A Comprehensive Survey on Chinese to Malay Speech Translation System Based on Fine-Tuning Large Multilingual Translation Models

SurveyForge

Abstract— This survey explores the advancements and challenges in Chinese to Malay speech translation systems, emphasizing the role of fine-tuned large multilingual models. Such systems are critical for facilitating communication in low-resource language settings where traditional approaches struggle due to data scarcity and linguistic complexity. The paper examines methodologies for leveraging pre-trained multilingual models through fine-tuning and transfer learning to enhance translation accuracy, particularly by addressing challenges such as tonal variations in Chinese and syntactic simplicity in Malay. It also highlights emerging techniques like cross-lingual transfer learning, domain adaptation, and data augmentation, which showcase the potential to overcome resource constraints and improve output fluency. Key application areas, including education, tourism, healthcare, and e-commerce, illustrate the real-world impact of these systems in fostering intercultural communication. However, challenges such as dialectal diversity, subtle contextual nuances, and data imbalance persist, calling for further advancements in multimodal approaches and adaptive learning. Future research directions include developing more inclusive datasets, refining linguistic modeling for low-resource languages, and integrating multimodal translation techniques to bolster performance and applicability in diverse settings, ultimately advancing the field of multilingual speech translation.

Index Terms—multilingual translation models, cross-lingual transfer, speech-to-text adaptation

1 Introduction

The burgeoning necessity for effective communication between Chinese and Malay speakers accentuates the critical role of advanced speech translation systems. In an increasingly globalized world, where cultural and economic interactions rely on seamless communication, the development of a dedicated Chinese to Malay speech translation system is pivotal. Such systems enable real-time translation, enhancing understandings in diverse contexts such as education, tourism, and commerce. Existing statistics show that nearly 30

The linguistic transition from Chinese to Malay, however, is fraught with challenges stemming from the distinct phonetic and syntactical structures inherent to both languages. The tonal nature of Chinese significantly complicates the translation process, as tonal distinctions can alter meanings entirely, presenting a unique array of hurdles [1]. Moreover, Malay syntax, which often employs simpler grammatical constructions, introduces additional complexities when applying standard translation algorithms tailored primarily to tonal languages [2]. Therefore, an understanding of these differences is essential for developing models that can accurately reflect meaning and intent across languages.

Prominent methodologies in contemporary translation systems hinge on the deployment of large, pre-trained multilingual models that utilize extensive training on diversified datasets [3]. These systems simplify the translation process by allowing knowledge transfer between languages, essentially enabling zero-shot tasks where the model translates languages without explicit prior training on those pairs—an area that has gained traction in recent literature [4]. For instance, the use of shared embeddings across languages has demonstrated empirical benefits by leveraging similarities that reduce computational overhead while maintaining high accuracy, echoing performance enhancements seen in studies comparing bilingual and multilingual frameworks [5].

Nonetheless, the application of advanced models is not without limitations. Issues such as data scarcity, especially for low-resource languages like Malay, can affect the robustness of translations. Furthermore, the interaction of varying dialects and registers can lead to biases in translation quality, often disadvantaging less-represented language pairs [6]. In navigating these multifaceted challenges, future research directions should focus on the integration of adversarial training and adaptive mechanisms that account for linguistic diversity, thus improving output quality. Enhancements will not only serve pragmatic translation functions but facilitate deeper cultural exchange. Consequently, the evolution of Chinese to Malay speech translation models stands to redefine communication paradigms, reflecting the evolving landscape of multilingual interactions both locally and globally.

2 LARGE MULTILINGUAL TRANSLATION MODELS

2.1 Architectural Foundations of Multilingual Translation Models

The architectural foundations of large multilingual translation models are primarily built on transformer architectures, leveraging their capacity for parallelization and efficiency in handling diverse language pairs. The transformer model, with its encoder-decoder structure and self-attention mechanisms, allows for better context understanding and translation accuracy across languages. The introduction of attention heads enables models to focus on various positions in a source sentence, effectively capturing long-range dependencies in languages with complex syntactic structures [2].

In the context of multilingual translation, the ability to share parameters across languages presents a distinct advantage. Multilingual models like those discussed in [7] utilize a shared embedding layer to process multiple languages, promoting knowledge transfer and reducing training complexity. However, the challenge lies in the potential trade-off between performance and the risk of negative interference among languages, particularly when dealing with typologically distinct language pairs [5]. This is evidenced by findings that indicate shared parameters can lead to subpar performance if not carefully managed. For instance, [5] underscores that the full parameter sharing approach can enhance BLEU scores primarily when the involved languages belong to similar families, highlighting the importance of language similarity in design choices.

Recent advancements have introduced techniques such as Mixture of Experts (MoE), which dynamically activate subsets of model parameters tailored for specific tasks or languages. This approach mitigates the limitations of fixed parameter sharing while optimizing computational resources. Research indicates that MoE frameworks yield significant improvements in translation quality by allowing models to leverage a focused set of parameters for particular languages, thereby increasing both efficiency and effectiveness [8].

Furthermore, the growing interest in Neural Architecture Search (NAS) methodologies facilitates the exploration of optimal architectural combinations that enhance multilingual task performance. By evaluating diverse configurations, researchers can identify structures that maximize model efficacy while addressing the unique features inherent in each language. This aligns with findings from [3], where empirical analyses reveal that thoughtful architectural adjustments are crucial for modeling low-resource languages.

In summary, the architectural foundations of multilingual translation models hinge upon the judicious integration of transformers, parameter sharing mechanisms, expert systems, and innovative architectural search techniques. As translation requirements evolve, further exploration of hybrid architectures combining diverse approaches may unlock new efficiencies and performance enhancements. Emphasizing language diversity in training and evaluation will be pivotal in overcoming the challenges associated with multilingual translation, propelling the development of truly universal models adept at bridging linguistic divides.

2.2 Pre-training and Fine-tuning Methodologies

Pre-training and fine-tuning methodologies are pivotal in enhancing the performance of multilingual translation models, particularly in bridging the linguistic gap between low-resource languages like Malay and high-resource languages such as Chinese. The predominant approach to pre-training involves leveraging massive multilingual datasets, allowing models to learn and capture diverse linguistic structures and semantic representations. This foundational learning process is crucial for identifying cross-linguistic patterns and information, enabling the generation of coherent translations across multiple language pairs. A prime example of this strategy is mBART, which utilizes large quantities of both monolingual and parallel data to enhance translation capabilities in resource-constrained settings [9].

Subsequent fine-tuning adapts these pre-trained models to specific translation tasks. This stage often employs supervised learning on task-specific datasets, fostering a greater sensitivity to the unique characteristics of the target languages. For instance, fine-tuning a model on extensive Chinese-Malay parallel data can significantly improve its understanding of idiomatic expressions and contextual nuances peculiar to this language pair [3]. The dual-stage approach of pre-training followed by fine-tuning thus represents a robust strategy to achieve state-of-the-art performance, with empirical results demonstrating substantial gains in both translation accuracy and fluency.

Recent advancements in self-supervised learning have further enriched fine-tuning methodologies. Techniques such as denoising autoencoders and contrastive learning capitalize on large amounts of unannotated data for pre-training, reducing reliance on costly labeled datasets while still refining the model's translation performance [10]. This trend is particularly beneficial for languages with limited resources, as it facilitates the leveraging of monolingual data from related languages to enhance overall translation quality.

Nonetheless, several challenges persist within this framework. The intricate balance between model complexity and computational efficiency often raises concerns about overfitting, particularly when fine-tuning on smaller datasets. Additionally, the "curse of multilinguality" can result in negative interference, which can impede performance across different languages due to the shared parameters used among diverse inputs [11]. To address these issues, adaptations such as lightweight adapters and language-specific sub-networks have emerged as promising solutions, optimizing resource utilization while mitigating risks associated with overfitting [12], [13].

As these methodologies continue to evolve, future research should explore enhanced integration of multi-modal learning. This approach could facilitate better contextual understanding by incorporating audio and visual data alongside text-based training, significantly improving translation efficacy in real-world applications. Moreover, the development of broader datasets that encapsulate cultural and dialectical variations is essential for refining the nuance and precision of translations. Collectively, these insights underscore the necessity for ongoing advancements in pre-training and fine-tuning frameworks within multilingual contexts.

2.3 Data Utilization and Augmentation Techniques

The efficacy of large multilingual translation models heavily relies on the quality, quantity, and diversity of training data. In particular, the challenge of data scarcity is pronounced for low-resource languages like Malay, where the availability of high-quality parallel corpora is limited. This necessitates innovative approaches to optimize data utilization and augmentation techniques.

One prominent strategy involves the construction and curation of multilingual datasets that leverage various sources of data, such as web-mined content, community contributions, and existing translation memory systems. Techniques such as cross-lingual data leveraging allow models trained on high-resource language pairs, like Chinese and English, to enhance translation capabilities for low-resource languages by enabling them to learn from indirectly related languages. For instance, pivot-based strategies have shown promise in improving source-target translations by utilizing intermediary languages, thus mitigating performance hurdles observed in direct translation tasks, as highlighted by the findings in [14].

Data augmentation techniques play a critical role in expanding training datasets. Two common methods include back-translation and paraphrasing. Back-translation generates additional parallel data by translating target language outputs back into the source language, creating a cycle that amplifies training resources. Empirical studies have demonstrated that back-translation can substantially boost translation performance across numerous language pairs, effectively enriching the learning environment for machine translation models [15]. Additionally, paraphrasing generates semantically similar sentences, thereby enhancing model robustness through exposure to diverse linguistic structures [16].

The integration of synthetic data further complements existing datasets, particularly in the context of low-resource languages where actual speech data is scarce. Synthetic data generation approaches, such as text-to-speech (TTS) systems that convert text into audio formats, can expand the training data pool considerably. Studies have illustrated the positive impact of synthetic data on improving model performance in speech translation tasks, allowing systems to generalize better in real-world applications [17].

However, while these techniques effectively mitigate data scarcity issues, they come with trade-offs. For instance, incorporating synthetic data without rigorous quality control may lead to the development of biased models, inadvertently reinforcing existing errors in translation, as observed in models trained with poor-quality augmentations [18]. Furthermore, the risks associated with overfitting to artificially generated instances underscore the need for careful monitoring and validation.

As we look to the future, it is essential to refine these data utilization methods by integrating more advanced self-supervised learning techniques. Such strategies have been shown to leverage vast amounts of unannotated data effectively, thus expanding the datasets without explicit manual intervention [19]. This evolving landscape promises enhanced performance for multilingual models, particularly as researchers increasingly explore collaboration across disciplines to cultivate comprehensive datasets that reflect the rich complexity of human language. By continuing to innovate in data utilization and augmentation practices, the multilingual translation community can tackle the ongoing challenges posed by resource constraints and strive towards achieving universal language translation capabilities.

2.4 Cross-Lingual Transfer Learning Approaches

Cross-lingual transfer learning approaches have emerged as pivotal mechanisms in enhancing translation tasks for low-resource language pairs, particularly within the framework of large multilingual models. These strategies leverage the wealth of resources available in high-resource languages to facilitate better performance in languages with limited training data. One effective methodology is pivot-based translation, where intermediary languages with abundant resources serve as bridges for translation tasks involving low-resource languages. For instance, utilizing a high-resource language such as English, models can translate to and from a pivot language before reaching the target low-resource language. This indirect transfer methodology enhances translation quality despite the absence of direct training data for that language pair, as evidenced by recent studies demonstrating significant performance gains in such low-resource settings [20].

In addition, parameter-efficient transfer techniques have gained traction, particularly with the introduction of adapters—small modules that can be integrated into pre-existing multilingual models. These adapters enable the selective adaptation of specific model layers to new languages while retaining the foundational knowledge acquired from high-resource counterparts. By allowing for efficient fine-tuning without the overhead of retraining entire models, adapters prove particularly beneficial when computational resources are limited. This targeted training approach demonstrates improved translation accuracy for low-resource languages [19].

Another critical aspect of cross-lingual transfer learning is zero-shot translation. In these scenarios, models utilize shared representations learned during training across multiple languages to perform translations in language pairs they have never explicitly encountered. This capability is made possible through shared embedding spaces within multilingual models, which foster a generalized understanding across diverse linguistic structures. Recent findings indicate that zero-shot performance can be significantly enhanced by pre-training with extensive multilingual datasets that enforce consistent semantic relations across languages [21].

Despite the promise that these approaches hold, several challenges persist regarding data imbalance and the quality of available resources. Structural and syntactic differences among many high-resource languages can complicate the transfer of linguistic patterns to low-resource languages. Additionally, maintaining the integrity of language-specific idiomatic expressions and cultural nuances during transfer learning processes remains an ongoing challenge. However, emerging methods that incorporate self-supervised learning and back-translation techniques show potential for mitigating these limitations, enabling models to generate pseudo-parallel data that enrich training datasets with contextual relevance [22].

In summary, while cross-lingual transfer learning presents a robust framework for enhancing low-resource language translations, it is imperative that the ongoing evolution of methodologies—such as pivot methods, adapters, and zero-shot capabilities—continues to adapt to the intricacies of linguistic diversity and resource constraints. Future research directions will likely focus on the integration of multimodal data, further enhancing the interactive capabilities of multilingual models, and refining techniques to generate diverse linguistic representations that respect the nuances of both high- and low-resource languages.

2.5 Impact of Multilingual Datasets on Model Performance

The impact of multilingual datasets on the performance of translation models is multifaceted, encompassing the richness of linguistic variety as well as the inherent challenges posed by data imbalances and quality disparities. Large-scale multilingual corpora are critical for training robust translation systems as they enable models to learn intricate language features and grammar rules across diverse languages, thus facilitating better generalization to low-resource scenarios. For instance, the exposure of a model to a variety of languages aids in the acquisition of shared representations, which can significantly enhance the model's zero-shot translation capabilities, where translations are generated for unseen language pairs [23].

However, the effectiveness of multilingual datasets is contingent upon their composition and quality. Data scarcity in certain languages can lead to what is known as "data imbalance," where certain language pairs are disproportionately represented. This imbalance can degrade translation quality for underrepresented languages due to the model's tendency to favor the majority languages during the training process [24]. Research indicates that when models are trained on unbalanced datasets, performance can significantly diminish for low-resource languages, highlighting the necessity for innovative data augmentation techniques such as back-translation or synthetic data generation [25].

Furthermore, the quality of the multilingual dataset plays a vital role; noisy or poorly translated data can introduce errors that propagate through the training phase, adversely affecting model performance []. Effective training strategies, such as leveraging high-quality parallel corpora or utilizing robust preprocessing techniques, can mitigate these issues. For example, methods that incorporate adversarial training or regularization techniques can help align the source and target distributions, thus improving translation accuracy across languages [27].

Emerging trends indicate a shift towards integrating self-supervised learning approaches with existing multilingual datasets to enhance performance in low-resource settings. The use of unlabeled data to supplement training can facilitate better representation learning, thereby improving the model's capabilities for languages with limited paired data [28]. Additionally, recent explorations into data representation and alignment suggest that enhancing cross-lingual structures within pretrained models can further empower models to yield high-quality outputs across various translations [29].

In conclusion, while multilingual datasets present substantial opportunities for improving translation model performance through diverse linguistic exposure, inherent challenges such as data imbalance, quality issues, and the complexities of cross-language alignment require ongoing attention. Future research should prioritize developing methodologies that capitalize on the strengths of multilingual datasets while addressing their limitations, potentially leveraging hybrid approaches that combine supervised and unsupervised learning to enrich the performance of translation systems across a broader spectrum of languages.

2.6 Future Directions in Multilingual Translation

The landscape of multilingual translation systems is rapidly evolving, propelled by advancements in large multilingual models and enhanced computational resources. A pivotal trend within this evolution is the integration of multimodal inputs, enabling models to utilize not just text but also speech and images to enrich contextual understanding and improve translation tasks. Recent studies have demonstrated that combining speech-to-text recognition with translation processes can enhance overall accuracy and fluency, particularly in complex or ambiguous contexts, ultimately delivering a superior user experience [30].

Another promising direction involves refining cross-lingual alignment through fine-tuning multilingual models on task-specific datasets. This targeted approach often results in significantly improved performance for low-resource languages. The introduction of language adapters—lightweight modules that facilitate fine-tuning of specific layers to accommodate language nuances without extensive retraining—exemplifies this strategy. These methods have shown effective outcomes in tailoring models to various dialects or accents while preserving foundational multilingual capabilities and minimizing the overhead associated with large-scale re-training [12].

Moreover, the exploration of ethical AI considerations is increasingly shaping the future of multilingual translation systems. Addressing biases inherent in training datasets is essential to ensure fair representation across languages, especially in low-resource environments. This effort includes the creation of diverse and high-quality multilingual datasets that accurately reflect various sociolinguistic contexts. Recent benchmarks, such as FLORES-101, provide frameworks for evaluating multilingual models across a broader linguistic spectrum, thereby emphasizing the necessity for more inclusive datasets to mitigate bias [31].

Additionally, the continued effectiveness of zero-shot and few-shot learning techniques presents an area of active research. These approaches empower models to transfer knowledge between languages without direct training examples, making them particularly valuable for under-represented languages. The performance of multilingual models under these conditions reveals substantial potential, enabling them to generalize effectively even in previously unseen language pairs [32]. Recent findings indicate that the capabilities of these models can be significantly enhanced through the thoughtful selection of multilingual instruction during fine-tuning processes [33].

Lastly, the rise of large, openly available multilingual datasets is catalyzing significant advancements in this field. The synthesis of data from diverse sources fosters models that better comprehend the nuanced interplay between languages. However, challenges persist in managing the volume and variety of data to prevent compromising quality in favor of quantity [6]. Striking a balance in dataset preparation, model architecture design, and training methodologies presents an exciting paradigm for the development of future multilingual translation systems.

In conclusion, the convergence of multimodal data integration, cross-lingual transfer learning, ethical considerations, and innovative learning techniques represents promising pathways for the enhancement of multilingual translation systems. Ongoing research in these areas is anticipated to foster significant progress, addressing the challenges associated with low-resource languages and improving overall translation quality in real-world applications.

3 Fine-Tuning Techniques for Speech Translation

3.1 Transfer Learning for Low-Resource Scenarios

Transfer learning plays a pivotal role in the development of speech translation systems, particularly in low-resource scenarios such as translating from Chinese to Malay. Given the scarcity of high-quality parallel corpus for the Malay language, leveraging knowledge from high-resource languages emerges as a compelling strategy to enhance translation performance. This subsection explores several methodologies including transfer learning, parameter-efficient approaches, and knowledge distillation that capitalize on existing models and datasets to improve the efficacy of translation tasks.

One prominent approach is the adaptation of models pre-trained on high-resource languages, such as English, to the task of translating low-resource languages. By fine-tuning these models on limited annotated data from the target language, researchers can retain substantial generalization capabilities while addressing linguistic idiosyncrasies specific to Malay. For example, previous work demonstrates that fine-tuning a transformer architecture on a smaller Malay dataset can yield significant improvements in translation fluency and accuracy, effectively adapting the model's previously acquired knowledge to the new domain [34].

Parameter-efficient techniques, such as adapter layers, offer an alternative that facilitates targeted fine-tuning without extensive retraining of the entire model. By introducing lightweight layers into a pre-trained architecture, only specific regions of the model are adjusted during training, significantly reducing computational overhead while maintaining overall model integrity. This approach has been shown to allow for specific adaptations with minimal loss of performance across primary tasks, showing promise in scenarios with limited computational resources [7].

Knowledge distillation is another effective method in improving the robustness of translation models for low-resource languages. In this setting, a large teacher model trained on ample data from high-resource languages informs the training of a smaller student model tailored for low-resourced ones. Through this process, the student learns to mimic the teacher's outputs, preserving essential translation features while reducing the model's size and computational demands, making it particularly useful in mobile or low-power applications [35].

Despite these promising methodologies, challenges remain, particularly regarding data imbalance and the adaptation of translation models to capture the nuanced syntactic and semantic structures unique to each language pair. Empirical evidence suggests that while transfer learning can mitigate the issues related to resource scarcity, it necessitates careful selection of data and rigorous evaluation methods to ensure performance across unseen language pairs [1].

Emerging trends such as multilingual models capable of zero-shot translation also highlight the potential benefits of expansive data integration across various languages. These models leverage shared representations and language similarities to facilitate translations without direct training data for specific language pairs, showcasing a noteworthy shift toward a more inclusive approach to global language translation [23]. As the field evolves, continued exploration of these techniques, along with advances in data collection methods and model architectures, will be essential in further bridging the gap in translation quality for low-resource languages like Malay, ultimately enhancing cross-cultural communication and accessibility on a broader scale.

3.2 Domain Adaptation Techniques

Domain adaptation techniques are essential for fine-tuning multilingual models to handle specific speech translation tasks, particularly in the context of the diverse dialects, accents, and contextual language usage prevalent in Chinese and Malay. This subsection explores various methodologies that support multi-layered adaptation strategies, enhancing the robustness and fluidity of speech translation systems.

A key approach to domain adaptation is context-specific adaptation, where training data is tailored to reflect particular domains such as casual conversation, business terminology, or technical jargon. By incorporating these targeted datasets into the model training process, the multilingual system significantly improves its proficiency in capturing domain-specific lexicon and colloquialisms that are crucial for effective communication [36]. Prior studies emphasize that utilizing domain-rich datasets closely aligns the model's semantic understanding with practical applications, facilitating smoother conversational flows and reducing translation errors.

In addition to contextual considerations, the importance of addressing accent and dialect variations cannot be overstated, as phonetic differences can greatly impact comprehension and translation accuracy. Techniques such as collecting diverse speech samples from various demographic groups enhance the model's sensitivity to regional accents and variations. This practice supports findings from [13], which highlights the necessity of training on diverse acoustic datasets to ensure that models are attuned to different pronunciation patterns. By implementing training strategies that aggregate this data, models can achieve a comprehensive representation of the language landscape, making them more adept at managing dialectical intricacies.

Another vital domain adaptation strategy involves the use of adaptive learning protocols, such as multi-task learning frameworks. In this scenario, the model is trained simultaneously on related tasks, including speech recognition alongside translation. This multi-task approach fosters a deeper integration of linguistic features across tasks and facilitates shared learning that captures both the nuances of spoken language and the dynamics of translation [37]. Recent studies advocate for augmenting these multi-task systems with specialized attention mechanisms that dynamically allocate resources, enriching contextual understanding when faced with varied input formats.

Despite the advancements in domain adaptation, challenges remain, particularly for low-resource language pairs. Variability in data availability across dialects and contexts can lead to issues such as overfitting or inadequately representative models when trained on insufficiently diverse datasets. One promising solution to mitigate these challenges is adversarial training, where perturbed inputs are introduced to test the model's generalization capabilities [38]. This framework encourages models to maintain robustness amid fluctuations in input characteristics, thus enhancing their adaptability across various dialects.

Looking ahead, emerging trends are increasingly focused on integrating synthetic data generation techniques, including adversarial examples and back-translation, into the domain adaptation process. These innovations hold the potential to create enriched training datasets, allowing models to learn from both real and simulated speech inputs. Future research may explore hybrid frameworks that blend supervised and unsupervised adaptation methods, enabling more efficient learning from smaller datasets while maintaining high levels of accuracy across diverse speech contexts. This direction could significantly address current limitations and bolster the performance of Chinese to Malay speech translations, expanding their applicability in real-world scenarios.

3.3 Data Augmentation Strategies

Data augmentation has emerged as a pivotal strategy in enhancing training datasets for low-resource languages, particularly for underrepresented language pairs such as Chinese and Malay. The inherent scarcity of high-quality parallel corpora necessitates innovative approaches to generate supplementary data that can significantly improve model performance. This subsection delves into a variety of data augmentation techniques, evaluating their strengths, limitations, and the practical implications for speech translation systems.

One prevalent strategy is synthetic data generation through back-translation. This method involves translating a sentence from the target language (Malay) back into the source language (Chinese) and leveraging the resulting bilingual pairs as additional training samples. Studies, including those by [15], have shown that back-translation can enhance translation quality across various languages by increasing the diversity and volume of training data. However, the effectiveness of this strategy often hinges on the quality of the translation models employed, which can lead to the introduction of noise into the dataset if not adequately managed.

Another innovative approach involves augmentative noise injection, where various types of noise—like distortions or interruptions in the speech signal—are systematically introduced into the training data. This technique helps train models to be robust against real-world conditions where audio quality may be compromised. The incorporation of noise aims to mimic natural variances in speech, enhancing the adaptability of models in diverse environments. The asymmetrical benefits of this approach have been highlighted in papers like [39], where it was demonstrated that noise-enhanced datasets can improve model generalization without extensive labeled data.

Additionally, speech synthesis techniques utilizing text-to-speech (TTS) systems have added another dimension to data augmentation. By generating synthetic audio recordings from textual data, TTS can create new training examples, particularly important in cases where original audio data is limited. The effectiveness of TTS as a data augmentation tool has been supported by results from [40], which emphasized the potential of combining various modalities for comprehensive training solutions.

Furthermore, recent advancements leverage multimodal transfer techniques, where models are trained to utilize both text and speech data in parallel, enhancing the overall learning process. This approach not only elevates data efficiency but also bridges the gap between textual and auditory comprehension. New frameworks articulated in [41] propose integrating diverse language inputs to create robust multilingual models, emphasizing that the model's ability to handle cross-modal data can significantly propel low-resource translations beyond traditional limitations.

Looking ahead, while current augmentation strategies provide considerable improvements, challenges remain in ensuring the scalability of these techniques across an ever-expanding array of languages and dialects. Future research should focus on developing more sophisticated augmentation methods that consider linguistic nuances, incorporating contextual elements from both source and target languages to achieve a higher fidelity in generated data. Leveraging emerging technologies such as advanced generative models may offer promising avenues for addressing these challenges, potentially enabling the creation of tailored datasets that reflect the intricate dynamics of speech translation between Chinese and Malay.

3.4 Fine-Tuning and Hyperparameter Optimization

The process of fine-tuning large multilingual models for speech translation, particularly in translating from Chinese to Malay, requires a meticulous approach to hyperparameter optimization. The performance of neural machine translation (NMT) models is significantly influenced by the selection of hyperparameters, including learning rates, batch sizes, and model architectures. A well-structured fine-tuning strategy not only allows the model to adapt to the unique characteristics of the target language but also ensures that it retains valuable knowledge acquired during pre-training. For example, careful tuning of the learning rate can lead to improved convergence and enhanced translation accuracy, where the optimal range often necessitates empirical validation through techniques such as grid or random search [22].

While traditional optimization methods typically involve exhaustive searches, recent trends have shifted toward more efficient techniques like Bayesian optimization and Hyperband. These methods are designed to dynamically allocate resources to the most promising hyperparameter configurations, thereby optimizing performance even within limited computational resources [42]. Furthermore, the careful adjustment of dropout rates is critical to preventing overfitting, especially in low-resource settings where the risk of learning noise from limited training data is prevalent [25].

An emerging trend within the literature is the integration of continuous learning with hyperparameter optimization. This approach allows models to incrementally adapt to new data, minimizing catastrophic forgetting while enhancing performance on specific tasks. Notably, frameworks like elastic weight consolidation can be advantageous in maintaining robustness against sudden shifts in task distributions, ensuring that multilingual models do not suffer from performance degradation [43].

However, hyperparameter optimization presents its own challenges and trade-offs. While extensive searches for optimal configurations can lead to notable improvements in translation quality, they often require significant computational resources and time, which may not always be feasible, especially for low-resource languages like Malay [44]. Moreover, integrating techniques such as knowledge distillation can help streamline model size and complexity, but may introduce latency during deployment, potentially impacting the effectiveness of real-time translation applications [45].

Empirical studies have demonstrated that fine-tuning on in-domain datasets yields better performance than using generic datasets, underscoring the importance of domain adaptation in this context [15]. From a synthesis perspective, future research could explore the role of transfer learning from high-resource languages to inform hyperparameter settings for low-resource languages. Such endeavors may yield insights that help bridge gaps in resource availability, aligning well with efforts to foster inclusivity in multilingual NMT systems and enhancing communication efficacy in diverse sociolinguistic contexts [46].

In conclusion, the integration of sophisticated fine-tuning methodologies and hyperparameter optimization strategies will be pivotal in maximizing the effectiveness of large multilingual models. As the field of multilingual speech translation continues to evolve, such innovations are likely to reshape the landscape of NMT for low-resource languages, ultimately fostering enhanced global interoperability and understanding.

3.5 Benchmark Datasets and Evaluation Techniques

Evaluating speech translation systems necessitates the establishment of robust benchmark datasets and evaluation metrics to ensure consistent assessment of model performance across various contexts and applications. This subsection delves into the critical frameworks and methodologies that underpin the evaluation of Chinese to Malay speech translation systems, highlighting the interplay between dataset selection and evaluation criteria.

The efficacy of a speech translation model is often anchored in its training data, necessitating a diverse set of benchmark datasets that encapsulate various linguistic features and contextual nuances of both source (Chinese) and target (Malay) languages. Datasets like the Common Voice repository and the MultiLingual Speech Dataset have proven instrumental by providing multilingual audio corpora, which not only enhance model robustness through varied phonetic representations but also introduce challenges related to dialectal differences and accent variability [47]. Effective utilization of such datasets helps mitigate common pitfalls associated with underrepresented linguistic phenomena in low-resource language pairs.

Evaluation metrics fall into two primary categories: traditional metrics derived from statistical machine translation (SMT) and metrics specifically tailored for speech translation tasks. Metrics such as BLEU and METEOR have been widespread in the assessment of translation quality due to their capability to quantify n-gram overlaps and semantic relevance. Nonetheless, these metrics can be insufficient in capturing fluency and contextual appropriateness, particularly in spoken language translation contexts [48]. For this reason, specialized metrics like Word Error Rate (WER) and Character Error Rate (CER) have emerged as indispensable for evaluating output fidelity in speech tasks, offering insights into phonetic accuracy rather than lexical precision alone [49].

Recent advancements also indicate the value of human evaluation methodologies, which can complement automated metrics by providing qualitative insights into translation adequacy, naturalness, and overall user experience [50]. These approaches acknowledge the subjective aspects of translation that automated evaluations often miss, thus advocating for a hybrid evaluation framework that integrates both qualitative and quantitative measures.

Emerging trends in evaluation methodologies are notably shaped by the need to adapt to rapid developments in multilingual models and their capabilities in zero-shot and few-shot learning contexts [32]. The introduction of benchmarks that assess cross-lingual performance across diverse tasks, including those that specifically target low-resource languages, indicates a shift towards more holistic evaluation practices. Models that can effectively utilize synthetic data or are fine-tuned on small parallel corpora have demonstrated substantial improvements in zero-shot settings, underscoring the continuing relevance of efficient data utilization [25].

In conclusion, the evolution of benchmark datasets and evaluation techniques presents a pathway for future research to focus on refining evaluation frameworks that leverage both quantitative metrics and qualitative insights. Such a multidisciplinary strategy could enhance the accuracy, fluency, and contextual appropriateness of translation models, fostering improvements in real-world applications of speech translation systems across diverse language pairs.

4 Evaluation Metrics and Performance Assessment

4.1 Conventional Evaluation Metrics for Speech Translation

Conventional evaluation metrics play a critical role in assessing the performance of speech translation systems. Established metrics like BLEU, METEOR, and TER have become standard tools for evaluating translation quality by providing quantitative measures of accuracy, fluency, and overall effectiveness. The BLEU (Bilingual Evaluation Understudy) score, widely recognized for its simplicity and practicality, quantifies the overlap of n-grams between the translated output and reference translations. It is expressed mathematically as:

$$BLEU = BP \cdot \exp\left(\sum_{n=1}^{N} \frac{1}{N} \log P_n\right)$$

where BP stands for the brevity penalty, which addresses the issue of short translations, and P_n is the precision of n-grams. However, while BLEU remains a dominant metric, it has notable limitations, particularly its inability to account for semantic equivalence or synonyms, which can lead to an underestimation of translation quality for less direct equivalents [34].

In contrast, METEOR (Metric for Evaluation of Translation with Explicit ORdering) seeks to remedy these shortcomings by incorporating stemming and synonym matching, thus providing a more flexible evaluation framework. Designed to capture semantic nuances, METEOR calculates its score based on the alignment of words and considers grammar factors such as word order and semantic meaning, making it a preferred metric in situations where fluency and naturalness are vital [7].

Another widely utilized metric is the Translation Error Rate (TER), which measures the edit distance between the machine-translated text and the reference text. Defined as the proportion of edits required to change the machine output into one of the references, TER computes the necessary insertions, deletions, and substitutions. While TER provides valuable insights into the nature of errors, its primary drawback is that it may not effectively gauge fluency, particularly in translations that are grammatically correct but lack the desired stylistic attributes [51].

Emerging evaluation methods, like COMET (Contextualized Embedding Metric), leverage neural networks and pretrained models to predict semantic similarity, translating the qualitative aspects of translation into quantifiable scores. This innovation addresses the traditional limitations of n-gram-based metrics, offering a more robust evaluation that correlates well with human judgments on translation quality. COMET's performance suggests an important future direction for the field—incorporating more sophisticated metrics that can understand and evaluate the subtleties inherent in speech translation [52].

As the landscape of speech translation systems evolves, there is a pressing need to refine these evaluation metrics to better accommodate the intricacies of natural language processing tasks—especially in the context of low-resource languages like Malay, where traditional metrics may underperform. The future could benefit from hybrid evaluation approaches that merge automated metric scores with insights drawn from human evaluations to form a comprehensive framework for assessing translation efficacy. Such developments are crucial for establishing benchmarks that accurately reflect the capabilities of innovative multilingual translation models while ensuring alignment with user expectations and contextual appropriateness in real-world applications [53].

4.2 Human Evaluation and Qualitative Assessment

Human evaluation methodologies are essential for assessing translation quality, as they provide nuanced insights that automated metrics often overlook. While quantitative evaluations like BLEU and METEOR offer valuable measures of translation output, they fall short in capturing critical dimensions such as fluency, adequacy, and emotional resonance. In a multilingual context like Chinese to Malay, human assessments enhance our understanding of translation quality through expert and crowdsourced evaluations.

Expert assessments, conducted by professional linguists, meticulously evaluate translations and offer detailed critiques concerning faithfulness to the source material and situational appropriateness. This method typically results in high reliability in quality judgments since experts can discern subtle syntactical and cultural nuances that machines may misinterpret. For instance, professionals possess the ability to identify idiomatic expressions or culturally significant references often lost in automated translations. However, this approach may lack scalability due to the substantial time and resources required, which can constrain the breadth of testing across diverse translation contexts.

Conversely, crowdsourced evaluations leverage non-expert users who rate translation quality based on criteria such as clarity and naturalness. This democratic approach captures a broader spectrum of user experiences and perceptions; however, it introduces potential inconsistencies and biases, as evaluators might not possess the linguistic proficiency necessary for informed judgments. By combining insights from both qualitative methods, a more holistic understanding of translation quality can emerge. Recent innovations suggest integrating A/B testing, where users choose between multiple translations, to illuminate user preferences and highlight qualitative aspects not captured by single-score evaluations. Such a feedback loop can iteratively refine models to align more closely with user expectations.

Emerging trends in human evaluation increasingly integrate qualitative assessments with automated metrics to create a complementary evaluation framework. For example, new approaches are quantifying emotional resonance in translations, utilizing human judgments alongside computational measures to establish a more nuanced assessment matrix. Integrated frameworks, such as the COMET model, have demonstrated closer alignment with human evaluations by incorporating contextual embeddings into the assessment of translation quality, proving to be more effective than traditional metrics alone [54].

Despite these advancements, challenges persist—particularly concerning the scalability of human evaluation in real-time applications, such as live speech translation. Addressing these challenges may necessitate the exploration of hybrid evaluation methodologies that balance the strengths of both automated and human assessments while mitigating their weaknesses. As the field evolves, continuous feedback systems that incorporate user experiences will be crucial for maintaining and improving translation quality. Ultimately, embracing a multidimensional evaluation strategy will enhance the reliability of translation systems across diverse multilingual contexts, enabling more intuitive interactions between technology and its users.

4.3 Impact of Evaluation Metrics on Model Development

The choice and implementation of evaluation metrics play a pivotal role in shaping the development and optimization strategies of speech translation models. Metrics such as BLEU, TER, and COMET are traditionally employed to appraise translation quality. However, reliance on these metrics can significantly impact model architecture, training regimens, and ultimately, translation efficacy. For instance, BLEU, which calculates the n-gram overlap between machine-generated and reference translations, while widely used, has limitations in reflecting the fluency and contextual appropriateness of translations. Studies have shown that metrics like COMET, which leverage contextual embeddings to assess semantic similarity, correlate better with human judgments than BLEU and METEOR, suggesting a paradigm shift towards more semantically aware evaluations [16].

The implementation of these metrics influences critical design decisions in model development. For example, when the objective is to maximize BLEU scores, developers may prioritize precision over recall, leading to a potentially over-optimized model that compromises fluency. This trade-off can be particularly detrimental in speech translation, where maintaining the naturalness of spoken language is vital. Adapter tuning methods, which allow models to be specialized for specific tasks while retaining a core pre-trained architecture, have emerged as effective strategies that reflect this balance [12]. By optimizing hyperparameters to enhance particular evaluation outcomes, developers encounter the challenge of aligning training processes with desired metric scores, which may not always reflect user satisfaction or real-world applicability.

Moreover, the emergence of zero-shot translation capabilities—where models are evaluated on language pairs they have not specifically been trained on—has prompted the exploration of different evaluation methodologies. Recent advancements in multilingual NMT, which leverage shared representations across languages, emphasize the necessity for robust evaluation frameworks that can uniformly assess model performance across diverse linguistic contexts [3]. Such frameworks encourage innovation in model design by accounting for cross-lingual transfer effects, where knowledge gleaned from high-resource languages can bolster translation for low-resource counterparts.

As the community moves forward, the integration of human-in-the-loop evaluations alongside automated metrics may present a holistic approach for assessing translation models. Combining qualitative insights from user feedback with quantitative metrics can guide the refinement process, ensuring adaptations resonate with end-users. Future research should focus on developing unified evaluation metrics that encapsulate both linguistic accuracy and contextual relevance, thereby fostering innovations in model design that cater to the nuances of speech translation across languages. This dual focus will promote the creation of models that not only perform well against set benchmarks but also enhance communication across linguistic barriers effectively and naturally.

4.4 Contemporary Benchmarks and Emerging Trends

Recent advancements in speech translation evaluation benchmarks have dramatically transformed the landscape of machine translation research, particularly within low-resource language contexts such as Chinese to Malay. These benchmarks serve not only as quantitative measures of translation efficacy but also as catalysts for methodological innovations and enhancements in quality assessment practices. For instance, prominent endeavors like the CoVoST 2 dataset [55] provide expansive multilingual speech translation resources that encompass a variety of languages, emphasizing diverse accents and contexts. This rich variety is crucial for training robust models capable of accurately capturing linguistic variances encountered in real-world scenarios.

In parallel, the emergence of benchmarks focused on zero-shot and few-shot capabilities has highlighted the challenges associated with achieving performance in low-resource language tasks without extensive training data. These benchmarks significantly inform model adaptability and robustness, as evidenced by approaches that emphasize the role of transfer learning in cross-lingual scenarios [46]. Evaluating translation performance under these conditions necessitates nuanced metrics that encompass both the precision of generated translations and their contextual appropriateness.

Moreover, the integration of human-in-the-loop evaluation methods innovates traditional assessment paradigms by facilitating continuous feedback mechanisms that weave qualitative insights with quantitative metrics [56]. This paradigm shift is pivotal in recognizing the inherent complexities of human language, where established automated evaluations, such as BLEU or METEOR scores, may fall short of capturing nuances reflecting fluency or emotional resonance. Empirical research indicates that human assessments often correlate poorly with automated metrics, leading to significant overhauls in evaluation strategies [33].

Emerging trends increasingly focus on integrating diverse data modalities, such as acoustic features alongside text, into translation models, fostering a more holistic approach to speech translation [57]. This integration aligns seamlessly with the progress towards multimodal models capable of interpreting contextual clues from various inputs, thus significantly enhancing the model's performance in dynamic environments.

As evaluation benchmarks continue to evolve, addressing existing challenges—such as model biases in language representation and the resource constraints prevalent in low-resource settings—becomes increasingly pressing. New strategies for leveraging synthetic data, back-translation, and cross-lingual embeddings are emerging, indicating a potential shift in focus towards increasing dataset diversity and quality rather than sheer quantity [22], [58].

Looking ahead, embracing interdisciplinary insights and fostering collaboration across academic and industry sectors can unlock innovative evaluation methodologies that align with usability and user satisfaction standards. This progression is likely to delineate future pathways that enhance the adaptability and effectiveness of multilingual translation models while addressing inherent disparities across various languages. Such concerted efforts will ultimately pave the way for more equitable access to translation technologies, thereby improving communication barriers for languages that have, until now, been overlooked by technological advancements.

4.5 Challenges and Future Directions

Evaluating speech translation systems, specifically for Chinese to Malay, presents several challenges that necessitate innovative solutions and future research avenues. One of the primary challenges is the scarcity of high-quality, contextually relevant datasets for both languages, which significantly complicates performance measurement. High-resource languages have well-established corpora, while low-resource languages like Malay often lack sufficient aligned data for effective training and evaluation [24]. This disparity leads to evaluation methods that can be ineffectual; for instance, traditional metrics such as BLEU and METEOR, which rely heavily on textual fidelity, may not adequately capture the nuances of spoken language translations, particularly in informal settings [59].

Moreover, existing evaluation metrics often fail to address different dimensions of translation quality, such as contextual relevance and emotional resonance in dialogues [27]. While human evaluations are indispensable, they are limited by scalability issues and subjectivity []. Incorporating automated metrics augmented with machine learning models, such as COMET, could provide more nuanced insights by linking quality assessments to contextual embeddings from pre-trained models [51]. However, these rely on the availability of robust training data, which amplifies the importance of addressing resource scarcity.

Future directions for evaluation frameworks could focus on establishing comprehensive benchmarks that incorporate a wider array of performance indicators—beyond mere accuracy metrics—effectively capturing aspects like fluency and contextuality [60]. Developing adaptive evaluation systems that dynamically integrate user feedback could also refine the assessment processes, enhancing relevance to real-world applications [53].

Emerging trends such as the integration of cross-lingual capabilities with zero-shot translation methodologies present opportunities to augment evaluation. By leveraging frameworks like multilingual models that are trained to handle unseen language pairs, it is possible to enhance understanding of how different languages can infer meaning from one another without direct translation pairs [61]. This could lead to the establishment of innovative metrics that assess not only immediate translation accuracy but also the ability to maintain the semantic integrity across diverse language pairs.

The cross-disciplinary collaboration between linguists, data scientists, and computational linguists is critical in sculpting evaluation paradigms that accurately reflect the dimensionality and complexity of human communication [62]. By addressing these multifaceted challenges, the future of speech translation evaluation can evolve into a more robust and contextually aware model, fostering advancements in multilingual communication technologies.

5 Applications and Real-World Implementations

5.1 Speech Translation in Education

Speech translation systems that facilitate the transition from Chinese to Malay have emerged as pivotal tools in educational settings, enabling enhanced language acquisition and intercultural communication. These systems utilize state-of-the-art multilingual translation models which have been fine-tuned to cater to the specific needs of language learners and educators. By providing instantaneous translation, they create an immersive and interactive learning environment that fosters bilingual fluency among students.

In language acquisition, the significance of immediate feedback cannot be overstated. Students practicing their spoken language skills can receive real-time translations, allowing them to refine their pronunciation and grammatical structures. This process mirrors the findings reported in the literature regarding the efficacy of immediate corrective feedback in language education, highlighting improved outcomes in both spoken and written tasks [63]. Additionally, technology that supports language learning is shown to enhance motivation and engagement, which are critical factors in successful language acquisition [53].

Educators also benefit from these systems as they can develop bilingual teaching materials that integrate English, Malay, and Chinese, thereby accommodating diverse student populations. Such inclusive practices not only promote language skills but also cultivate cultural awareness and sensitivity, which are vital in today's globalized world. Importantly, multilingual speech translation tools allow teachers to engage non-Malay-speaking students effectively, ensuring they can participate in classroom interactions without language barriers [34].

However, the implementation of these systems is not without challenges. For instance, the quality of translations can vary depending on the complexity of phrases and contextual nuances, potentially leading to misunderstandings that hinder learning [3]. Furthermore, reliance on automated translation systems may result in oversimplified language understanding among learners if not accompanied by structured guidance from educators. The technical limitations present trade-offs that educators must navigate, weighing the facilitative benefits against the potential for incorrect or misleading translations.

Despite these challenges, there is a clear trajectory towards improved integration of speech translation technology in educational contexts. Emerging trends include the use of hybrid models that combine traditional language instruction with AI-driven translation tools to balance direct and indirect language learning methods [6]. Future research may focus on the development of context-aware translation models, which can adapt based on the specific educational circumstances and learner profiles.

In conclusion, Chinese to Malay speech translation systems hold considerable promise for revolutionizing educational practices, not only in language learning but also in fostering deeper intercultural understanding. Ongoing advancements in machine learning, particularly in multilingual models, suggest that continued improvements and broader acceptance of these technologies will further enhance their applicability in diverse educational environments, paving the way for innovative pedagogical strategies.

5.2 Applications in Tourism and Hospitality

The implementation of Chinese to Malay speech translation systems represents a significant advancement in the tourism and hospitality sector, facilitating seamless communication between Chinese-speaking tourists and Malay service providers. This dynamic interaction enriches the overall travel experience while fostering cross-cultural engagement. Historically, language barriers have hindered effective communication in multicultural environments, often resulting in misunderstandings and dissatisfaction among tourists. By harnessing speech translation technologies, service providers can offer real-time assistance that enhances customer service quality.

A notable application of these systems lies in tourist assistance. With real-time translation capabilities, tour guides and hotel staff can provide accurate information about local attractions, customs, and services, crucial for creating a welcoming atmosphere for visitors. For example, a translation system can swiftly convert inquiries from Chinese tourists about menu items or directions into Malay, facilitating smoother interactions in the hospitality context. Emerging studies indicate that multilingual translation systems significantly improve user satisfaction and operational efficiency in service-oriented environments, as highlighted by findings in [30].

Moreover, platforms that integrate advanced translation capabilities empower tourists to contribute user-generated content in their native languages. This involvement encompasses reviews and personal experiences shared across various online platforms, which can be translated into Malay to ensure broader accessibility. Such contributions not only guide future travelers in their decision-making processes but also enrich the cultural tapestry of local tourism industries. As noted in analyses [49], leveraging user-generated content can markedly enhance engagement and build trust in hospitality services.

However, the implementation of these systems does face challenges. Ensuring accuracy and contextual relevance in translated content remains a significant concern. Inaccuracies could lead to misinterpretations that impact tourist safety and satisfaction, underscoring the necessity for ongoing evaluation and refinement of translation models. Additionally, regional dialect variations within Malay present a challenge that translation systems need to address to ensure effective communication. Techniques such as language-specific tuning and data augmentation can significantly enhance translation effectiveness; as detailed in [5], the accuracy of translations often improves when accounting for localized dialects.

Looking ahead, there is an increasing necessity for adaptive systems that can learn from conversational exchanges and user feedback, thereby improving over time. Innovations in artificial intelligence and machine learning, including reinforcement learning, can support this adaptive learning model, equipping translation systems to become more contextually aware of cultural nuances and situational variances. Furthermore, as the global tourism landscape evolves, the integration of multilingual support in various applications—ranging from mobile tourism applications to interactive kiosks—will likely transform into a fundamental expectation, rather than an added value. Overall, the potential for Chinese to Malay speech translation systems to revolutionize the tourism and hospitality sector is immense, paving the way for more personalized and enriched experiences for all stakeholders involved.

5.3 Business and E-Commerce Applications

The application of Chinese to Malay speech translation systems has become increasingly vital in the realms of business and e-commerce, particularly as globalization intensifies and international markets become more interconnected. These translation systems facilitate effective communication, which is crucial for businesses aiming to penetrate Malay-speaking markets or engage in cross-border trade with Chinese firms. Speech translation technologies not only streamline negotiations but also bolster customer support, ultimately enhancing user experience and satisfaction.

The importance of real-time speech translation in business negotiations cannot be overstated. Effective dialogue between Chinese and Malay speakers can significantly reduce misunderstandings and foster trust in professional relations. Research shows that utilizing high-quality translation systems can overcome the common pitfalls associated with cultural and linguistic barriers [64]. For instance, the continual improvement of multilingual neural machine translation (NMT) models demonstrates their capacity to facilitate nuanced, context-rich conversations that traditional software often fails to capture [65]. Moreover, these models have been shown to simplify complex terminology, thus aiding in clearer communication during contract negotiations and business strategizing [66].

In e-commerce specifically, Chinese to Malay speech translation technologies empower businesses to effectively engage with a broader customer base. By enabling seamless integration of translation systems into e-commerce platforms, companies can provide voice-activated customer service options that cater to Malay-speaking consumers. This is especially applicable given the growing trend of voice commerce, where users prefer to interact using voice commands instead of conventional text-based methods. Enhanced user engagement through voice interfaces could potentially drive sales and improve overall customer satisfaction [17].

However, challenges remain, particularly around accuracy and contextuality in translations. The risk of errors in real-time applications can lead to significant misinterpretations, which can jeopardize business relations. Additionally, the computational resources required for advanced multilingual models can be daunting for smaller enterprises. Despite these challenges, the framework of adapter tuning has emerged as a promising approach, allowing businesses to customize large translation models efficiently without requiring extensive computational overhead [12]. This not only streamlines the translation process but also significantly reduces costs, making it more accessible for small to medium enterprises.

Emerging trends indicate that future innovations in translation technology will focus on enhancing the contextual understanding of phrases through better pre-training methodologies. Hybrid approaches that combine supervised and unsupervised methodologies could yield further improvements in translation quality, particularly for low-resource languages like Malay [19]. Such advancements may allow for more personalized translation experiences, further bolstered by user data analytics to tailor translations to specific business contexts.

As businesses increasingly rely on multilingual translation systems to navigate complex global markets, there is a significant impetus for ongoing research to address existing limitations. Important future directions include exploring efficient methods for low-resource language adaptation and developing systems that can dynamically learn from user interactions. In doing so, the ultimate goal should not only be improved translation quality but also fostering an inclusive business environment where diverse linguistic identities are respected and understood. The convergence of these innovations will likely define the next phase of global commerce, particularly within contexts characterized by rapid technological evolution and increasing linguistic diversity.

5.4 Healthcare Communication

The integration of speech translation systems in healthcare settings is pivotal for enhancing communication between healthcare providers and patients who speak different languages. These systems facilitate critical interactions during medical consultations, diagnoses, and treatments, often marked by complexities arising from linguistic barriers. Effective communication in healthcare is not merely an operational necessity; it is a fundamental aspect of patient care that profoundly influences medical outcomes, patient satisfaction, and overall healthcare efficiency.

Research demonstrates that speech translation systems can significantly streamline the process of gathering patient medical histories. For instance, real-time translation tools allow healthcare practitioners to pose vital questions and receive detailed responses from Chinese-speaking patients, which is essential for accurate diagnoses and personalized treatment plans. This capability relies on the precision of the speech recognition technology coupled with multilingual translation, as highlighted in the study "Improving Neural Machine Translation Models with Monolingual Data," which underscores the importance of leveraging high-quality, context-appropriate data to optimize fluency in translations relevant to healthcare vocabularies [67].

Despite these advancements, the efficacy of these systems is not without challenges. A primary concern lies in the accuracy of translations, particularly regarding technical medical terminology crucial for informed consent processes. If patients are unable to completely understand their treatment options due to translation errors, ethical dilemmas surrounding patient autonomy arise. This challenge is further complicated by the diverse dialects and specialized medical jargon present in the Chinese language, requiring advanced fine-tuning strategies for speech translation systems that can adapt to specific regional variations while maintaining contextual integrity [15], [22].

Emerging trends suggest a shift towards more sophisticated deep learning models that utilize extensive multilingual datasets to enhance translation accuracy in medical contexts. Techniques such as back-translation have demonstrated potential in enriching training datasets, particularly for low-resource languages, thereby providing a more robust foundation for the development of healthcare-specific translation models [58], [19]. Additionally, integrating feedback loops from real healthcare interactions can further improve model performance by capturing the nuances and terminology unique to medical dialogues.

Nevertheless, several challenges continue to impede progress. Ensuring data privacy and security in sensitive healthcare environments during the deployment of these systems poses a significant concern. Furthermore, continuous training and adaptation of models in response to evolving language in the medical field are imperative for sustaining reliability and effectiveness. A comprehensive strategy that fosters interdisciplinary collaboration among healthcare professionals, linguists, and AI technologists is crucial for the ongoing improvement and refinement of these systems.

Looking ahead, future research should explore the development of multimodal systems that incorporate visual aids alongside verbal translations. This approach could significantly enhance comprehension by providing contextual visuals that accompany translated instructions or descriptions. Initial studies indicate the promise of such integrations to enrich diagnostic discussions and treatment engagements. Customized applications tailored to the specific nuances of various healthcare contexts may emerge as key focal points for ongoing research and development in healthcare communication technologies [46].

By addressing these challenges and harnessing emerging technologies, speech translation systems have the potential to redefine patient-provider interactions, ultimately fostering more equitable healthcare delivery and improving patient outcomes across linguistic barriers.

5.5 Integration with Technology and Social Media

The intersection of Chinese to Malay speech translation systems with technology and social media platforms plays a pivotal role in democratizing communication and enhancing user engagement. By integrating these translation systems into widely-used applications like social media, messaging services, and content creation platforms, we can bridge linguistic barriers and foster more inclusive online environments. The advancements in natural language processing (NLP) fueled by large multilingual models enable real-time translation capabilities, allowing for fluid interactions between speakers of disparate languages. This is particularly relevant in social media, where the need for immediate, context-aware translations has never been greater for global discourse.

Technological integrations such as voice-activated tools and chatbots that utilize speech translation ensure that users can conduct multilingual interactions seamlessly. For instance, platforms like WeChat and WhatsApp can leverage speech translation services to allow users to send voice messages in their native language, which are then rendered into the recipient's language in real time. This functionality not only enhances personal communication but also allows businesses to engage with a broader audience, transcending geographical and linguistic boundaries. The work presented in [17] demonstrates that effective speech translation can be achieved with limited fine-tuning, improving translation fidelity and response times in applications that demand efficiency.

Moreover, the incorporation of speech translation into social media encourages user-generated content in multiple languages, promoting cultural exchange and broader participation. Users can comment, share posts, and create content in different languages without the fear of miscommunication. This integration is supported by advancements in machine learning that optimize translation accuracy even in noisy environments, as discussed in [25], demonstrating the robustness needed for practical applications in social media settings.

However, challenges remain. The use of translation systems must address issues related to linguistic diversity and regional dialects. Speech translation models often grapple with varying accents and cultural contexts, which can dilute the accuracy of translations [68]. Moreover, ethical considerations around the handling of user data and ensuring the privacy of conversations remain critical, particularly given the scale at which these technologies operate in social media environments.

Looking forward, emerging trends indicate a shift towards incorporating multimodal features—integrating text, voice, and visual elements seamlessly. This would enable more nuanced communication by factoring in contextual cues that enhance understanding and fluidity in interactions across language barriers.

In conclusion, the synergy between Chinese to Malay speech translation systems and technology facilitates not just enhanced communication, but also embodies a significant step towards a more interconnected global community. Continued innovations in this domain will require addressing the technical challenges of linguistic variation and maintaining user privacy, but the potential impact on multilingual engagement in digital spaces is immense.

5.6 Challenges and Future Directions

The implementation of Chinese to Malay speech translation systems encounters several significant challenges rooted in resource limitations and the intricate nature of language processing. A primary concern is the scarcity of high-quality parallel datasets necessary for training robust models. In contrast to languages with abundant resources, both Chinese and Malay face data limitations that hinder the development of effective speech translation applications. This scarcity necessitates the exploration of alternative data augmentation and synthesis techniques. Previous work has highlighted strategies such as back-translation and synthetic data generation, which can augment existing datasets and thereby contribute to improved model performance under low-resource conditions [22].

In addition to data scarcity, the fine-tuning of large multilingual models presents challenges regarding dialectal variation within Malay. The ability of speech translation systems to accurately distinguish between different dialects is crucial for achieving precise translations. Literature indicates that parameter-efficient methods, such as adapter tuning, could provide pathways to enhance model specificity without necessitating extensive retraining on large datasets, effectively balancing the trade-off between model performance and computational efficiency [12].

Emerging trends in cross-lingual transfer learning offer promising avenues for improving translation capabilities. By utilizing approaches that exploit similarities between high-resource and low-resource languages, particularly through zero-shot learning methods, the performance of translation systems can be significantly enhanced while reducing the data burden on less-represented language pairs [3]. Furthermore, advancements in multilingual instructional tuning suggest that models trained with a blend of languages can effectively improve translation capabilities across all involved languages, hinting at potential research directions that transcend traditional English-centric models [33].

Moreover, it is imperative to integrate ethical considerations into the development of these systems. Given the complex socio-cultural landscape within which these translations occur, an awareness of bias and the potential misrepresentation of languages due to model limitations is essential. Future efforts must emphasize the creation of fair and inclusive systems that fully account for the nuances of different dialects and sociolinguistic contexts.

In summary, addressing the challenges of resource scarcity, dialectal adaptation, and ethical considerations constitutes the foundational framework for future implementations of Chinese to Malay speech translation systems. By fostering continued innovation in data utilization, exploring cross-lingual approaches, and maintaining sensitivity to socio-cultural nuances, researchers can significantly enhance the efficacy and applicability of these systems in real-world settings. This, in turn, will support the overarching goal of promoting inclusivity and accuracy in multilingual communication, as highlighted in the previous discussions on technological integration and the evolving landscape of digital interaction.

6 Conclusion

The ongoing developments in Chinese to Malay speech translation systems underscore a pivotal juncture in multilingual communication, facilitated chiefly by advancements in large multilingual translation models. This survey has explored various methodologies, revealing that fine-tuning such models enhances translation accuracy and fluency significantly. The utilization of pre-trained multilingual models demonstrates an efficient pathway to robust language translation, particularly for low-resource language pairs such as Chinese and Malay, illustrating a distinct advantage in leveraging shared linguistic features via transfer learning methodologies, as evidenced in works such as [34] and [7].

Despite the promising landscape, notable challenges remain; issues of resource scarcity and linguistic diversity continue to pose significant barriers. The disparity in training data availability has been a major limitation in optimizing model performance across various dialects and accents, as outlined in [42]. Moreover, existing models often grapple with the intricacies of encoded grammar structures and tonal variations prevalent in the source languages. This gap necessitates further enhancement in model adaptability, ensuring that nuanced translations capture the contextual richness required for practical applications in diverse environments.

Emerging trends indicate a promising shift towards integrating multimodal approaches that incorporate speech-to-text, text-to-speech, and cross-modal translation capabilities, promoting seamless communication in contexts such as education and healthcare, as suggested by [69]. Furthermore, the prospect of utilizing synthetic datasets and advanced augmentation techniques is poised to mitigate the inadequacies stemming from limited parallel corpora, as demonstrated in [70].

Future research should focus on enhancing the interoperability of translation systems across a spectrum of dialectal expressions and idiomatic nuances. Building ethical frameworks for AI that promote inclusivity and reduce bias in translation will be essential, particularly as highlighted in [52]. Furthermore, dedicated efforts towards constructing larger and more diverse repositories of training data will likely mitigate the disparities faced by low-resource languages and further enable the accessibility of translation technologies across global contexts.

In summary, the potential of fine-tuned multilingual models to transform communication between Chinese and Malay speakers is undeniable. To capitalize on this momentum, the translation community must adopt comprehensive strategies that address existing limitations while fostering innovative methodologies for a future of enriched multilingual dialogue.

REFERENCES

[1] M. Sperber and M. Paulik, "Speech translation and the end-to-end promise: Taking stock of where we are," in *Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 7409–7421. 1, 5

[2] S. M. Lakew, M. Cettolo, and M. Federico, "A comparison of transformer and recurrent neural networks on multilingual neural machine translation," *ArXiv*, vol. abs/1806.06957, 2018. 1, 2

[3] "Edinburgh research explorer improving massively multilingual neural machine translation and zero-shot translation." 1, 2, 9, 10, 13

[4] X. Zhang, S. Li, B. Hauer, N. Shi, and G. Kondrak, "Don't trust chatgpt when your question is not in english: A study of multilingual abilities and types of llms," in *Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 7915–7927. 1

[5] D. S. Sachan and G. Neubig, "Parameter sharing methods for multilingual self-attentional translation models," in *Conference on Machine Translation*, 2018, pp. 261–271. 1, 2, 11

[6] A. Bapna, I. Caswell, J. Kreutzer, O. Firat, D. Esch, A. Siddhant, M. Niu, P. Baljekar, X. García, W. Macherey, T. Breiner, V. Axelrod, J. Riesa, Y. Cao, M. Chen, K. Macherey, M. Krikun, P. Wang, A. Gutkin, A. Shah, Y. Huang, Z. Chen, Y. Wu, and M. Hughes, "Building machine translation systems for the next thousand languages," *ArXiv*, vol. abs/2205.03983, 2022. 1, 5, 10

[7] X. Tan, Y. Ren, D. He, T. Qin, Z. Zhao, and T.-Y. Liu, "Multilingual neural machine translation with knowledge distillation," *ArXiv*, vol. abs/1902.10461, 2019. 2, 5, 8, 14

[8] N. Arivazhagan, A. Bapna, O. Firat, D. Lepikhin, M. Johnson, M. Krikun, M. Chen, Y. Cao, G. F. Foster, C. Cherry, W. Macherey, Z. Chen, and Y. Wu, "Massively multilingual neural machine translation in the wild: Findings and challenges," *ArXiv*, vol. abs/1907.05019, 2019. 2

[9] E. Lee, S. Thillainathan, S. Nayak, S. Nayak, S. Ranathunga, D. I. Adelani, R. Su, and A. D. McCarthy, "Pre-trained multilingual sequence-to-sequence models: A hope for low-resource language translation?" in *Findings*, 2022, pp. 58–67. 2

[10] A. Ustun, A. Berard, L. Besacier, and M. Gallé, "Multilingual unsupervised neural machine translation with denoising adapters," *ArXiv*, vol. abs/2110.10472, 2021. 2

[11] U. Shaham, M. Elbayad, V. Goswami, O. Levy, and S. Bhosale, "Causes and cures for interference in multilingual translation," *ArXiv*, vol. abs/2212.07530, 2022. 2

[12] H. Le, J. Pino, C. Wang, J. Gu, D. Schwab, and L. Besacier, "Lightweight adapter tuning for multilingual speech translation," *ArXiv*, vol. abs/2106.01463, 2021. 2, 4, 9, 12, 13

[13] Z. Lin, L. Wu, M. Wang, and L. Li, "Learning language specific sub-network for multilingual machine translation," *ArXiv*, vol. abs/2105.09259, 2021. 2, 5

[14] Y. Kim, P. Petrov, P. Petrushkov, S. Khadivi, and H. Ney, "Pivot-based transfer learning for neural machine translation between non-english languages," in *Conference on Empirical Methods in Natural Language Processing*, 2019, pp. 866–876. 3

[15] Z. Zhang, S. Liu, M. Li, M. Zhou, and E. Chen, "Joint training for neural machine translation models with monolingual data," *ArXiv*, vol. abs/1803.00353, 2018. 3, 6, 7, 12

[16] Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, and L. Zettlemoyer, "Multilingual denoising pre-training for neural machine translation," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 726–742, 2020. 3, 9

[17] X. Li, C. Wang, Y. Tang, C. Tran, Y. Tang, J. Pino, A. Baevski, A. Conneau, and M. Auli, "Multilingual speech translation from efficient finetuning of pretrained models," in *Annual Meeting of the Association for Computational Linguistics*, 2021, pp. 827–838. 3, 11, 13

[18] Y. Wang, C. Zhai, and H. Awadalla, "Multi-task learning for multilingual neural machine translation," *ArXiv*, vol. abs/2010.02523, 2020. 3

[19] A. Siddhant, A. Bapna, Y. Cao, O. Firat, M. Chen, S. Kudugunta, N. Arivazhagan, and Y. Wu, "Leveraging monolingual data with self-supervision for multilingual neural machine translation," in *Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 2827–2835. 3, 12

[20] M. Fang and T. Cohn, "Model transfer for tagging low-resource languages using a bilingual dictionary," *ArXiv*, vol. abs/1705.00424, 2017. 3

[21] X. Wang, S. Ruder, and G. Neubig, "Expanding pretrained models to thousands more languages via lexicon-based adaptation," in *Annual Meeting of the Association for Computational Linguistics*, 2022, pp. 863–877. 3

[22] F. Burlot and F. Yvon, "Using monolingual data in neural machine translation: a systematic study," *ArXiv*, vol. abs/1903.11437, 2018. 4, 6, 9, 12, 13

[23] M. Johnson, M. Schuster, Q. V. Le, M. Krikun, Y. Wu, Z. Chen, N. Thorat, F. Viégas, M. Wattenberg, G. S. Corrado, M. Hughes, and J. Dean, "Google's multilingual neural machine translation system: Enabling zero-shot translation," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 339–351, 2016. 4, 5

[24] B. Zoph, D. Yuret, J. May, and K. Knight, "Transfer learning for low-resource neural machine translation," *ArXiv*, vol. abs/1604.02201, 2016. 4, 10

[25] Y. Kim, Y. Gao, and H. Ney, "Effective cross-lingual transfer of neural machine translation models without shared vocabularies,"

in *Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 1246–1257. 4, 7, 13

[26] M. Costa-jussà, Q. CarlosA.Henríquez, and R. E. Banchs, “Evaluating indirect strategies for chinese - spanish statistical machine translation: Extended abstract,” *ArXiv*, vol. abs/1402.0563, 2012.

[27] Y. Yang, A. Eriguchi, A. Muzio, P. Tadepalli, S. Lee, and H. Hassan, “Improving multilingual translation by representation and gradient regularization,” *ArXiv*, vol. abs/2109.04778, 2021. 4, 10

[28] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, “Unsupervised cross-lingual representation learning at scale,” in *Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 8440–8451. 4

[29] N. Foroutan, M. Banaei, R. Lebret, A. Bosselut, and K. Aberer, “Discovering language-neutral sub-networks in multilingual language models,” in *Conference on Empirical Methods in Natural Language Processing*, 2022, pp. 7560–7575. 4

[30] H. Inaguma, K. Duh, T. Kawahara, and S. Watanabe, “Multilingual end-to-end speech translation,” *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 570–577, 2019. 4, 11

[31] N. Goyal, C. Gao, V. Chaudhary, P.-J. Chen, G. Wenzek, D. Ju, S. Krishnan, M. Ranzato, F. Guzmán, and A. Fan, “The flores-101 evaluation benchmark for low-resource and multilingual machine translation,” *Transactions of the Association for Computational Linguistics*, vol. 10, pp. 522–538, 2021. 4

[32] D. Wu, S. Tan, Y. Meng, D. Stap, and C. Monz, “How far can 100 samples go? unlocking overall zero-shot multilingual translation via tiny multi-parallel data,” *ArXiv*, vol. abs/2401.12413, 2024. 4, 7

[33] U. Shaham, J. Herzig, R. Aharoni, I. Szpektor, R. Tsarfaty, and M. Eyal, “Multilingual instruction tuning with just a pinch of multilinguality,” *ArXiv*, vol. abs/2401.01854, 2024. 5, 9, 13

[34] P. Nakov and H. T. Ng, “Improving statistical machine translation for a resource-poor language using related resource-rich languages,” *J. Artif. Intell. Res.*, vol. 44, pp. 179–222, 2012. 5, 8, 10, 14

[35] J. M. Eisenschlos, S. Ruder, P. Czapla, M. Kardas, S. Gugger, and J. Howard, “Multifit: Efficient multi-lingual language model fine-tuning,” in *Conference on Empirical Methods in Natural Language Processing*, 2019, pp. 5701–5706. 5

[36] Y. Tang, C. Tran, X. Li, P.-J. Chen, N. Goyal, V. Chaudhary, J. Gu, and A. Fan, “Multilingual translation with extensible multilingual pretraining and finetuning,” *ArXiv*, vol. abs/2008.00401, 2020. 5

[37] G. W. Blackwood, M. Ballesteros, and T. Ward, “Multilingual neural machine translation with task-specific attention,” *ArXiv*, vol. abs/1806.03280, 2018. 6

[38] T. Pires, R. M. Schmidt, Y.-H. Liao, and S. Peitz, “Learning language-specific layers for multilingual machine translation,” *ArXiv*, vol. abs/2305.02665, 2023. 6

[39] A. Rouditchenko, S. Khurana, S. Thomas, R. Feris, L. Karlinsky, H. Kuehne, D. F. Harwath, B. Kingsbury, and J. Glass, “Comparison of multilingual self-supervised and weakly-supervised speech pre-training for adaptation to unseen languages,” in *Interspeech*, 2023, pp. 2268–2272. 6

[40] N.-Q. Pham, A. Waibel, and J. Niehues, “Adaptive multilingual speech recognition with pretrained models,” in *Interspeech*, 2022, pp. 3879–3883. 6

[41] X. Li, Y. Yao, X. Jiang, X. Fang, C. Wang, X. Liu, Z. Wang, Y. Zhao, X. Wang, Y. Huang, S. Song, Y. Li, Z. Zhang, B. Zhao, A. Sun, Y. Wang, Z. He, Z. Wang, X. Li, and T. Huang, “Tele-flm technical report,” *ArXiv*, vol. abs/2404.16645, 2024. 6

[42] X. Wang, Y. Tsvetkov, and G. Neubig, “Balancing training for multilingual neural machine translation,” *ArXiv*, vol. abs/2004.06748, 2020. 7, 14

[43] R. Östling and J. Tiedemann, “Neural machine translation for low-resource languages,” *ArXiv*, vol. abs/1708.05729, 2017. 7

[44] B. Haddow, R. Bawden, A. V. M. Barone, J. Helcl, and A. Birch, “Survey of low-resource machine translation,” *Computational Linguistics*, vol. 48, pp. 673–732, 2021. 7

[45] F. Stahlberg, J. Cross, and V. Stoyanov, “Simple fusion: Return of the language model,” *ArXiv*, vol. abs/1809.00125, 2018. 7

[46] X. Chen, A. H. Awadallah, H. Hassan, W. Wang, and C. Cardie, “Multi-source cross-lingual model transfer: Learning what to share,” *ArXiv*, vol. abs/1810.03552, 2018. 7, 9, 12

[47] H. Yadav and S. Sitaram, “A survey of multilingual models for automatic speech recognition,” in *International Conference on Language Resources and Evaluation*, 2022, pp. 5071–5079. 7

[48] A. Fan, S. Bhosale, H. Schwenk, Z. Ma, A. El-Kishky, S. Goyal, M. Baines, O. Çelebi, G. Wenzek, V. Chaudhary, N. Goyal, T. Birch, V. Liptchinsky, S. Edunov, E. Grave, M. Auli, and A. Joulin, “Beyond english-centric multilingual machine translation,” *J. Mach. Learn. Res.*, vol. 22, pp. 107:1–107:48, 2020. 7

[49] C. Baziotis, M. Artetxe, J. Cross, and S. Bhosale, “Multilingual machine translation with hyper-adapters,” in *Conference on Empirical Methods in Natural Language Processing*, 2022, pp. 1170–1185. 7, 11

[50] A. Ebrahimi and K. Kann, “How to adapt your pretrained multilingual model to 1600 languages,” *ArXiv*, vol. abs/2106.02124, 2021. 7

[51] A. Pu, H. W. Chung, A. P. Parikh, S. Gehrmann, and T. Sellam, “Learning compact metrics for mt,” in *Conference on Empirical Methods in Natural Language Processing*, 2021, pp. 751–762. 8, 10

[52] N. M. Guerreiro, D. M. Alves, J. Waldendorf, B. Haddow, A. Birch, P. Colombo, and A. Martins, “Hallucinations in large multilingual translation models,” *Transactions of the Association for Computational Linguistics*, vol. 11, pp. 1500–1517, 2023. 8, 14

[53] A. Siddhant, A. Bapna, O. Firat, Y. Cao, M. Chen, I. Caswell, and X. García, “Towards the next 1000 languages in multilingual machine translation: Exploring the synergy between supervised and self-supervised learning,” *ArXiv*, vol. abs/2201.03110, 2022. 8, 10

[54] R. Rei, C. A. Stewart, A. C. Farinha, and A. Lavie, “Comet: A neural framework for mt evaluation,” *ArXiv*, vol. abs/2009.09025, 2020. 8

[55] C. Wang, A. Wu, and J. Pino, “Covost 2 and massively multilingual speech-to-text translation,” *arXiv: Computation and Language*, 2020. 9

[56] D. Vilar, M. Freitag, C. Cherry, J. Luo, V. Ratnakar, and G. F. Foster, “Prompting palm for translation: Assessing strategies and performance,” *ArXiv*, vol. abs/2211.09102, 2022. 9

[57] E. Yilmaz, H. V. D. Heuvel, and D. V. Leeuwen, “Acoustic and textual data augmentation for improved asr of code-switching speech,” *ArXiv*, vol. abs/1807.10945, 2018. 9

[58] A. Poncelas, D. Shterionov, A. Way, G. M. de Buy Wenniger, and P. Passban, “Investigating backtranslation in neural machine translation,” *ArXiv*, vol. abs/1804.06189, 2018. 9, 12

[59] S. Gu and Y. Feng, “Improving zero-shot multilingual translation with universal representations and cross-mappings,” *ArXiv*, vol. abs/2210.15851, 2022. 10

[60] R. Dabre, C. Chu, and A. Kunchukuttan, “A comprehensive survey of multilingual neural machine translation,” *ArXiv*, vol. abs/2001.01115, 2020. 10

[61] X. V. Lin, T. Mihaylov, M. Artetxe, T. Wang, S. Chen, D. Simig, M. Ott, N. Goyal, S. Bhosale, J. Du, R. Pasunuru, S. Shleifer, P. S. Koura, V. Chaudhary, B. O'Horo, J. Wang, L. Zettlemoyer, Z. Kozareva, M. T. Diab, V. Stoyanov, and X. Li, “Few-shot learning with multilingual language models,” *ArXiv*, vol. abs/2112.10668, 2021. 10

[62] L. Chai, J. Yang, T. Sun, H. Guo, J. Liu, B. Wang, X. Liang, J. Bai, T. Li, Q. Peng, and Z. Li, “xcot: Cross-lingual instruction tuning for cross-lingual chain-of-thought reasoning,” *ArXiv*, vol. abs/2401.07037, 2024. 10

[63] A. Berard, O. Pietquin, C. Servan, and L. Besacier, “Listen and translate: A proof of concept for end-to-end speech-to-text translation,” *ArXiv*, vol. abs/1612.01744, 2016. 10

[64] C. Wang, Y. Wu, S. Liu, Z. Yang, and M. Zhou, “Bridging the gap between pre-training and fine-tuning for end-to-end speech translation,” *ArXiv*, vol. abs/1909.07575, 2019. 11

[65] A. Bapna, N. Arivazhagan, and O. Firat, “Simple, scalable adaptation for neural machine translation,” in *Conference on Empirical Methods in Natural Language Processing*, 2019, pp. 1538–1548. 11

[66] S. Wu, A. Conneau, H. Li, L. Zettlemoyer, and V. Stoyanov, “Emerging cross-lingual structure in pretrained language models,” in *Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 6022–6034. 11

[67] R. Sennrich, B. Haddow, and A. Birch, “Improving neural machine translation models with monolingual data,” *ArXiv*, vol. abs/1511.06709, 2015. 12

[68] S. Kudugunta, A. Bapna, I. Caswell, N. Arivazhagan, and O. Firat, “Investigating multilingual nmt representations at scale,” in *Conference on Empirical Methods in Natural Language Processing*, 2019, pp. 1565–1575. 13

[69] S. Communication, L. Barrault, Y.-A. Chung, M. C. Meglioli, D. Dale, N. Dong, M. Duppenthaler, P.-A. Duquenne, B. Ellis, H. ElSahar, J. Haaheim, J. Hoffman, M.-J. Hwang, H. Inaguma, C. Klaiber, I. Kulikov, P. Li, D. Licht, J. Maillard, R. Mavlyutov, A. Rakotoarison, K. R. Sadagopan, A. Ramakrishnan, T. Tran, G. Wenzek, Y. Yang, E. Ye, I. Evtimov, P. Fernandez, C. Gao, P. Hansanti, E. Kalbassi, A. Kallet, A. Kozhevnikov, G. M. Gonzalez, R. S. Roman, C. Touret, C. Wong, C. Wood, B. Yu, P. Andrews, C. Balioglu, P.-J. Chen, M. Costa-jussà, M. Elbayad, H. Gong, F. Guzm'an, K. Heffernan, S. Jain, J. T. Kao, A. Lee, X. Ma, A. Mourachko, B. Peloquin, J. Pino, S. Popuri, C. Ropers, S. Saleem, H. Schwenk, A. Y. Sun, P. Tomasello, C. Wang, J. Wang, S. Wang, and M. Williamson, "Seamless: Multilingual expressive and streaming speech translation," *ArXiv*, vol. abs/2312.05187, 2023. 14

[70] A. Kocabiyikoglu, L. Besacier, and O. Kraif, "Augmenting librispeech with french translations: A multimodal corpus for direct speech translation evaluation," *ArXiv*, vol. abs/1802.03142, 2018. 14