10.1016/j.heliyon.2022.e10375

Hu, S., Li, X., Bai, J., Lei, H., Qian, W., Hu, S., Zhang, C., Akpatsa, S., Qiu, Q., Zhou, Y., & Yang, S. (2023). Neural machine translation by fusing key information of text. *Computers, Materials & Continua*, *74*, 2803–2815. https://doi.org/10.32604/cmc.2023.032732

Huang, F., & Papineni, K. (2007). Hierarchical system combination for machine translation. In J. Eisner (Ed.), *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)* (pp. 277–286). Association for Computational Linguistics. https://aclanthology.org/D07-1029

Kakum, N., Laskar, S. R., Sambyo, K., & Pakray, P. (2023). Neural machine translation for limited resources English-Nyishi pair. *Sādhanā, 48*(4), 237. https://doi.org/10.1007/s12046-023-02308-8

Karyukin, V., Rakhimova, D., Karibayeva, A., Turganbayeva, A., & Turarbek, A. (2023). The neural machine translation models for the low-resource Kazakh–English language pair. *PeerJ Computer Science*, *9*, e1224. https://doi.org/10.7717/peerj-cs.1224

KchaouSaméh, BoujelbaneRahma, & HadrichLamia. (2023). Hybrid pipeline for building Arabic Tunisian dialect-standard Arabic neural machine translation model from scratch. *ACM Transactions on Asian and Low-Resource Language Information Processing*. https://doi.org/10.1145/3568674

Kirchhoff, K., Capurro, D., & Turner, A. M. (2014). A conjoint analysis framework for evaluating user preferences in machine translation. *Machine Translation*, *28*(1), 1–17. https://doi.org/10.1007/s10590-013-9140-x

Kraus, S., Breier, M., & Dasí-Rodríguez, S. (2020). The art of crafting a systematic literature review in entrepreneurship research. *International Entrepreneurship and Management Journal*, *16*(3), 1023–1042. https://doi.org/10.1007/s11365-020-00635-4

Kri, R., & Sambyo, K. (2024). Comparative study of low-resource Digaru language using SMT and NMT. *International Journal of Information Technology*, *16*(4), 2015-2024.

Kumar, A., Parida, S., Pratap, A., & Singh, A. K. (2023). Machine translation by projecting text into the same phonetic-orthographic space using a common encoding. *Sādhanā*, *48*(4), 238. https://doi.org/10.1007/s12046-023-02275-0

Lalrempuii, C., & Soni, B. (2023). Extremely low-resource multilingual neural machine translation for Indic Mizo language. *International Journal of Information Technology*, *15*(8), 4275–4282. https://doi.org/10.1007/s41870-023-01480-8

Lankford, S., Afli, H., & Way, A. (2022). Human evaluation of English–Irish transformer-based NMT. *Information*, *13*(7), Article 7. https://doi.org/10.3390/info13070309

Lee, S., Lee, J., Moon, H., Park, C., Seo, J., Eo, S., Koo, S., & Lim, H. (2023). A survey on evaluation metrics for machine translation. *Mathematics*, *11*(4), Article 4. https://doi.org/10.3390/math11041006

Li, Y., Li, X., Yang, Y., & Dong, R. (2020). A diverse data augmentation strategy for low-resource neural machine translation. *Information*, *11*(5), Article 5. https://doi.org/10.3390/info11050255

Lihua, Z. (2022). The relationship between machine translation and human translation under the influence of artificial intelligence machine translation. *Mobile Information Systems*, *2022*, 1–8. https://doi.org/10.1155/2022/9121636

Liu, H., Ye, Z., Zhao, H., & Yang, Y. (2023). Chinese text de-colloquialization technique based on back-translation strategy and end-to-end learning. *Applied Sciences*, *13*(19), Article 19. https://doi.org/10.3390/app131910818

Liu, Z., Chen, Y., & Zhang, J. (2023). Neural machine translation of electrical engineering based on integrated convolutional neural networks. *Electronics*, *12*(17), Article 17. https://doi.org/10.3390/electronics12173604

Lo, C. (2019). YiSi—A unified semantic MT quality evaluation and estimation metric for languages with different levels of available resources. In O. Bojar, R. Chatterjee, C. Federmann, M. Fishel, Y. Graham, B. Haddow, M. Huck, A. J. Yepes, P. Koehn, A. Martins, C. Monz, M. Negri, A. Névéol, M. Neves, M. Post, M. Turchi, & K. Verspoor (Eds.), *Proceedings of the Fourth Conference on Machine Translation* (Volume 2: Shared Task Papers, Day 1) (pp. 507–513). Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-5358

Lo, C., & Wu, D. (2011). MEANT: An inexpensive, high-accuracy, semi-automatic metric for evaluating translation utility based on semantic roles. In D. Lin, Y. Matsumoto, & R. Mihalcea (Eds.), *Proceedings of the 49th Annual Meeting of the Association for*

Computational Linguistics: Human Language Technologies (pp. 220–229). Association for Computational Linguistics. https://aclanthology.org/P11-1023

Lommel, A. (2018). Metrics for translation quality assessment: A case for standardising error typologies. *Translation quality assessment: From principles to practice*, 109–127.

Ma, Q., Bojar, O., & Graham, Y. (2018). Results of the WMT18 metrics shared task: Both characters and embeddings achieve good performance. In O. Bojar, R. Chatterjee, C. Federmann, M. Fishel, Y. Graham, B. Haddow, M. Huck, A. J. Yepes, P. Koehn, C. Monz, M. Negri, A. Névéol, M. Neves, M. Post, L. Specia, M. Turchi, & K. Verspoor (Eds.), *Proceedings of the Third Conference on Machine Translation: Shared Task Papers* (pp. 671–688). Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-6450

Ma, Q., Graham, Y., Wang, S., & Liu, Q. (2017). Blend: A novel combined MT metric based on direct assessment — CASICT-DCU submission to WMT17 metrics task. In O. Bojar, C. Buck, R. Chatterjee, C. Federmann, Y. Graham, B. Haddow, M. Huck, A. J. Yepes, P. Koehn, & J. Kreutzer (Eds.), *Proceedings of the Second Conference on Machine Translation* (pp. 598–603). Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-4768

Macháček, M., & Bojar, O. (2014). Results of the WMT14 metrics shared task. In O. Bojar, C. Buck, C. Federmann, B. Haddow, P. Koehn, C. Monz, M. Post, & L. Specia (Eds.), *Proceedings of the Ninth Workshop on Statistical Machine Translation* (pp. 293–301). Association for Computational Linguistics. https://doi.org/10.3115/v1/W14-3336

Mahanty, M., Vamsi, B., & Madhavi, D. (2023). A corpus-based auto-encoder-and-decoder machine translation using deep neural network for translation from English to Telugu language. *SN Computer Science*, *4*(4), 354. https://doi.org/10.1007/s42979-023-01678-4

Mahata, S. K., Garain, A., Das, D., & Bandyopadhyay, S. (2022). Simplification of English and Bengali sentences for improving quality of machine translation. *Neural Processing Letters*, *54*(4), 3115–3139. https://doi.org/10.1007/s11063-022-10755-3

Mahsuli, M. M., Khadivi, S., & Homayounpour, M. M. (2023). LenM: Improving low-resource neural machine translation using target length modeling. *Neural Processing Letters*, *55*(7), 9435–9466. https://doi.org/10.1007/s11063-023-11208-1

Maučec, M. S., & Donaj, G. (2019). Machine translation and the evaluation of its quality. *Recent trends in computational intelligence*, 143.

Mathur, N., Baldwin, T., & Cohn, T. (2020). Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 4984–4997). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.448

Mengist, W., Soromessa, T., & Legese, G. (2019). Method for conducting systematic literature review and meta-analysis for environmental science research. *MethodsX*, *7*, 100777. https://doi.org/10.1016/j.mex.2019.100777

Mirela. (2024, January 9). The role of high-resource languages in NLP and localization. *POEditor Blog*. https://poeditor.com/blog/high-resource-languages/

Muftah, M. (2022). Machine vs human translation: A new reality or a threat to professional Arabic–English translators. *PSU Research Review*, ahead-of-print(ahead-of-print). https://doi.org/10.1108/PRR-02-2022-0024

Nagi, K. A. (2023). Arabic and English relative clauses and machine translation challenges. *Journal of Social Studies*, *29*(3), Article 3. https://doi.org/10.20428/jss.v29i3.2180

Nasir, M. U., & Mchechesi, I. A. (2022). Geographical distance is the new hyperparameter: A case study of finding the optimal pre-trained language for English-isiZulu machine translation. *arXiv preprint arXiv:2205.08621*.

Nayak, P., Haque, R., Kelleher, J. D., & Way, A. (2022). Investigating contextual influence in document-level translation. *Information*, *13*(5), Article 5. https://doi.org/10.3390/info13050249

Ngo, T.-V., Nguyen, P.-T., Nguyen, V. V., Ha, T.-L., & Nguyen, L.-M. (2022). An efficient method for generating synthetic data for low-resource machine translation: An empirical study of Chinese, Japanese to Vietnamese neural machine translation. *Applied Artificial Intelligence*, *36*(1), 2101755. https://doi.org/10.1080/08839514.2022.2101755

Nguyen, P., Vo, A.-D., Shin, J.-C., & Ock, C.-Y. (2018). Effect of word sense disambiguation on neural machine translation: A case study in Korean. *IEEE Access*, *PP*, 1–1. https://doi.org/10.1109/ACCESS.2018.2851281

Nguyen, T., Nguyen, H., & Tran, P. (2021). Sublemma-based neural machine translation. *Complexity*, *2021*, e5935958. https://doi.org/10.1155/2021/5935958

Nightingale, A. (2009). A guide to systematic literature reviews. *Surgery (Oxford)*, *27*(9), 381–384. https://doi.org/10.1016/j.mpsur.2009.07.005

O'Shea, J., Sosoni, V., & Stasimioti, M. (2023). Translating law: A comparison of human and post-edited translations from Greek to English. *78*, 92–12. https://doi.org/10.2436/rld.i78.2022.3704

Pan, Y., Li, X., Yang, Y., & Dong, R. (2020). Multi-source neural model for machine translation of agglutinative language. *Future Internet*, *12*(6), Article 6. https://doi.org/10.3390/fi12060096

Panić, M. (2020). Everything you need to know about DQF. *TAUS*. https://www.taus.net/resources/blog/everything-you-need-to-know-about-dqf

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: A method for automatic evaluation of machine translation. In P. Isabelle, E. Charniak, & D. Lin (Eds.), *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (pp. 311–318). Association for Computational Linguistics. https://doi.org/10.3115/1073083.1073135

Pathak, A., & Pakray, P. (2019). Neural machine translation for Indian languages. *Journal of Intelligent Systems, 28*(3), 465–477. https://doi.org/10.1515/jisys-2018-0065

Peng, R., Hao, T., & Fang, Y. (2021). Syntax-aware neural machine translation directed by syntactic dependency degree. *Neural Computing and Applications, 33*(23), 16609–16625. https://doi.org/10.1007/s00521-021-06256-4

Pham, N. L., Vinh Nguyen, V., & Pham, T. V. (2023). A data augmentation method for English-Vietnamese neural machine translation. *IEEE Access, 11*, 28034–28044. https://doi.org/10.1109/ACCESS.2023.3252898

Popel, M., Tomkova, M., Tomek, J., Kaiser, Ł., Uszkoreit, J., Bojar, O., & Žabokrtský, Z. (2020). Transforming machine translation: A deep learning system reaches news translation quality comparable to human professionals. *Nature Communications, 11*(1), Article 1. https://doi.org/10.1038/s41467-020-18073-9

Popović, M. (2018). Error classification and analysis for machine translation quality assessment. In *Translation quality assessment: From principles to practice* (pp. 129–158).

Popović, M., & Arčan, M. (2015). Identifying main obstacles for statistical machine translation of morphologically rich South Slavic languages. In İ. D. El-Kahlout, M. Özkan, F. Sánchez-Martínez, G. Ramírez-Sánchez, F. Hollowood, & A. Way (Eds.), *Proceedings of the 18th Annual Conference of the European Association for Machine Translation* (pp. 97–104). https://aclanthology.org/W15-4913

Qing-dao-er-ji, R., Cheng, K., & Pang, R. (2022). Research on traditional Mongolian-Chinese neural machine translation based on dependency syntactic information and transformer model. *Applied Sciences, 12*(19), Article 19. https://doi.org/10.3390/app121910074

Rei, R., Stewart, C., Farinha, A. C., & Lavie, A. (2020). COMET: A neural framework for MT evaluation. In B. Webber, T. Cohn, Y. He, & Y. Liu (Eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 2685–2702). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.213

Reiter, E., & Belz, A. (2009). An investigation into the validity of some metrics for automatically evaluating natural language generation systems. *Computational Linguistics*, *35*(4), 529–558. https://doi.org/10.1162/coli.2009.35.4.35405

Rivera-Trigueros, I. (2022). Machine translation systems and quality assessment: A systematic review. *Language Resources and Evaluation*, *56*(2), 593–619. https://doi.org/10.1007/s10579-021-09537-5

Sanchez-Torron, M., & Koehn, P. (2016, November). Machine translation quality and post-editor productivity. In *Twelfth Conference of the Association for Machine Translation in the Americas* (pp. 16–26). Association for Machine Translation in the Americas, AMTA.

Sismat, M. A. H. (2020). Analysing patterns of errors in neural and statistical machine translation of Arabic and English. *JALL| Journal of Arabic Linguistics and Literature*, *2*(2), 126–142.

Shukshina, E. (2019). The impact of some linguistic features on the quality of neural machine translation. *Journal of Applied Linguistics and Lexicography*, *1*, 365–370. https://doi.org/10.33910/2687-0215-2019-1-2-365-370

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., & Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, 223–231. https://aclanthology.org/2006.amta-papers.25

Singh, T. D., & Hujon, A. V. (2020, July). Low resource and domain specific English to Khasi SMT and NMT systems. In *2020 International Conference on Computational Performance Evaluation (ComPE)* (pp. 733–737). IEEE.

Stanojević, M., & Sima'an, K. (2015). BEER 1.1: ILLC UvA submission to metrics and tuning task. In O. Bojar, R. Chatterjee, C. Federmann, B. Haddow, C. Hokamp, M. Huck, V. Logacheva, & P. Pecina (Eds.), *Proceedings of the Tenth Workshop on Statistical Machine Translation* (pp. 396–401). Association for Computational Linguistics. https://doi.org/10.18653/v1/W15-3050

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. https://doi.org/10.1016/j.aiopen.2020.11.001

Tukeyev, U., Karibayeva, A., & Abduali, B. (2019). Neural machine translation system for the Kazakh language based on synthetic corpora. *MATEC Web of Conferences*, *252*, 03006. https://doi.org/10.1051/matecconf/201925203006

Tukeyev, U., Karibayeva, A., & Zhumanov, Z. H. (2020). Morphological segmentation method for Turkic language neural machine translation. *Cogent Engineering*, *7*(1), 1856500. https://doi.org/10.1080/23311916.2020.1856500

Turian, J. P., Shen, L., & Melamed, I. D. (2003, September 23). Evaluation of machine translation and its evaluation. *Proceedings of Machine Translation Summit IX: Papers*. MTSummit 2003, New Orleans, USA. https://aclanthology.org/2003.mtsummit-papers.51

Uguet, C. S., & Aranberri, N. (2023). Exploring politeness control in NMT: Fine-tuned vs. multi-register models in Castilian Spanish. *Procesamiento del Lenguaje Natural*, *70*(0), Article 0.

Ulitkin, I., Filippova, I., Ivanova, N., & Poroykov, A. (2021). Automatic evaluation of the quality of machine translation of a scientific text: The results of a five-year-long experiment. *E3S Web of Conferences*, *284*, 08001. https://doi.org/10.1051/e3sconf/202128408001

Vardaro, J., Schaeffer, M., & Hansen-Schirra, S. (2019). Translation quality and error recognition in professional neural machine translation post-editing. *Informatics*, *6*(3), Article 3. https://doi.org/10.3390/informatics6030041

Vigier-Moreno, F., & Macías, L. (2022). Assessing neural machine translation of court documents: A case study on the translation of a Spanish remand order into English. *Revista de Llengua i Dret*, *78*, 73–91. https://doi.org/10.2436/rld.i78.2022.3691

Vilar, D., Xu, J., D'Haro, L. F., & Ney, H. (2006). Error analysis of statistical machine translation output. In N. Calzolari, K. Choukri, A. Gangemi, B. Maegaard, J. Mariani, J. Odijk, & D. Tapias (Eds.), *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2006/pdf/413_pdf.pdf

Wan, Y., Yang, B., Wong, D. F., Chao, L. S., Yao, L., Zhang, H., & Chen, B. (2022). Challenges of neural machine translation for short texts. *Computational Linguistics, 48*(2), 321–342. https://doi.org/10.1162/coli_a_00435

Wang, P. (2022). A study of an intelligent algorithm combining semantic environments for the translation of complex English sentences. *Journal of Intelligent Systems*, *31*, 623–631. https://doi.org/10.1515/jisys-2022-0048

Way, A. (2018). Quality expectations of machine translation: From principles to practice (pp. 159–178). https://doi.org/10.1007/978-3-319-91241-7_8

Webster, R., Fonteyne, M., Tezcan, A., Macken, L., & Daems, J. (2020). Gutenberg goes neural: Comparing features of Dutch human translations with raw neural machine translation outputs in a corpus of English literary classics. *Informatics, 7*(3), Article 3. https://doi.org/10.3390/informatics7030032

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., Dean, J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv*. https://doi.org/10.48550/arXiv.1609.08144

Xie, S., Xia, Y., Wu, L., Huang, Y., Fan, Y., & Qin, T. (2022). End-to-end entity-aware neural machine translation. *Machine Learning*, *111*(3), 1181–1203. https://doi.org/10.1007/s10994-021-06073-9

Yagahara, A., Masahito, U., & Yokoi, H. (2024). Evaluation of machine translation accuracy focused on the adverse event terminology for medical devices. *Studies in Health Technology and Informatics*, *310*, 1450–1451. https://doi.org/10.3233/SHTI231239

Yan, L. (2022). Real-time automatic translation algorithm for Chinese subtitles in media playback using knowledge base. *Mobile Information Systems*, *2022*, 1–11. https://doi.org/10.1155/2022/5245035

Yang, Y., Liu, R., Qian, X., & Ni, J. (2023). Performance and perception: Machine translation post-editing in Chinese-English news translation by novice translators. *Humanities and Social Sciences Communications*, *10*(1), Article 1. https://doi.org/10.1057/s41599-023-02285-7

Yuan, W., Neubig, G., & Liu, P. (2021). BARTScore: Evaluating generated text as text generation. *Advances in Neural Information Processing Systems*, *34*, 27263–27277. https://proceedings.neurips.cc/paper_files/paper/2021/hash/e4d2b6e6fdeca3e60e0f1a62fee3d9dd-Abstract.html

Zhang, J., & Matsumoto, T. (2019). Corpus augmentation for neural machine translation with Chinese-Japanese parallel corpora. *Applied Sciences*, *9*(10), Article 10. https://doi.org/10.3390/app9102036

Zhang, J., Tian, Y., Mao, J., Han, M., Wen, F., Guo, C., Gao, Z., & Matsumoto, T. (2023). WCC-JC 2.0: A web-crawled and manually aligned parallel corpus for Japanese-Chinese neural machine translation. *Electronics, 12*(5), Article 5. https://doi.org/10.3390/electronics12051140

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020). BERTScore: Evaluating text generation with BERT. *arXiv*. https://doi.org/10.48550/arXiv.1904.09675

Zhang, W., Dai, L., Liu, J., & Wang, S. (2023). Improving many-to-many neural machine translation via selective and aligned online data augmentation. *Applied Sciences, 13*(6), Article 6. https://doi.org/10.3390/app13063946

Zhang, X., Malkov, Y., Florez, O., Park, S., McWilliams, B., Han, J., & El-Kishky, A. (2023, August). Twhin-bert: A socially-enriched pre-trained language model for multilingual tweet representations at Twitter. In *Proceedings of the 29th ACM SIGKDD conference on knowledge discovery and data mining* (pp. 5597-5607).

Zouhar, V., Tamchyna, A., Popel, M., & Bojar, O. (2021). Neural machine translation quality and post-editing performance. *arXiv*. https://doi.org/10.48550/arXiv.2109.05016