Journal of Literature and Linguistics Studies
ISSN: 3078-4832
DOI: 10.61424/jlls
Journal Homepage: www.bluemarkpublishers.com/index.php/JLLS

JLLS
BLUEMARK PUBLISHERS

| RESEARCH ARTICLE

Advancements in Machine Translation and Cross-Language Computational Applications: Techniques, Challenges, and Future Directions

Irshad A. Naikoo[1] and Parvaiz A. Ganai[2] ✉
[12]Department of Foreign Languages, College of Arts and Humanities, Jazan University, Jazan, Saudi Arabia
Corresponding Author: Parvaiz A. Ganai, **E-mail**: pganai@jazanu.edu.sa

| ABSTRACT

Over the past few years, there has been spectacular growth in Machine Translation (MT) and cross-lingual computational processes based on developments in neural network architecture and the availability of large data sets. This paper presents a detailed overview of recent work, identifies major challenges, and proposes potential directions for future research. The discussion of the latest developments in machine translation begins with an overview of neural machine translation (NMT) and transformer-based models. It also addresses long-term challenges from low-resource language translation to persistent translation quality and intricate cross-linguistic variation control. Furthermore, the paper explores trending research directions and emerging trends, including zero-shot translation, cross-lingual embeddings, and interdisciplinary synergies among machine translation and other NLP tasks. By integrating such results, this paper intends to contribute to existing research and innovation, triggering further advancements in MT and cross-language technology.

| KEYWORDS

Machine Translation, Computational applications, Cross-lingual embedding, Natural Language process

| ARTICLE INFORMATION

ACCEPTED: 11 April 2025 **PUBLISHED:** 02 May 2025 **DOI:** 10.61424/jlls.v3.i2.270

1. Introduction

The concept of Machine Translation (MT) was first proposed by Warren Weaver in 1947 [1], just a year after the invention of the world's first computer—the Electronic Numerical Integrator and Computer (ENIAC). From its beginning, MT has been viewed as one of the most complex and demanding challenges in Natural Language Processing (NLP). Since then, MT has gone through a considerable evolution, pushed by developments in computational capabilities, expansion in linguistic theory, and the increasing availability of data. What began as a conceptual idea has now grown into a practical and widely used tool with multiple applications across many domains.

Traditionally, MT techniques have been divided into rule-based and corpus-based approaches. Rule-Based Machine Translation (RBMT) remained dominant between the 1950s and 1990s. It utilized bilingual dictionaries and manually crafted linguistic rules to translate text from a source language to a target language. MT was originally motivated by military needs, and a milestone event occurred in 1954 when Georgetown University and IBM conducted the first Russian-to-English translation using the IBM-701 computer, validating MT's feasibility and prompting optimism for over a decade. However, the ALPAC report in 1966 [2] did not approve its development, and the momentum declined. Despite this, the 1970s and 1980s were a period of consolidation for RBMT systems, with SYSTRAN becoming a leading commercial MT vendor in 1978, even providing Google's translation services until 2007. During that time, NLP research remained focused on rule-based parsing and translation, as seen at the first International Conference on Computational Linguistics in 1965. The second part of the 20th century brought a paradigm shift with the development of corpus-based methods, encouraged by the availability of bilingual corpora. The methods such as example-based machine translation (EBMT), statistical machine translation (SMT), and neural machine translation (NMT) later became prevalent. EBMT, which came into being during the mid-1980s, was based on bilingual corpora for retrieving similar sentence pairs to translate [3], with good performance regarding quality but limited by corpus coverage, and hence apt for

computer-assisted systems. The 2000s saw the popularization of SMT, with Google launching its internet translation service in 2006 with phrase-based SMT, followed by Microsoft and Baidu. However, hybrid systems combining multiple MT models were created for performance improvement [4] because of the complexity of the translation task. The breakthrough came in 2014 with Bahdanau et al.'s [5] and Sutskever et al.'s [6] NMT employing end-to-end neural networks and attention mechanisms to translate source text into dense semantic representations. This, along with Dong et al.'s multilingual system [7], changed the landscape. Baidu (2015) [8] and Google (2016) [9] were some of the companies that ran NMT systems into production. Innovations, such as convolutional sequence-to-sequence models [10] and the Transformer model [11], further pushed translation quality towards a higher level, pushing speculation as to whether MT could be approaching equality with human translation.

Meanwhile, spoken language translation research has progressed greatly, ranging from domain-limited systems to open-domain spontaneous translation systems [17]-[21]. It started with a small experimental system demonstrated at the International Telecommunication Union (ITU) exposition in 1983 [15]. It continued with milestones like advancement in the Speech Trans System in 1988 [16], which laid the foundation for today's speech-to-speech (S2S) translation technologies. The MT field continues to evolve rapidly, driven by sensational advances such as the advent of massive-scale pre-trained language models such as OpenAI's GPT-3 [22] and Google's T5 [23], which leverage gigantic datasets and computational resources to achieve state-of-the-art performance in multilingual translation. A second crucial progress is the advent of unsupervised and semi-supervised NMT approaches, which relax the requirement for large parallel corpora by leveraging monolingual data and advanced alignment techniques, such as those developed by Lample et al. [12] and Artetxe et al. [13]. In low-resource language pairs, these approaches have been working correctly and are being merged into real-world MT systems with increasing regularity. The Transformer model [11] remains the basis for state-of-the-art NMT systems, with improvements like Efficient Transformer variants [24] and sparse attention mechanisms [25] enhancing efficiency and scalability. Multilingual NMT models such as Google's M4 [26] and Facebook's M2M-100 [27] became commercially practical because of their ability to translate multiple languages using one model that can excel in zero-shot and few-shot configurations. For speech translation, end-to-end models like Facebook's S2ST [28] and Google's Translatotron [29] have achieved unprecedented progress in fluency and accuracy and enabled real-time and high-quality speech translation use cases. Moreover, integrating AI ethics with fairness in MT studies has also been a great necessity, where scientists have been involved in overcoming problems like biased translation models, ethics in machine-based translation, and the need for inclusive systems for underrepresented languages [30], [31].

This paper outlined the advancements, difficulties, and prospects of MT and cross-language computer applications. By summarizing previous milestones and recent breakthroughs, the present paper aims to provide a fair overview of the discipline and its ongoing development.

2. Techniques in Machine Translation

2.1 Neural Machine Translation (NMT)

Neural Machine Translation (NMT) has made significant improvements over the last few years, transforming the field of machine translation [25], [26]. Unlike earlier techniques, such as Rule-Based Machine Translation (RBMT) and Statistical Machine Translation (SMT), NMT utilizes deep learning techniques to mimic the translation process more naturally and efficiently. The two main key elements of an NMT system are an encoder network and a decoder network. The source sentences are encoder inputs and are converted into a compact, real-valued vector, commonly known as a "thought vector" or "context vector," which represents the meaning of the sentence. The decoder network then takes this vector as input and outputs the target sentence word by word in a word-by-word simulation of how a human translator would initially comprehend the source text and then output the translation word by word. The NMT has important advantages over previous methods in the end-to-end learning paradigm. Unlike RBMT, which is based on manually designed linguistic rules, or SMT, which requires statistical alignments and feature engineering, NMT trains to translate directly from large parallel corpora end-to-end. It learns translation patterns and semantic representations automatically through training and doesn't need explicit human intervention in rule or feature design. NMT models tend to be built on sequence-to-sequence (Seq2Seq) frameworks, occasionally with the addition of attention mechanisms. Attention allows the model to focus on specific positions in the source sentence when generating each word in the target sentence. This overcomes the limitation of fixed-size context vectors and allows the model to process long sentences. Even more recently, transformer-based models have continued to significantly boost the performance of NMT through leveraging self-attention mechanisms that enable parallelization and more accurately capture long-distance dependencies.

NMT's success is based on a variety of different factors: a) Higher Fluency and Accuracy – NMT produces more fluent and accurate contextualized translations compared to RBMT and SMT; b) Scalability – NMT models can be trained on huge corpora because there exist huge parallel corpora; c) Flexibility – NMT systems are very flexible to be adapted for application in a specific domain or language and hence are very flexible. With these advantages, NMT became the de facto standard for machine translation, introducing new levels of translation quality and impacting low-resource language translation, multilingual models,

and zero-shot translation research, among many more. While, hitherto, there has been incredible success in machine translation, the issues yet to be tackled are how to handle low-frequency words, coherence for large sections of text, and mitigating biases in training data. These issues have yet to be researched by scholars, who continue to solidify NMT as the optimal norm in machine translation.

2.2 Multilingual Translation

Multilingual translation is challenging and labor-intensive in that there is a built-in diversity within the morphologies, structures, and syntactic rules of languages. English and Chinese, for instance, are subject-verb-object (SVO) languages, while Korean and Japanese are subject-object-verb (SOV) languages. Translation across languages with such structural differences often entails long-distance word or phrase reordering. Languages also vary in terms of their morphological features. Chinese is an isolating language with minimal inflectional change, whereas Japanese is agglutinative with dense morphological changes. The contrast between linguistic features of the two languages renders multilingual MT extremely difficult, not only for machine systems but even for human translators. Parallel corpora in large volumes are used by data-driven MT models, such as Statistical Machine Translation (SMT) and Neural Machine Translation (NMT), to acquire translation patterns. Overall, translation quality improves with an increasing amount of training data. Koehn and Knowles [32], for instance, demonstrated that an increase in English–Spanish training data from 0.4 million words to 385.7 million words resulted in a 30% absolute improvement in BLEU score, a measure of translation quality. However, most of the languages are "resource-poor", which means there is insufficient parallel data to enable high-quality MT systems. According to InternetWorldStats, (2020), the top ten languages (e.g., English, Chinese, Spanish, etc.) account for roughly 77% of internet users, with English and Chinese together contributing 25.9% and 19.4%, respectively. The remaining together represent merely 23.1% of users. Though resource-rich languages like English and Chinese have the richness of billions of parallel sentence pairs, resource-poor language pairs like Chinese–Kiswahili or Chinese–Hindi can possess thousands of sentence pairs or even fewer, and hence, it is a challenging attempt to create fluent MT systems.

To solve the problem of data sparsity, studies have suggested novel means for getting the most out of existing data. One of these involves back-translation [33], [34], which employs monolingual data to augment training corpora. Here, the first NMT model is trained on a small parallel corpus and used to translate huge quantities of target-language monolingual text into the source language, creating a "pseudo-parallel corpus". This synthetic data is used to train the model again with considerably improved translation quality. A worst-case scenario where parallel data is not available has seen unsupervised methods in translation being investigated. For example, Lample et al. [12] proposed a method for mapping sentences from various languages to a shared latent space and learning translation through reconstructing sentences. In the same direction, Artetxe et al. [13] created improved unsupervised NMT through model initialization via sophisticated SMT techniques. Another direction involves approaches by Song et al. [14], Conneau and Lample [35], and Ren et al. [36] that leverage pretraining to facilitate unsupervised translation. Another aim of research for upgrading multilingual translation is utilizing resource-rich languages to ease the translation of resource-poor languages. This idea is not novel to the SMT era, and one of the popular approaches is pivot-based translation. In pivot-based translation, a high-resource language (e.g., English) is employed as a bridge between two low-resource languages. For example, for Chinese–German translation, English can be utilized as the pivot language, leveraging existing Chinese–English and English–German parallel corpora. The most straightforward pivot-based approach is the transfer approach, employing two cascaded systems: one translating from the source to the pivot and another from the pivot to the target language. While this is simple to use, the approach is susceptible to error propagation. Wu and Wang [37], [38] and Cohn and Lapata [39] avoided this by employing the triangulation approach, deriving phrase-level translation data from source–pivot and pivot–target models. Recently, multilingual NMT models have been proposed where translation of several languages is permitted with a single system. One such instance is the research by Johnson et al. [40], where a simple yet efficient method to add a special token in the source sentence to mark the target language has been provided. This allows the model to learn shared representations without modifying the structure for closely related languages. To further improve performance, Tan et al. [41] investigated clustering languages into clusters based on linguistic similarity and trained a separate model for each cluster. In practice, hybrid methods of translation are commonly used, which integrate multiple approaches to trade off translation quality, efficiency, and deployment cost. Owing to these advances, contemporary MT systems can facilitate translation between hundreds of languages. For example, Arivazhagan et al. [42] trained a very large multilingual MT model using over 50 billion parameters on more than 25 billion sentence pairs across 103 languages. Similarly, Fan et al. [43] introduced the M2M-100 model, which interprets any two languages among 100 languages without taking English as a pivot language with 7.5 billion sentence pairs. Despite these advances, there exists a challenge for multilingual translation, particularly for low-resource languages and linguistically distant language pairs. Ongoing research is aimed at improving data efficiency, reducing biases, and building the generalization ability of MT models, with the goal of making translation systems more complete and accurate in the future.

2.3 Simultaneous Translation (ST)

Simultaneous Translation (ST) is a very challenging task that attempts to translate with little delay in real-time from the source language speech to the translation, sacrificing not high translation quality. Compared to full-sentence translation, in which the

entire source sentence is listened to before the target translation is output, ST compels the system to translate incrementally while the source sentence is being spoken. This creates extremely strict challenges as the system must strike a balance between the quality of the translation and latency. ST research accommodates two general paradigms: the cascaded (pipeline) methods and the end-to-end approaches. Both have strengths and weaknesses, and both are being pursued enthusiastically for improving the art of real-time translation technology.

2.3.1 Simultaneous Speech-to-Speech (S2S) Translation Pipeline

There are three main elements of a typical cascaded ST system: a) Automatic Speech Recognition (ASR) – ASR converts speech from a source language into streaming text in real time; b) Machine Translation (MT) – The MT system translates the transcribed source text to the target language text; c) Text-to-Speech (TTS) – This module transforms the translated text into speech in the target language. It is an optional module and depends on whether the output needed is speech or text in specific application scenarios.

A cascaded approach is widely used in real-world systems because of its modularity and deployability. All the modules (ASR, MT, TTS) can be optimized independently, and advanced models can be inserted into the pipeline. This has the disadvantage of error propagation, where an error in the ASR module adversely impacts the MT and TTS modules and leads to low-quality translation. Moreover, the linearity of the pipeline introduces latency, which is a critical consideration in real-time translation.

2.3.2. Toward End-to-End Simultaneous Translation

Developing end-to-end ST systems that will translate source speech directly into target speech or text without any intermediate process is the main goal of ST research. The following are the potential gains: potentially reducing error propagation, improving efficiency, and reducing latency. End-to-end ST is very challenging to attain due to the following reasons: 1. Sparsity of Data: ST training data end-to-end are sparse. There are a few corpora with a maximum of hundreds of speech hours, e.g., Japanese–English [44], [45], and European languages [46], [47]. In the Chinese–English translation task, Baidu introduced an open corpus with 70 speech hours, transcription, and translation [48], but it was still too small to train strong models. 2. Complexity of integration: Speech recognition and translation combined in one framework is not a simple task. The system must work on acoustic, linguistic, and translational characteristics simultaneously, and this requires sophisticated modeling expertise.

Despite such problems, impressive progress has been made in end-to-end ST research: a) Pretraining and Multitask Learning – Multitask learning, or pretraining, has been utilized in studies to improve translation quality. For example, Bansal et al. [49] demonstrated that it is possible to do end-to-end speech translation without source transcriptions by leveraging a pre-trained encoder trained on ASR data. To improve the performance of speech translation, text translation data has been utilized in other studies [50], [51], and [52]; b) Knowledge Distillation – Liu et al. [53] employed knowledge distillation to transfer the high-quality MT model's knowledge into an end-to-end ST model with improved translation accuracy; c)Two-Stage Models – For handling the issue of sharing information among tasks, two-stage models [54], [55], and [56] have been introduced. The first-stage decoder performs the speech recognition and generates a hidden state, which is taken by the second-stage decoder to perform the translation; d) Interactive Models: Liu et al. [57] introduced an interactive end-to-end ST model where speech recognition and MT interact dynamically with one another and enhance performance on both tasks; e) Direct Speech-to-Speech Translation – Recent research [58], [59] has tried direct S2S translation with the abandonment of text-based intermediates. However, the performance in such methods lags due to rare training examples and the inherent problem of manipulating more than a single modality.

Most practical ST systems nowadays rely on cascaded systems since they are easier to implement and can generate high-quality translations. For instance, Xiong et al. [48] compared a pipeline ST system with human interpreters and found that while human interpreters tend to skip irrelevant information to reduce latency, ST systems provide more adequate translations. Similarly, Shimizu et al. [44] pointed out that even less skilled interpreters lose information while interpreting, stressing the difficulty of the task for both humans and machines. Despite the advances, ST remains a difficult task considering the demands of low latency, high accuracy, and robustness to diverse linguistic and acoustic conditions.

3. Applications of Machine Translation (MT)

Machine Translation (MT) is an inevitable tool in most sectors owing to its cost-effectiveness, speed, and continuously increasing translation quality. The costs of human translation range between $0.08 and $0.25 per word, depending on the translator's level of expertise, the language pair, and the complexity level of the text. On the other hand, MT systems are several orders of magnitude less expensive, often as low as $0.00001 per word, and hence a very cost-efficient option for large-scale translation works. This tremendous cost advantage, combined with the quick advances in the quality of translation, has led to the

widespread adoption of MT across many industries and domains. This significant cost advantage, coupled with rapid advances in translation quality, has led to the widespread acceptance of MT both in everyday and professional applications. A good example is Baidu Translate, which provides translations from any two of over 200 languages and processes over 100 billion characters a day. Fig. 1 illustrates the distribution of translation across the top eight domains in Baidu Translate, which indicates the diverse applications of MT. These domains include:

1. E-commerce: making cross-border trade possible by providing simple communication between traders and customers who are different linguistically.
2. Travel and Tourism: Supporting real-time translation of signs, menus, and conversations of tourists, enhancing their overall experience.
3. Education: Supporting language learning and access to learning materials in different languages.
4. Healthcare: Supporting communication between physicians and patients with different languages, increasing access to healthcare.
5. Legal and Government: Supporting ease in translating legal documents, policy, and international agreements.
6. Media and Entertainment: Making films, TV shows, and news reports available to international markets by localizing the content.
7. Customer Support: Dispensing multilingual customer support through chatbots and auto-translate interfaces.
8. Technology and Software: Translating software user interfaces, user guides, and technical writing to international markets.

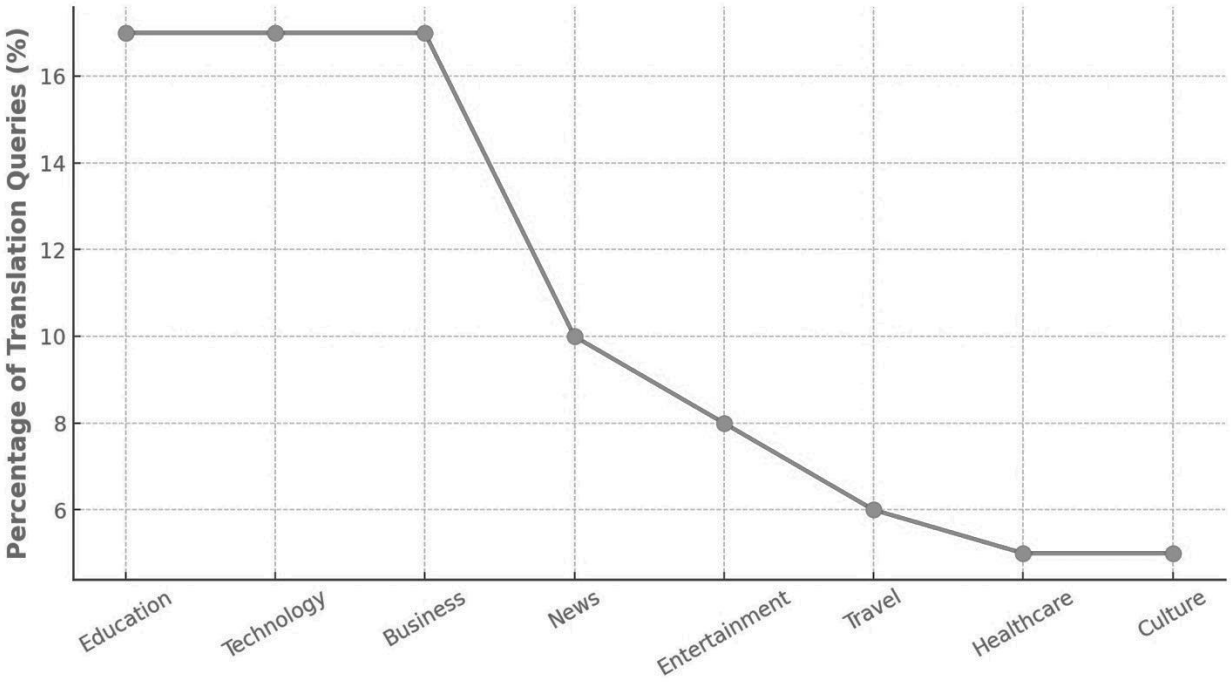

Fig. 1: Translation Distribution of Baidu Translate

The fact that it can handle such varied applications is a tribute to the MT system's versatility and growing maturity. It has also become more interfaced with real-time communication software, i.e., video conferencing software and messaging applications, in order to facilitate seamless cross-lingual communication at both the interpersonal and public sphere levels. MT can have many strengths, yet it remains weak in domain-specific terms, cultural nuance, and low-resource languages. Research is still engaged in addressing these weaknesses and making the accuracy and variety of MT systems even better. With improvements in technology in MT, its application is set to increase even more, bridging the language divide and facilitating global interconnectedness in an increasingly integrated world.

3.1 Text Translation

The most typical application of machine translation (MT) is perhaps the translation of the text. It can be utilized to translate text from the source language into the target language, supporting cross-linguistic interaction and information exchange. Some of the most obvious applications of text translation, with full descriptions of how they are utilized and their impacts, are as follows:

a) Webpage Translation: As the rapid pace of globalization speeds up, the need for immediate and efficient access to foreign-language information keeps growing. While using human translators to translate enormous volumes of web pages takes time and money, MT is a viable and scalable solution. The user only needs to cut and paste the content of a webpage or input its Uniform Resource Locator (URL) into an MT system to view its content in the preferred language. Advanced MT systems of today are integrated into browsers so that web pages can be translated in real-time without destroying the original page layout. This is achieved through advanced algorithms that preserve the page layout, formatting, and hyperlinks while translating the text. Additionally, deep learning-based NMT models have perfected webpage translation fluency and accuracy to the point of being nearly unrecognizable from human translations in most cases.

b) Scientific Literature Translation: Researchers, engineers, and graduate students routinely use MT to read scientific texts, such as research articles and patents, in their native languages. Conversely, they also use MT to translate their work into other languages so that it can be accessed by an international audience. A case in point is that the COVID-19 pandemic accelerated the translation of biomedical texts in a bid to support international collaboration to combat the virus. Scientific texts tend to have extremely dense, highly specialized vocabularies and complex sentence structures. In response to this, domain adaptation techniques are employed to counter this issue. These involve training a translation model on a general-purpose corpus at scale and then fine-tuning it on a smaller domain-specific corpus. This enables the model to handle technical terms and context-dependent nuances better. In addition, formatted document translation has gained traction, enabling users to translate documents in file formats like PowerPoint, Excel, Word, and PDF and maintaining formatting styles like font size, color, and layout.

c) E-commerce Translation: MT has become the most important part of the global e-commerce platform. MT tools are used by sellers to translate their website information, product descriptions, and user manuals into other languages to reach maximum customers. Buyers, on the other hand, are able to effortlessly browse and buy from international sellers without any language barriers. Besides product content, MT is also used in customer service to achieve maximum communication efficiency. For example, chatbots that are MT-enabled can provide instant real-time customer feedback in various languages, leading to increased customer satisfaction and business efficiency. Advanced MT technology in e-commerce is typically combined with NLP tools so that not only precise but also culturally appropriate translations are provided, considering regional dialects as well as customer behavior.

d) Language Learning: Modern MT systems have been blessed with many features that make them ideal tools for learners of languages. They encompass translation, quality dictionaries, sentence pair examples, and grammar-checking programs. Learners can use such features to look up unfamiliar words or phrases as well as see how they come across in context. For instance, students input full paragraphs to MT systems so that they assist in comprehension reading. They can also use sentence pair examples to improve their writing skill by comparing professionally translated sentences with their sentences. Sophisticated MT systems even offer interactive learning modules, such as vocabulary quizzes and pronunciation guides, to further help language learning. All these developments highlight the groundbreaking importance of MT in enabling language bridge construction across industries, corporate and educational settings, medicine, and international diplomacy. As AI continues to evolve, the dependability, efficiency, and ease of use of MT systems are only set to improve further, thereby making them an absolute imperative for the modern-day globally integrated world. In addition to the translation of the text, image translation and speech translation have been widely used in real applications after recent advancements in artificial intelligence techniques.

3.2. Image Translation

Image translation is an effective application of machine translation (MT) and computer vision technologies. It takes images as input, extracts text or visual information, and translates them into languages of relevance. Such a capability is increasingly crucial in today's economy with globalization when visual content is often required to be understood across linguistic and cultural boundaries. Two significant applications of image translation are outlined below, along with advanced discussions on their application and utilization:

a) Multilingual Image Captioning: Multilingual image captioning is a sophisticated MT that generates image descriptions in multiple languages. The technology not only gives descriptions of the images but also supports visual question answering (VQA), where users can ask questions about an image and receive answers in their preferred language. Recent research has made significant progress in this direction, employing neural machine translation (NMT) architectures to achieve state-of-the-art performance [60], [61], [62]. In multilingual image captioning, the encoder consumes an image and generates a target-language textual description as output. This is achieved using encoder-decoder models in which the encoder (often a convolutional neural network or CNN) operates on the visual representations of the image, and the decoder (oftentimes a recurrent neural network or transformer) emits the text in response. Subsequent versions can create captions in other languages for a single picture, so the technology is particularly beneficial for cross-linguistic studies and multilingual content production.

b) Optical Character Recognition (OCR) Translation: OCR translation is a powerful application of MT that involves recognizing text within images, translating the text, and inserting the translated text into the image in its original form. It is extremely convenient for translating menus, signs, product labels, and other text-heavy images one finds in foreign environments. For example, a foreign visitor can read a restaurant menu instantly using OCR translation. The recent advances in OCR translation have focused on modeling layout and text information in document images at the same time [63]. This allows MT systems to not only translate text but also maintain the original formatting, such as font styles, sizes, and spatial layouts. Modern OCR translation systems support transformer-based models and attention mechanisms for improved accuracy and efficiency. They are trained in large volumes of data with diverse document types, enabling them to handle complex layouts and multilingual documents. End-to-end models that combine OCR and MT into a single pipeline have also been suggested, which reduces errors and improves translation quality.

In addition to the above uses, image translation is also used for augmented reality (AR) and virtual reality (VR) platforms. As an example, AR glasses with image translation capabilities can project translated text on real-world objects in real time, making the user experience more enhanced in foreign locations. Besides, multimodal translation models that combine image, text, and speech translation are emerging as a new field of AI innovation. These models would aim to provide fluent translation across modes, where the user can interact with multilingual content more naturally. As an example, a user could take a photo of a sign in a different language, listen to it read aloud in translation, and have the translation drawn over the image within one application.

3.3 Speech Translation

Speech translation is an effective application of speech processing and machine translation (MT) technology. It takes spoken input in a source language and generates text or speech output in a target language. Speech translation has witnessed remarkable growth in recent years with advancements in automatic speech recognition (ASR), neural machine translation (NMT), and text-to-speech (TTS) systems. Some of the most significant applications and advancements in speech translation are discussed below:

a) Simultaneous Translation (ST): As explained in Section 2.3, there has been a remarkable improvement in simultaneous translation (ST), which provides real-time spoken language translation. Such technology is now implemented in different products and services so that it becomes available to have a broad category of use cases.

Speech-to-Text (S2T) Translation systems convert spoken words into written words in real-time. As a user comfort option, the ASR output (text rendering of the speech as heard first) and the target texts are usually shown at a single location. Due to limitations on how much can be shown on the screen, one can show only a single pair of languages in a single session in an S2T system; thus, scaling up is a limitation in multi-language environments. Speech-to-Speech (S2S) Translation systems counter the limitation of S2T since they translate spoken words directly into translated speech. The users can listen to the translated text in their preferred language using devices such as smartphones or headphones. In global conferences, for example, speakers from different countries can choose to listen to the translation in their native language. S2S systems are now an essential tool in global events, enabling simple cross-lingual communication. ST systems are also now integrated into conferencing services like Zoom, Microsoft Teams, and Google Meet, which support real-time translation for the participants. During the COVID-19 pandemic, the adoption of virtual conferences and meetings has accelerated. ST applications and plugins also allow the user to view foreign-language videos, such as movies, lectures, and live streams, with real-time voiceover or subtitles in the desired language. Progress in end-to-end neural models has acutely reduced the latency for ST systems, thereby making them more

suitable for real-time applications. Techniques such as streaming ASR and incremental NMT enable faster and improved translation, even for long and complex sentences.

b) Portable Translation Devices: Handheld translation devices have been extremely popular over the last few years because of their portability and versatility. These devices use edge computing and AI chips to provide instant and accurate translations with or without internet connectivity. Offline translation of multiple languages is also supported on some devices, making them reliable where internet connectivity is poor. These portable devices are equipped with sophisticated speech translation technology, so they prove useful in a vast range of different contexts: i) Language Learning: Mobile devices facilitate students practicing pronunciation and comprehension through instant feedback and translation. ii) Overseas Travel: These devices enable tourists to speak with the people, read the signs, and explore alien surroundings without language being a problem. iii) Business Negotiations: Experts use hand-held translators to facilitate cross-language communication during meetings, negotiations, and networking meetings.

c) Creative Applications of MT: Poem and Couplet Generation: In addition to their utilitarian applications, MT models have also been creatively employed for aesthetic and cultural purposes, e.g., poem generation. MT models are capable of generating poems by formulating the process as a sequence-to-sequence task. The process is one line generated at a time; and the previous line is the source for each line, and the next line is the target. More sophisticated models, such as GPT-based models, have been trained on large poetry corpora and fine-tuned to come up with coherent and aesthetically satisfying lines.

4. Challenges and Future Directions

Though an incredible amount of improvement has occurred in machine translation (MT), plenty of room remains for progress. At meetings such as the Workshop on Statistical Machine Translation (WMT), it is at times portrayed that human translators are outperformed by machines. Scores such as BLEU (Bilingual Evaluation Understudy), WER (Word Error Rate), and METEOR (Metric for Evaluation of Translation with Explicit Ordering) [64], [65], [66] seem to prove this statement. However, these numbers may not necessarily represent the guidelines for an effective translation. A proper translation should exhibit some significant qualities: adequacy (faithfulness to the original work) and fluency (naturalness in the target work). While neural machine translation (NMT) methods have achieved significant results for certain language pairs and topics, particularly text translation, they are far from ideal in a lot of areas, particularly speech translation (ST) applications. The following are the main challenges and future directions of MT.

4.1 The Need for Better Evaluation Metrics

Current measures like BLEU and WER are more focused on lexical and syntactic correctness but not on other critical aspects of translation quality. For instance, human interpreters also consider contextual appropriateness, emphasis, and timing in simultaneous interpretation. They know how to omit less important details to reduce latency and how to emphasize more important issues to convey the intent of the speaker. On the other hand, MT systems tend to translate everything literally, which results in inefficiencies and a lack of nuance.

Emphasis and Prosody: More recent research has explored the use of acoustic signals (i.e., stress, tone, and pitch) to identify emphasis in speech and echo this in the translated text. However, time-synching translation with a speaker's body language and prosody remains problematic. For example, a presenter's hand movements or indicating on slides during a presentation are often lost in translation. Future MT systems will have to integrate multimodal inputs (e.g., visual and audio signals) in an attempt to more accurately capture and reflect these nuances.

Latency and Synchronization: For software such as speech translation (ST), synchronization and latency are paramount. Systems that exhibit unacceptable delay need to be penalized by measures, and the measures need to synchronize translations with the rhythm of the speaker as well as visual aids (e.g., slides). To measure these factors, new paradigms that are beyond current metrics with an emphasis on completeness rather than usability are required.

Context and Understanding: The ideal translation would balance understanding against word-for-word precision. Metrics would penalize systems producing unnecessary or redundant information but favor systems with a focus on contextually relevant content. Developing context-aware evaluation metrics is a promising way forward.

4.2 Improving Robustness

A common problem in the MT systems is the robustness; a small change to the source text, like adding a word or punctuation mark, can produce a translation that has remarkable differences. Contrarily, humans have a high capacity for error tolerance, which allows them to deal with ambiguities, mistakes, and non-standard language phenomena with ease.

Explainable MT: Creating explainable MT techniques that offer insight into the translation creation process is one way to address this issue. This would make it possible to find and fix the system's flaws, making it stronger and more reliable.

Adversarial Training: MT systems can become more resilient to noisy or imperfect inputs by employing adversarial training techniques. Models can be made more resilient to variation in the real world by being exposed to a range of difficult situations during training.

Error Correction Mechanisms: Adding error detection and correction modules to MT pipelines can also make them more resilient. Before translating, these modules can find and fix any errors or inconsistencies in the original text.

4.3 Addressing Data Sparsity in Low-Resource Scenarios

NMT models require massive parallel training data, typically tens or hundreds of millions of sentence pairs. In the case of low-resource language pairs and domain-limited pairs, though, such data is typically not available. Humans, on the other hand, are effective learners, even in a few instances. It is one of the hardest challenges that MT research has to overcome to close this gap.

Data Augmentation: Back-translation, synthetic data generation, and noise injection have been employed to augment training data for low-resource languages. These can be utilized to improve translation quality but are not as effective across languages and domains.

Multitask and Transfer Learning: Multitask learning and transfer learning techniques use information from related tasks or languages to improve performance on low-resource tasks. Pretraining models on high-resource languages and then adapting them to low-resource languages, for instance, has demonstrated encouraging outcomes.

Few-Shot and Zero-Shot Learning: Few-shot and zero-shot learning advancements will assist in enabling MT systems to be effective with minimal or no parallel data. They operate through multilingual pretraining (e.g., on mBART or mT5 models) and cross-lingual transfer for language generalizability.

Unsupervised and Semi-Supervised Methods: Unsupervised methods of MT that do not require parallel data and semi-supervised methods that use a small amount of labeled data along with enormous amounts of unlabeled data are also explored to address data sparsity.

4.4 Expanding to Multimodal and Multilingual Translation

Multimodality and multilinguality are the keys to the future of machine translation, where systems are capable of processing a variety of inputs and outputs with ease.

Multimodal Translation: Increasing research into the combination of text, speech, and visual inputs—e.g., text translation in images or videos. For example, video translation can offer real-time dubbing or subtitles for overseas-language videos, while image-to-text translation can translate texts into documents, menus, or billboards.

Multilingual Systems: The main goal is to develop machine translation (MT) systems that are capable of translating between hundreds of languages. Even though it is challenging at present to promise consistent quality across all language pairs, recent models like Google's Universal Speech Model (USM) and Meta's NLLB-200 have made great strides in this direction.

4.5 Ethical and Inclusive MT

Since MT systems are used at random, ethics and inclusivity have emerged as crucial issues to address. MT systems have the potential to inadvertently reinforce biases present in the training data. To ensure that the translations are impartial and fair, methods for identifying and combating bias must be developed.

Privacy and Security: For real-time applications like speech translation, user data must be protected. To protect sensitive data, different strategies, such as federated learning and differential privacy, can be developed.

Low-Resource Language Inclusivity: To maintain linguistic diversity and inclusivity worldwide, MT's capacity to incorporate unstudied and endangered languages must be expanded.

4.6 Human-AI Collaboration

Future machine translation (MT) systems should try to enhance human capabilities rather than replace them. For instance:
Post-Editing Tools: Creating sophisticated post-editing and machine-aided translation tools can increase the productivity of human translators.

Interactive MT: Translation quality can be enhanced, and output can be adjusted to user preferences by systems that let users offer comments or corrections while the translation is being done.

5. Conclusion

Machine Translation (MT) and cross-language computational software have achieved revolutionary advancements in the past few years, stimulated by breakthroughs in neural network architectures, the availability of large-scale data, and the use of multimodal technologies. From the early rule-based systems to the current era, when Neural Machine Translation (NMT) and transformer models rule the roost, MT has evolved into a flexible tool that transcends linguistic and cultural borders. Despite these unmatchable achievements, much is yet to be done, and the field continues to push the boundaries of what is possible. NMT brought a revolution with the capability of end-to-end training and producing more natural-sounding and contextually accurate translations than ever. The use of attention mechanisms, particularly in transformer-based models, has further enhanced the ability to cope with long-distance dependencies and complex linguistic patterns. Techniques such as back-translation, pivot translation, and unsupervised learning have gone a long way in addressing data scarcity for low-resource languages, while multilingual NMT models capable of translating between hundreds of languages using a single model have shown outstanding performance in zero-shot and few-shot scenarios. End-to-end and cascaded real-time translation systems have made significant progress in reducing delay and increasing accuracy and are now integrated into all sorts of applications, from international conferences to pocket-sized translation devices, to help enable seamless cross-lingual communication. The integration of text, speech, and vision inputs has given rise to new possibilities for MT, such as image translation of text contained within images and speech translation systems that provide real-time subtitles or voiceovers for movies and live shows. As MT systems become more widespread, work to address bias, privacy, and inclusion issues is becoming ever more central, with an emphasis on including underrepresented and vulnerable languages essential to linguistic diversity maintenance and fair access to translation technology.

Current measures like BLEU and WER are beneficial but still unable to properly encode subtleties of translation quality, including emphasis, context, and synchronization, so more comprehensive evaluation frameworks must be established to advance the field. MT systems are not strong with robustness, i.e., small differences in input cause radically different output, and thus, there is a requirement for higher error tolerance as well as explainable models. Even as transfer learning and data augmentation progress, low-resource language translation remains an issue, with few-shot and zero-shot learning approaches and unsupervised techniques offering promising directions that are yet to be optimized. Much progress has been made in unifying text, speech, and visual inputs; however, end-to-end multimodal translation remains a challenging problem. Future systems need to coordinate more between visual and audio inputs and their translations to enhance the user experience. Ensuring MT systems are free from bias, that they preserve user privacy, and that they support underrepresented languages is central to responsibly deploying them, with continued work in ethical AI and inclusive design being important.

Developing systems that learn to answer the personal taste, voice, and jargon of specific users will render MT more accessible and precise. Rather than replacing human translators, new-generation MT systems must supplement human abilities with advanced post-editing technology and interactive MT systems facilitating real-time feedback to improve translation quality and productivity. MT could revolutionize industries such as health care, law firms, and education by providing accurate and obtainable translations for professional content, with continuous research on domain adaptation and terminological management being the backbone. Advances in hardware and algorithms are enabling faster and more effective real-time MT systems, and portable devices that can function offline are becoming sought after, pushing MT into poorly connected domains. Besides utilitarian uses, MT is being employed in innovative artistic and cultural use cases, such as composing poetry and preserving the form of traditional poetry, exemplifying its role in developing cultural heritage. In addition, MT has proved to be a tool for promoting inclusivity, collaboration, and understanding across linguistic and cultural divides. As the field develops

further, its influence on global trade, education, and communication will only increase, making it an essential component of our increasingly interconnected world.

Funding: This research received no external funding.
Conflicts of Interest: The authors declare no conflict of interest.
ORCID ID: https://orcid.org/0009-0009-8797-3406

References

[1] Weaver, W. (1955). *Translation*. *Machine Translation and Linguistics*, 14, 15–23.
[2] Hutchins, W. J. (2003). ALPAC: The (in)famous report. In S. Nirenburg, H. L. Somers, & Y. A. Wilks (Eds.), *Readings in machine translation* (pp. [pages if known]). MIT Press.
[3] Nagao, M. (1984). A framework of a mechanical translation between Japanese and English by analogy principle. In A. Elithorn & R. Banerji (Eds.), *Proceedings of the International NATO Symposium on Artificial and Human Intelligence*(pp. 173–180). Elsevier North-Holland.
[4] Wang, H. (2006). Multi-strategy machine translation. In Y. Q. Cao & M. S. Sun (Eds.), *Frontiers of Chinese information processing* (pp. 45–52). Tsinghua University Press.
[5] Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, May 7–9, 2015, San Diego, USA.
[6] Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. *Advances in Neural Information Processing Systems*.
[7] Dong, D., Wu, H., He, W., Yu, D., & Wang, H. (2015). Multi-task learning for multiple language translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 1723–1732). Association for Computational Linguistics.
[8] Pouliquen, B. (2017). WIPO Translate: Patent neural machine translation publicly available in 10 languages. In *Machine Translation XVI*.
[9] Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, Ł., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Rudnick, A., Corrado, G., Hughes, M., & Dean, J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation.
[10] Gehring, J., Auli, M., Grangier, D., Yarats, D., & Dauphin, Y. N. (2017). Convolutional sequence to sequence learning. *Proceedings of the 34th International Conference on Machine Learning*, 70, 1243–1252.
[11] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. A., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS 2017)*, Long Beach, CA, USA, December 4–9, 2017.
[12] Lample, G., Conneau, A., Denoyer, L., & Ranzato, M. A. (2017). Unsupervised machine translation using monolingual corpora only. *Proceedings of the International Conference on Learning Representations (ICLR) 2018*, Vancouver, BC, Canada, April 30–May 3, 2018.
[13] Artetxe, M., Labaka, G., & Agirre, E. (2019). An effective approach to unsupervised machine translation. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*, Florence, Italy, July 28–August 2, 2019.
[14] Song, K., Tan, X., Qin, T., Lu, J., & Liu, T. Y. (2019). MASS: Masked sequence to sequence pre-training for language generation. *Proceedings of the 36th International Conference on Machine Learning (ICML 2019)*, Long Beach, CA, USA, June 9–15, 2019.
[15] Kato, Y. (1995). The future of voice-processing technology in the world of computers and communications. *Proceedings of the National Academy of Sciences of the United States of America (PNAS)*, 92(22), 10060–10063.
[16] Tomita, M., & Tomabechi, H. (1990). SpeechTrans: An experimental real-time speech-to-speech translation. *Language Resources and Evaluation*, 26(4), 663–672.
[17] Sumita, E., Shimizu, T., & Nakamura, S. (2007). NICT-ATR speech-to-speech translation system. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL 2007)*, Prague, Czech Republic, June 25–27, 2007.
[18] Fügen, C., Kolss, M., Paulik, M., Stüker, S., Schultz, T., & Waibel, A. (2006). Open domain speech translation: From seminars and speeches to lectures. In *Proceedings of the TC-STAR Workshop on Speech-to-Speech Translation*, Barcelona, Spain, June 19–21, 2006.
[19] Moser-Mercer, B., Künzli, A., & Korac, M. (1998). Prolonged turns in interpreting: Effects on quality, physiological, and psychological stress (pilot study). *Interpreting*, 3(1), 47–64.
[20] Wang, H., Wu, H., Hu, X., Liu, Z., Li, J., Ren, D., & Niu, Z. (2008). The TCH machine translation system for IWSLT 2008. In *Proceedings of the 5th International Workshop on Spoken Language Translation (IWSLT 2008)*, Waikiki, Hawaii, October 20–21, 2008.
[21] Nakamura, S., Markov, K., Nakaiwa, H., Kikui, G., Kawai, H., Jitsuhiro, T., & Yamamoto, H. (2006). The ATR multilingual speech-to-speech translation system. *IEEE Transactions on Audio, Speech, and Language Processing, 14*(2), 365–376.
[22] He, H., Boyd-Graber, J., & Daumé, H. III. (2016). Interpretese vs. translationese: The uniqueness of human strategies in simultaneous interpretation. In *Proceedings of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2016)* (pp. 1117–1126). Association for Computational Linguistics.
[23] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research, 21*(140), 1–67.
[24] Ma, M., Huang, L., Xiong, H., Zheng, R., Liu, K., Zhang, B., He, Z., Liu, H., Li, X., Wu, H., & Wang, H. (2019). STACL: Simultaneous translation with implicit anticipation and controllable latency using prefix-to-prefix framework. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)* (pp. 3025–3036). Association for Computational Linguistics.
[25] Zhang, J., & Zong, C. (2020). Neural machine translation: Challenges, progress and future. *Science China Technological Sciences, 63*(10), 2028–2050.
[26] Edunov, S., Ott, M., Auli, M., & Grangier, D. (2018). Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP 2018)* (pp. 489–500). Association for Computational Linguistics.

[27] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735–1780.
[28] Bengio, Y., Simard, P., & Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks, 5*(2), 157–166.
[29] He, W., He, Z., Wu, H., & Wang, H. (2016). Improved neural machine translation with SMT features. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI 2016)* (pp. 3023–3029). AAAI Press.
[30] Tu, Z., Lu, Z., Liu, Y., Liu, X., & Li, H. (2016). Modeling coverage for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)* (Vol. 1, pp. 76–85). Association for Computational Linguistics.
[31] Cheng, Y., Shen, S., He, Z., He, W., Wu, H., Sun, M., & Liu, Y. (2016). Agreement-based joint training for bidirectional attention-based neural machine translation. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI 2016)* (pp. 2761–2767). International Joint Conferences on Artificial Intelligence Organization.
[32] Koehn, P., & Knowles, R. (2017). Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation* (pp. 28–39). Association for Computational Linguistics.
[33] Sennrich, R., Haddow, B., & Birch, A. (2016). Improving neural machine translation models with monolingual data. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (Vol. 1, pp. 86–96). Association for Computational Linguistics.
[34] Poncelas, A., Shterionov, D., Way, A., Maillette de Buy Wenniger, G., & Passban, P. (2018). Investigating backtranslation in neural machine translation. *Proceedings of the 21st Annual Conference of the European Association for Machine Translation* (pp. 269–278). European Association for Machine Translation.
[35] Conneau, A., & Lample, G. (2019). Cross-lingual language model pretraining. In *Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019)* (pp. 7059–7069). Curran Associates, Inc.
[36] Ren, S., Wu, Y., Liu, S., Zhou, M., & Ma, S. (2019). Explicit cross-lingual pre-training for unsupervised machine translation. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP 2019)* (pp. 770–779). Association for Computational Linguistics.
[37] Wu, H., & Wang, H. (2007). Pivot language approach for phrase-based statistical machine translation. *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics* (pp. 856–863). Association for Computational Linguistics.
[38] Wu, H., & Wang, H. (2009). Revisiting pivot language approach for machine translation. *Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2009)* (pp. 154–162). Association for Computational Linguistics.
[39] Cohn, T., & Lapata, M. (2007). Machine translation by triangulation: Making effective use of multi-parallel corpora. *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics* (pp. 728–735). Association for Computational Linguistics.
[40] Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., Hughes, M., & Dean, J. (2017). Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics, 5*, 339–351.
[41] Tan, X., Lu, Z., Sun, H., & Liu, Q. (2019). Multilingual neural machine translation with language clustering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP 2019)* (pp. 963–973). Association for Computational Linguistics.
[42] Arivazhagan, N., et al. (2019). Massively multilingual neural machine translation in the wild: Findings and challenges. *arXiv:1907.05019*.
[43] Fan, A., et al. (2020). Beyond English-centric multilingual machine translation. *arXiv:2010.11125*.
[44] Shimizu, H., et al. (2014). Collection of a simultaneous translation corpus for comparative analysis. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation* (pp. 4577–4584). European Language Resources Association (ELRA).
[45] Toyama, H., et al. (2004). CIAIR simultaneous interpretation corpus. In *Proceedings of Oriental COCOSDA* (pp. 104–109). New Delhi, India.
[46] Sandrelli, A., & Bendazzoli, C. (2004). Tagging a corpus of interpreted speeches: The European Parliament interpreting corpus (EPIC). In *Proceedings of LREC 2004* (pp. 745–748). European Language Resources Association (ELRA).
[47] Di Gangi, M. A., et al. (2019). MuST-C: A multilingual speech translation corpus. In *Proceedings of the 17th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019)* (pp. 2019). Association for Computational Linguistics.
[48] Xiong, H., et al. (2019). DuTongChuan: Context-aware translation model for simultaneous interpreting. *arXiv:1907.12984*.
[49] Bansal, S., et al. (2018). Pre-training on high-resource speech recognition improves low-resource speech-to-text translation. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL 2018)* (pp. 301–312). Association for Computational Linguistics.
[50] Weiss, R. J., et al. (2017). Sequence-to-sequence models can directly translate foreign speech. In *Proceedings of the 18th Annual Conference of the International Speech Communication Association* (pp. 1465–1469). International Speech Communication Association (ISCA).
[51] Anastasopoulos, A., & Chiang, D. (2018). Leveraging translations for speech transcription in low-resource settings. *Proceedings of the 19th Annual Conference of the International Speech Communication Association*, 2018, 365–376.
[52] Bérard, A., et al. (2016). Listen and translate: A proof of concept for end-to-end speech-to-text translation. *Proceedings of the 30th Conference on Neural Information Processing Systems*, 2016, 1–10.
[53] Liu, Y., et al. (2019). End-to-end speech translation with knowledge distillation. *Proceedings of the 20th Annual Conference of the International Speech Communication Association*, 2019, 1–5.
[54] Kano, T., Sakti, S., & Nakamura, S. (2017). Structured-based curriculum learning for end-to-end English–Japanese speech translation. *Proceedings of the 18th Annual Conference of the International Speech Communication Association*, 2017, 1–5.
[55] Anastasopoulos, A., & Chiang, D. (2018). Tied multitask learning for neural speech translation. *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2018, 82–91.
[56] Sperber, M., Neubig, G., Niehues, J., & Waibel, A. (2019). Attention-passing models for robust and data-efficient end-to-end speech translation. *Transactions of the Association for Computational Linguistics*, 7, 313–325.

[57] Liu, Y., Zhang, J., Xiong, H., Zhou, L., He, Z., Wu, H., Wang, H., & Zong, C. (2020). Synchronous speech recognition and speech-to-text translation with interactive decoding. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05), 8417–8424.
[58] Jia, Y., Weiss, R. J., Biadsy, F., Macherey, W., Johnson, M., Chen, Z., & Wu, Y. (2019). Direct speech-to-speech translation with a sequence-to-sequence model. *Proceedings of the 20th Annual Conference of the International Speech Communication Association*, 2019, 1–5.
[59] Kano, T., Sakti, S., & Nakamura, S. (2021). Transformer-based direct speech-to-speech translation with transcoder. *Proceedings of the IEEE Spoken Language Technology Workshop*, 2021, 1–5.
[60] Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: A neural image caption generator. *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition*, 2015, 3156–3164.
[61] Lu, J., Xiong, C., Parikh, D., & Socher, R. (2017). Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 375–383).
[62] Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., & Zhang, L. (2018). Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 6077–6086).
[63] Xu, Y., Li, M., Cui, L., Huang, S., Wei, F., & Zhou, M. (2020). LayoutLM: Pre-training of text and layout for document image understanding. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 1192–1200). ACM.
[64] Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (pp. 311–318). Association for Computational Linguistics.
[65] Tomás, J., Llitjós, A. F., Carbonell, J., Lavie, A., & Dorr, B. (2003). A quantitative method for machine translation evaluation. In *Proceedings of the EACL 2003 Workshop on Evaluation Initiatives in Natural Language Processing: Are Evaluation Methods, Metrics, and Resources Reusable?* (pp. 51–58). Association for Computational Linguistics.
[66] Banerjee, S., & Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization* (pp. 65–72). Association for Computational Linguistics.

1. Dr. Irshad Ahmad Naikoo has pursued his doctorate (PhD) from the Department of Linguistics, University of Kashmir, Srinagar, India. He has presented/published many research papers at various national and international platforms. He has also published many articles on English, Kashmiri and Balti languages. He has been associated with the Central Institute of Indian Languages Mysore and Urdu Teaching and Research Centre Lucknow (Ministry of HRD, Govt. of India). Previously, he was teaching at the Department of English, Government Degree College (Boys), Baramulla, Kashmir. He has been teaching at Department of Foreign Languages, College of Arts and Humanities, Jazan University, Jazan, Saudi Arabia since 2014. He is a member of Translation and Quality & Accreditation Committee at Jazan University.

2. Dr. Parvaiz Ahmad Ganai has pursued his doctorate (PhD) from the Department of Linguistics, University of Kashmir, Srinagar, India. He has presented/published many research papers at various national and international platforms. He has also published many articles on English, Kashmiri, and Kohistani languages. He has been associated with the Central Institute of Indian Languages Mysore and Urdu Teaching and Research Centre Lucknow (Ministry of HRD, Govt. of India). He is a reviewer of Journal of Language, Literature and Education (JELLE). Previously, he was teaching at the Department of English, Government Degree College for Women, Baramulla, Kashmir. He has been teaching at Department of Foreign Languages, College of Arts and Humanities, Jazan University, Jazan, Saudi Arabia since 2014. He is a member of Planning and Development, Re-evaluation, and Quality Assurance and Development Unit at Jazan University.