Review

Not peer-reviewed version

Neural Machine Translation and Multilingual NLP: A Survey of Methods, Architectures, and Applications

Yao Yuna , Junhao Song * , Jing Qiao

Posted Date: 6 January 2026

doi: 10.20944/preprints202601.0244.v1

Keywords: neural machine translation; multilingual NLP; transformer models; attention mechanisms; cross-lingual transfer; low-resource languages; large language models; sequence-to-sequence learning

 doi:10.20944/preprints202601.0244.v1

Review

Neural Machine Translation and Multilingual NLP: A Survey of Methods, Architectures, and Applications

Yao Yuna [1], Junhao Song [2,*] and Jing Qiao [3]

[1] AppCubic, USA
[2] Imperial College London, London, UK
[3] University of California, Santa Cruz, CA, USA
* Correspondence: junhaosong23@imperial.ac.uk

Abstract

Neural machine translation (NMT) has revolutionized the field of natural language processing by enabling high-quality automatic translation between languages using deep neural networks. This comprehensive survey examines the evolution of machine translation from statistical methods to modern neural approaches, with particular emphasis on the transformer architecture and its variants that have dominated the field since 2017. We systematically review the fundamental architectures including encoder-decoder models, attention mechanisms, and transformer-based systems, analyzing their theoretical foundations and practical implementations. The survey explores critical challenges in multilingual NLP including low-resource translation, zero-shot learning, cross-lingual transfer, and multimodal translation. We investigate recent advances in massively multilingual models, examining architectures that can translate between hundreds of language pairs within a single model. Furthermore, we discuss the emergence of large language models in translation tasks, analyzing their capabilities and limitations compared to dedicated translation systems. The paper also addresses practical considerations including evaluation metrics, data augmentation techniques, and deployment strategies for production systems. We provide insights into current research trends including document-level translation, simultaneous translation, and neural translation with external knowledge. By synthesizing research from over 100 papers, this survey offers both theoretical foundations and practical guidance for researchers and practitioners working in neural machine translation and multilingual natural language processing.

Keywords: neural machine translation; multilingual NLP; transformer models; attention mechanisms; cross-lingual transfer; low-resource languages; large language models; sequence-to-sequence learning

1. Introduction

Automatic translation between human languages represents one of artificial intelligence's most enduring challenges and, increasingly, one of its greatest successes. Neural machine translation (NMT) has fundamentally transformed this landscape through a paradigm shift: instead of engineering translation through explicit rules or probabilistic models, we now directly learn to map source sentences to high-quality translations using deep neural networks trained on parallel corpora [1]. The transformer architecture's introduction [2] marked a watershed moment in the field, enabling unprecedented translation quality and catalyzing a broader revolution in multilingual natural language processing.

The historical trajectory of machine translation reveals successive paradigm shifts driven by fundamental constraints. Early neural approaches using recurrent neural networks demonstrated the potential of end-to-end learning but struggled fundamentally with long-range dependencies and computational efficiency [3,4]. The attention mechanism [1] provided relief to these constraints by enabling the model to selectively focus on relevant input portions during generation [5]. Yet it was the transformer's self-attention mechanism that truly unlocked neural translation's potential, replacing

 doi:10.20944/preprints202601.0244.v1

sequential computation with parallel operations that capture complex linguistic relationships more naturally.

Modern translation systems confront a landscape of profound opportunities and challenges. The explosive growth of digital content across languages demands systems capable of handling unprecedented diversity in domains, styles, and modalities. Social media introduces informal registers, code-switching, and cultural nuances that challenge traditional approaches. Simultaneously, the emergence of large language models raises fundamental questions about the future of dedicated translation systems: will general-purpose models with vast knowledge ultimately subsume specialized translation, or will hybrid approaches that combine their complementary strengths prove optimal [6]? Recent analyses examine this paradigm shift and how LLMs transform translation performance [7,8]. Furthermore, comparative studies of neural architectures and hybrid approaches have clarified the relative strengths of dedicated NMT systems versus general-purpose LLMs [9,10].

Beyond translation itself, multilingual NLP encompasses cross-lingual understanding, zero-shot transfer, and universal representations. Massively multilingual models like mBERT [11] and XLM-R [12] have demonstrated that a single neural system can learn representations transferable across hundreds of languages. This capability carries profound implications for linguistic equity: languages with limited resources can now benefit from transfer learning from high-resource languages, creating pathways for translation technologies to serve all humanity's linguistic diversity. Recent advances have shown how to enhance these capabilities through improved architectural designs and training methodologies [13].

This comprehensive survey synthesizes the state-of-neural machine translation and multilingual NLP, providing both historical perspective and forward-looking analysis. We examine the theoretical foundations underlying modern translation systems, from information-theoretic perspectives to learning-theoretic frameworks. Our treatment spans the full methodological spectrum: supervised learning with parallel corpora, semi-supervised approaches leveraging monolingual data, and truly unsupervised methods discovering translation through structure alone. Recent comprehensive reviews have systematized progress in evaluation methodologies and low-resource scenarios [14,15].

The practical deployment of translation systems introduces constraints and considerations beyond pure accuracy. These include handling rare words and unknown phenomena, maintaining terminology consistency across documents, ensuring cultural appropriateness, evaluating quality reliably, and deploying systems cost-effectively at scale. We examine how recent advances in architecture, training, and inference address these real-world requirements. By synthesizing evidence from over 100 papers, this survey provides both theoretical grounding and practical guidance for researchers and practitioners advancing neural machine translation and multilingual NLP.

2. Historical Evolution and Background

2.1. From Rule-Based to Statistical Methods

The journey of machine translation began in the 1950s with rule-based systems that relied on linguistic knowledge encoded as grammar rules and bilingual dictionaries. These systems, while interpretable and controllable, required extensive manual effort and struggled with the ambiguity and complexity of natural language. The Georgetown-IBM experiment in 1954 demonstrated automatic translation of 60 Russian sentences into English, sparking initial optimism that was later tempered by the realization of the task's complexity.

Statistical machine translation (SMT) emerged in the late 1980s and dominated the field for two decades [16,17]. SMT systems learned translation probabilities from parallel corpora, decomposing the translation process into multiple sub-problems including word alignment, phrase extraction, and language modeling [18]. The noisy channel model provided a principled probabilistic framework, treating translation as a problem of finding the most likely target sentence given a source sentence. Phrase-based SMT systems extended word-based models by translating sequences of words as units, better capturing local reordering and idiomatic expressions [19,20].

 doi:10.20944/preprints202601.0244.v1

The introduction of log-linear models allowed SMT systems to incorporate multiple features beyond simple translation probabilities. Features such as lexical weighting, phrase penalties, and reordering models were combined using weights learned through minimum error rate training. Hierarchical phrase-based models [21] introduced synchronous context-free grammars to handle long-distance reordering, while syntax-based SMT incorporated linguistic parse trees to guide translation decisions.

2.2. The Neural Revolution

The transition to neural methods began with the use of neural networks as components within SMT systems [22,23]. Neural language models improved fluency by better capturing long-range dependencies than n-gram models [24,25]. Joint models learned to score translation candidates using continuous representations, while neural reordering models predicted word order changes. These hybrid approaches demonstrated the potential of neural methods while maintaining the modular structure of SMT.

The breakthrough came with end-to-end neural machine translation, where a single neural network learned to map source sentences directly to target sentences [26]. The encoder-decoder architecture provided a simple yet powerful framework: an encoder network processes the source sentence into a fixed-dimensional representation, which a decoder network then uses to generate the target sentence. This approach eliminated the need for explicit alignment models, phrase tables, and other components of SMT systems.

Early NMT systems using recurrent neural networks showed promising results but faced significant challenges. The fixed-size bottleneck between encoder and decoder limited the amount of information that could be transmitted, particularly for long sentences. Training was computationally expensive due to the sequential nature of RNNs, and the models struggled with rare words and proper nouns. Despite these limitations, NMT quickly matched and then surpassed SMT performance on many language pairs [27,28].

2.3. The Attention Revolution

The introduction of attention mechanisms [1] fundamentally changed neural machine translation. Instead of compressing the entire source sentence into a fixed-size vector, attention allows the decoder to selectively focus on different parts of the source sentence at each decoding step. This mechanism addresses the information bottleneck problem and provides a form of soft alignment between source and target words.

The attention mechanism computes a context vector for each target position as a weighted sum of encoder hidden states. The weights are determined by an alignment model that scores the relevance of each source position to the current target position. This approach not only improves translation quality but also provides interpretable attention weights that reveal which source words influence each target word. Various attention variants emerged, including global versus local attention [5], multi-head attention, and self-attention mechanisms.

Attention mechanisms enabled several important capabilities in NMT systems. They facilitated better handling of long sentences by maintaining access to all source information throughout decoding. The soft alignment provided by attention weights offered insights into model behavior and aided in error analysis. Furthermore, attention mechanisms proved valuable beyond translation, becoming a fundamental component in various NLP tasks and ultimately leading to the transformer architecture.

3. Fundamental Architectures in Neural Machine Translation

3.1. Encoder-Decoder Framework

The encoder-decoder framework forms the foundation of modern neural machine translation systems. This architecture elegantly decomposes the translation process into two phases: understanding the source sentence and generating the target sentence. The encoder processes the source sequence

 doi:10.20944/preprints202601.0244.v1

$x = (x_1, ..., x_n)$ into a sequence of hidden representations $h = (h_1, ..., h_n)$, while the decoder generates the target sequence $y = (y_1, ..., y_m)$ conditioned on these representations.

In recurrent encoder-decoder models, both components typically use variants of RNNs such as Long Short-Term Memory (LSTM) or Gated Recurrent Unit (GRU) networks [29,30]. The encoder RNN processes the source sentence word by word, updating its hidden state at each step:

$$h_t = f_{enc}(x_t, h_{t-1}) \tag{1}$$

where f_{enc} represents the encoder's recurrent function. Bidirectional encoders process the sequence in both forward and backward directions, concatenating the hidden states to capture both past and future context [31,32].

The decoder generates the target sequence autoregressively, predicting one word at a time based on the encoder representation and previously generated words:

$$p(y_t|y_{<t}, x) = g_{dec}(y_{t-1}, s_t, c_t) \tag{2}$$

where s_t is the decoder's hidden state, c_t is the context vector (in attention-based models), and g_{dec} represents the decoder's output function.

3.2. Transformer Architecture

The transformer architecture [2] revolutionized NMT by replacing recurrence with self-attention, enabling parallel computation and capturing long-range dependencies more effectively. The transformer consists of stacked encoder and decoder layers, each containing multi-head self-attention and position-wise feed-forward networks.

The self-attention mechanism computes representations by relating different positions within a sequence. For each position, it calculates attention weights over all other positions and uses these weights to compute a weighted sum of value vectors:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{3}$$

where Q, K, and V represent queries, keys, and values derived from the input through learned linear transformations, and d_k is the dimension of the key vectors.

Multi-head attention extends this mechanism by computing multiple attention functions in parallel:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, ..., \text{head}_h)W^O \tag{4}$$

where each head computes attention with different learned projections, allowing the model to attend to different types of relationships simultaneously.

The transformer's position-wise feed-forward networks apply the same fully connected layers to each position independently:

$$\text{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2 \tag{5}$$

This simple structure, combined with residual connections and layer normalization, enables very deep models while maintaining stable training dynamics.

Positional encodings are crucial in transformers since the architecture lacks inherent position information. The original transformer uses sinusoidal positional encodings:

$$PE_{(pos,2i)} = \sin(pos/10000^{2i/d_{model}}) \tag{6}$$

$$PE_{(pos,2i+1)} = \cos(pos/10000^{2i/d_{model}}) \tag{7}$$

 doi:10.20944/preprints202601.0244.v1

These encodings allow the model to learn relative positions and generalize to sequence lengths not seen during training. The integration of these architectural components has been extensively studied in recent deep learning frameworks [33].

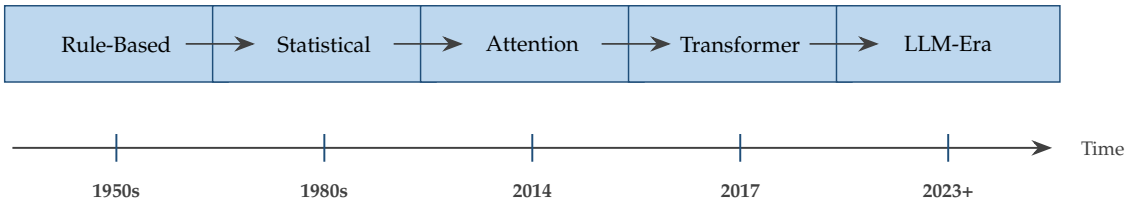

Figure 1. Evolution of machine translation systems from rule-based methods through statistical approaches to neural architectures. Major paradigm shifts are driven by fundamental innovations in handling sequence structure and long-range dependencies, culminating in transformer-based and large language model architectures.

3.3. Variants and Improvements

Numerous architectural variants have been proposed to improve upon the standard transformer. The Transformer-XL [34] introduces recurrence mechanisms to handle longer contexts by caching and reusing hidden states from previous segments. This approach enables learning dependencies beyond fixed-length contexts while maintaining the parallel computation advantages of transformers.

Table 1. Comparison of Transformer Architecture Variants for Machine Translation.

Architecture	Key Innovation	Complexity	Efficiency	Long Context
Vanilla Transformer	Multi-head attention	$O(n^2)$	Baseline	Limited
Transformer-XL	Segment recurrence	$O(n^2)$	Slightly lower	Better
Longformer	Sparse attention	$O(n \log n)$	Higher	Excellent
Reformer	LSH attention	$O(n \log n)$	High	Excellent
Linformer	Linear projection	$O(n)$	Very high	Good
Performer	Kernel methods	$O(n)$	Very high	Good

Sparse transformers reduce the quadratic complexity of self-attention by limiting which positions can attend to each other [35,36]. Patterns such as strided attention and fixed attention patterns maintain model expressiveness while improving efficiency [37,38]. The Reformer [39] uses locality-sensitive hashing to approximate full attention with logarithmic complexity, enabling processing of much longer sequences. Ongoing efforts to improve computational efficiency continue to yield practical improvements for deployment [40]. Recent convolutional approaches for efficient sequence processing have demonstrated competitive performance compared to attention-based methods [41–43].

Adaptive computation approaches dynamically adjust the model's computational depth based on input complexity. Universal Transformers [44] apply the same transformation repeatedly with adaptive halting, allowing different positions to undergo different numbers of transformations. This flexibility enables the model to allocate more computation to challenging parts of the input.

Cross-lingual and multilingual variants of transformers have become increasingly important. mBART [45] extends BART to multiple languages through multilingual denoising pre-training. XLM [46] uses cross-lingual language modeling objectives to learn universal representations. These models demonstrate that transformers can effectively share parameters across languages, learning language-agnostic representations that transfer across linguistic boundaries.

 doi:10.20944/preprints202601.0244.v1

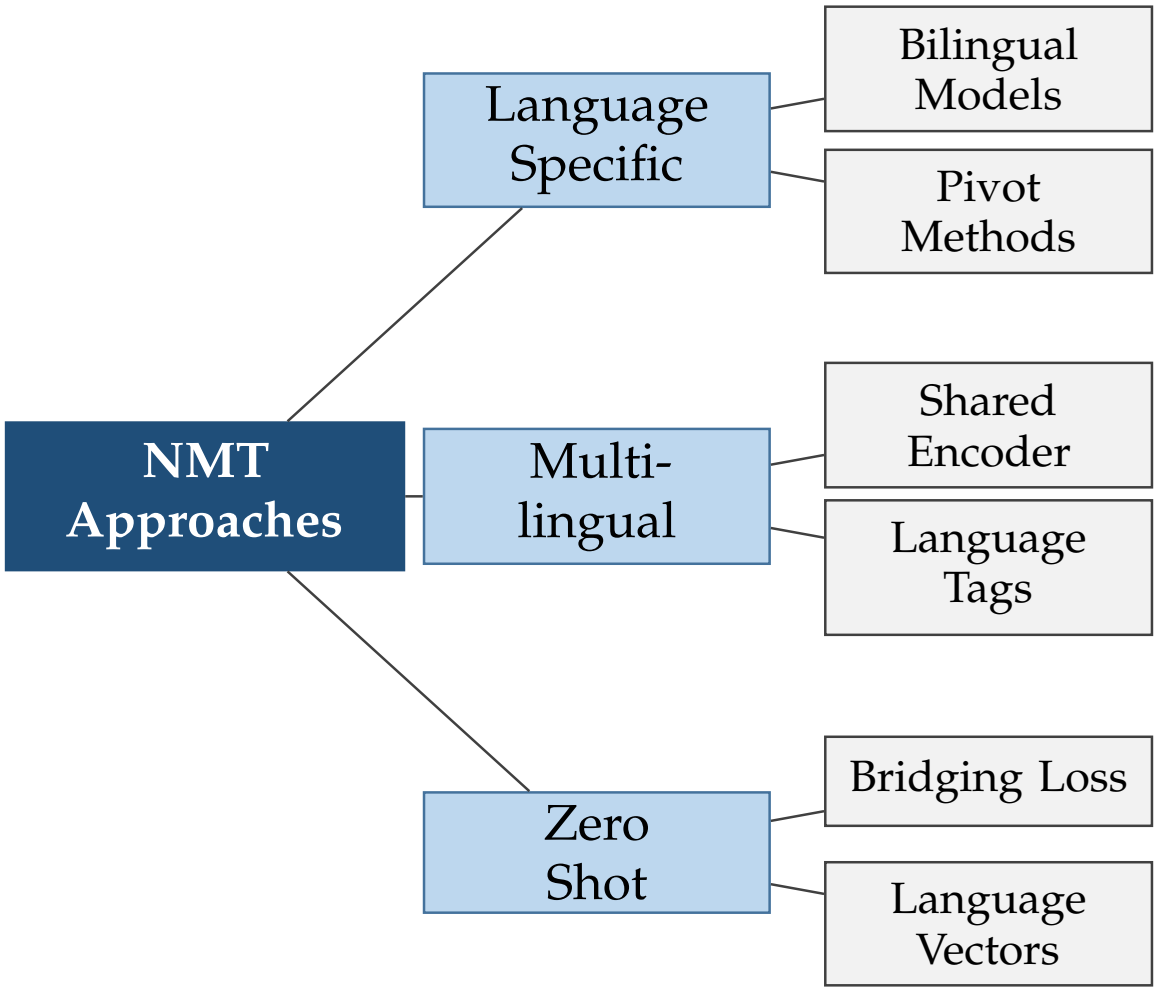

Figure 2. Taxonomy of neural machine translation approaches, ranging from language-specific models to massively multilingual systems with zero-shot translation capabilities. The hierarchical organization reflects the evolution from isolated translation models toward unified systems serving multiple language pairs simultaneously.

4. Multilingual Neural Machine Translation

4.1. Massively Multilingual Models

Massively multilingual neural machine translation represents a paradigm shift from training separate models for each language pair to training a single model capable of translating between many languages. This approach offers several advantages: parameter sharing across languages improves low-resource translation through transfer learning, the model can perform zero-shot translation between language pairs not seen during training, and deployment is simplified with a single model serving multiple language pairs.

Google's Multilingual Neural Machine Translation (GNMT) system [47] pioneered this approach by training a single model on multiple language pairs simultaneously. The key innovation was the use of a shared encoder-decoder architecture with language tags prepended to source sentences to indicate the target language. This simple modification enables the model to learn shared representations across languages while maintaining language-specific generation capabilities [48–51].

The mathematical formulation of multilingual NMT extends the standard translation objective to multiple language pairs:

$$\mathcal{L} = \sum_{(s,t)\in\mathcal{P}} \sum_{(x,y)\in D_{s,t}} -\log p(y|x,t;\theta) \tag{8}$$

where $\mathcal{P}$ represents the set of language pairs, $D_{s,t}$ is the parallel corpus for language pair (s,t), and t indicates the target language.

 doi:10.20944/preprints202601.0244.v1

Scaling to hundreds of languages presents unique challenges. The curse of multilinguality refers to the degradation in high-resource language performance as more languages are added [12,52]. This occurs due to interference between languages and limited model capacity [53,54]. Various solutions have been proposed, including language-specific parameters [55,56], adaptive capacity allocation [57,58], and mixture of experts approaches that route different languages through specialized sub-networks [59,60].

4.2. Zero-Shot and Few-Shot Translation

Zero-shot translation enables models to translate between language pairs never seen during training. This capability emerges from the shared multilingual representations learned by the model. For instance, a model trained on English-French and English-German can potentially translate directly between French and German without explicit training on this pair.

The quality of zero-shot translation depends on several factors including the similarity between languages, the amount of training data for related language pairs, and the model architecture. Encouraging language-agnostic representations through techniques like language adversarial training improves zero-shot performance [61,62]. Adding auxiliary tasks such as denoising autoencoding and cross-lingual language modeling also helps learn better multilingual representations. Contemporary work demonstrates substantial improvements in zero-shot cross-lingual transfer through advanced architectures [63].

Few-shot translation addresses scenarios where only limited parallel data is available for a language pair. Meta-learning approaches treat each language pair as a separate task and learn to quickly adapt to new language pairs from few examples. Techniques from prompt engineering and in-context learning enable large language models to perform translation with just a few demonstration examples [64,65]. Recent work on low-rank adaptation methods has shown promise in efficiently adapting models to new languages [66].

4.3. Cross-Lingual Transfer Learning

Cross-lingual transfer learning leverages knowledge from high-resource languages to improve performance on low-resource languages. This approach is particularly valuable for languages with limited parallel corpora but abundant monolingual data. Pre-trained multilingual models like mBERT and XLM-R learn universal language representations that can be fine-tuned for specific translation tasks [12,46].

Transfer learning strategies in NMT include several approaches. Sequential transfer learning involves first training on high-resource language pairs then fine-tuning on low-resource pairs [67,68]. Multi-task learning jointly trains on multiple language pairs with shared parameters [69,70]. Adapter modules add language-specific parameters to a shared backbone, enabling efficient adaptation while preserving the pre-trained knowledge [71–73]. Recent work has explored parameter-efficient methods that significantly reduce adaptation costs [66,74].

The theoretical understanding of cross-lingual transfer draws from domain adaptation and multi-task learning theory. The key insight is that languages share universal linguistic structures and semantic concepts that can be captured in shared representations. The degree of transfer depends on linguistic similarity, with closely related languages benefiting more from transfer [75]. However, even distant languages can benefit from transfer at higher levels of abstraction, such as syntactic structures and semantic roles. Advanced architectures with selective layer adaptation have shown promise for preserving language-specific phenomena while leveraging shared representations [76].

5. Advanced Techniques and Methods

5.1. Document-Level Translation

Document-level translation extends beyond sentence-by-sentence translation to consider broader context, maintaining consistency in terminology, style, and discourse structure across entire documents.

 doi:10.20944/preprints202601.0244.v1

This approach addresses limitations of sentence-level systems that can produce locally fluent but globally inconsistent translations. Recent advances incorporate explicit discourse coherence mechanisms to achieve more consistent document-level translations [77].

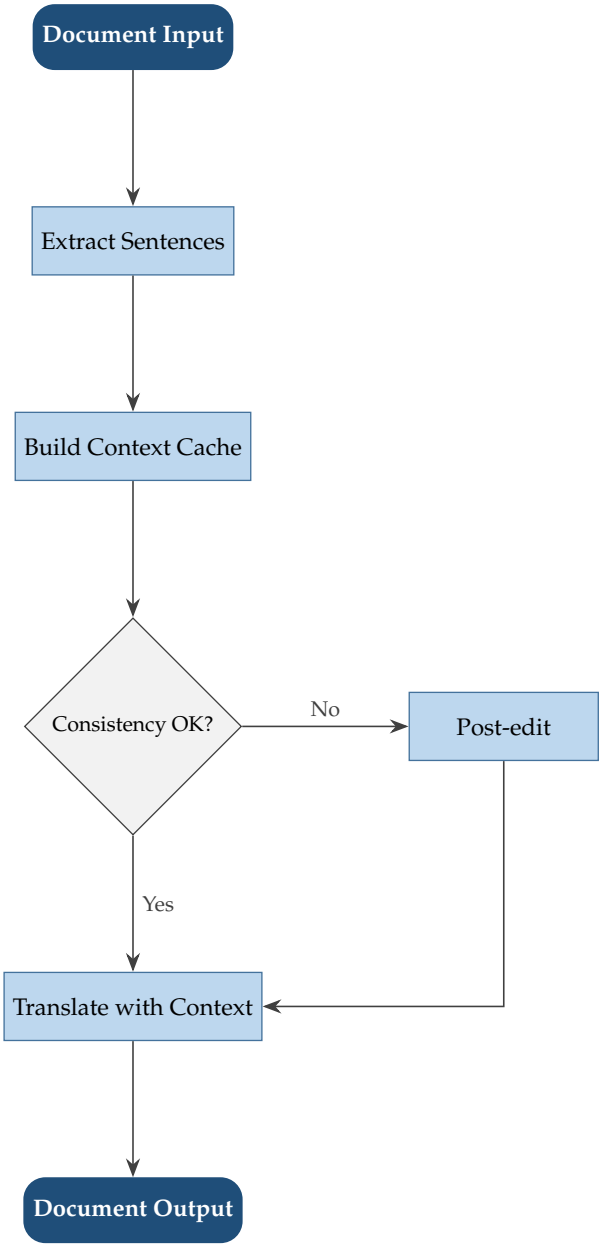

Figure 3. Processing pipeline for document-level neural machine translation. The system maintains context from previously translated sentences and dynamically checks for consistency violations. When detected, post-editing mechanisms ensure terminology and style consistency throughout the document.

Context-aware models incorporate information from surrounding sentences when translating. Hierarchical attention networks process documents at multiple granularities, with separate attention mechanisms for words, sentences, and paragraphs. Memory-augmented networks maintain a memory of previously translated content to ensure consistency. Cache-based models store recent translations and retrieve relevant information when translating new sentences.

Discourse phenomena present particular challenges for document-level translation. Pronoun resolution requires understanding antecedents that may appear in previous sentences. Maintaining consistent translation of terminology throughout a document requires tracking entity mentions. Preserving discourse connectives and maintaining coherent argumentation structure demands understanding of document-level relationships.

 doi:10.20944/preprints202601.0244.v1

The mathematical formulation of document-level translation extends the standard NMT objective to consider context:

$$p(y_i|x_i, C) = \prod_{t=1}^{|y_i|} p(y_{i,t}|y_{i,<t}, x_i, C; \theta) \tag{9}$$

where C represents the document context, which may include previous source sentences, previous target sentences, or both.

5.2. Multimodal Translation

Multimodal translation incorporates visual or audio information alongside text to improve translation quality. This approach is particularly valuable for translating image captions, video subtitles, and other content where visual context provides important disambiguation cues [78,79].

Visual grounding helps resolve ambiguities in translation. For instance, the word "bank" could refer to a financial institution or a riverbank, but accompanying images can clarify the intended meaning. Multimodal attention mechanisms allow the model to attend to relevant image regions when generating translations. Cross-modal alignment ensures that textual and visual representations are properly coordinated. Recent architectures for vision-language understanding have demonstrated improved results in bridging the gap between visual and textual representations [41].

The architecture for multimodal translation typically extends standard NMT models with visual encoders:

$$p(y|x, v) = \prod_{t=1}^{|y|} p(y_t|y_{<t}, x, v; \theta) \tag{10}$$

where v represents visual features extracted from images using convolutional neural networks or vision transformers.

Recent advances in vision-language models like CLIP have enabled better integration of visual and textual modalities. These models learn aligned representations across modalities, facilitating zero-shot transfer and improving translation quality when visual context is available. The emergence of large multimodal models that can process both text and images within a unified architecture opens new possibilities for multimodal translation. Studies on multimodal large language models have explored various architectural approaches to integrate different modalities [80], while specialized work on multimodal machine translation demonstrates integration techniques [81].

5.3. Domain Adaptation and Specialized Translation

Domain adaptation addresses the challenge of translating specialized text such as medical documents, legal contracts, or technical manuals. These domains often have specific terminology, stylistic conventions, and accuracy requirements that generic translation models struggle to handle.

Fine-tuning pre-trained models on domain-specific parallel corpora is the most straightforward approach. However, this can lead to catastrophic forgetting of general translation capabilities. Techniques like elastic weight consolidation and gradient episodic memory help preserve general knowledge while adapting to specific domains. Multi-domain models use domain tags or domain-specific parameters to handle multiple specialized domains within a single model.

Terminology consistency is crucial in specialized translation. Terminology databases can be integrated into NMT systems through constrained decoding, ensuring that specific terms are translated consistently. Placeholder mechanisms handle technical terms and named entities by replacing them with placeholders during translation and reinserting them afterwards. Copy mechanisms allow models to directly copy source terms when appropriate, particularly useful for technical terms that should not be translated.

 doi:10.20944/preprints202601.0244.v1

6. Large Language Models in Translation

6.1. Emergence of LLM-based Translation

Large language models have demonstrated remarkable translation capabilities despite not being explicitly trained for translation. Models like GPT-3 [6] and GPT-4 can perform high-quality translation through few-shot prompting, where a few translation examples are provided as context. This emergent capability suggests that sufficiently large models trained on diverse multilingual data implicitly learn to map between languages.

The translation capabilities of LLMs differ fundamentally from dedicated NMT systems. LLMs leverage their vast knowledge to produce translations that are not only linguistically accurate but also culturally appropriate and contextually aware. They can handle code-switching, idiomatic expressions, and cultural references more naturally than traditional NMT systems. However, they may also hallucinate information or produce overly verbose translations.

Prompt engineering plays a crucial role in LLM-based translation. The quality of translation can vary significantly based on how the task is presented to the model. Effective prompts include clear instructions, relevant examples, and specifications for desired output format. Chain-of-thought prompting, where the model is asked to explain its translation decisions, can improve accuracy for challenging translations [82,83]. Recent advances in securing and optimizing large language models have addressed some of the reliability concerns in translation applications [74].

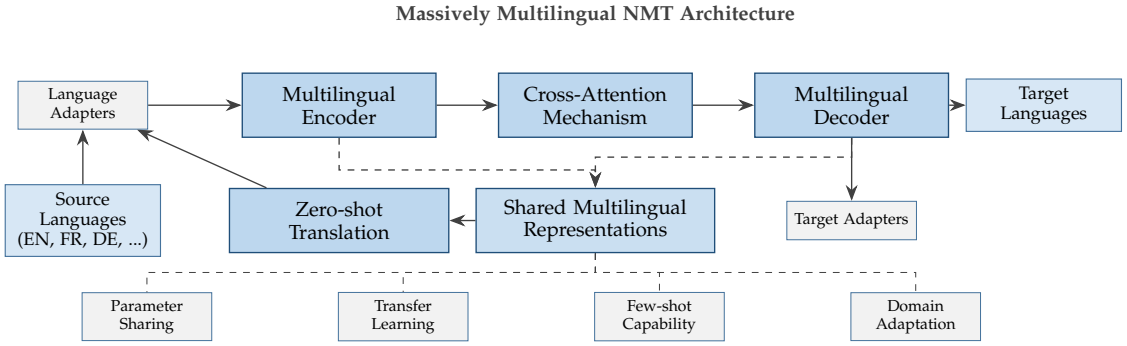

Figure 4. Architecture of massively multilingual neural machine translation systems showing shared representations, language-specific adapters, and zero-shot translation capabilities. The system uses a unified encoder-decoder framework with language-specific adapters enabling efficient parameter sharing across multiple language pairs.

6.2. Comparison with Dedicated NMT Systems

Comparing LLM-based translation with dedicated NMT systems reveals distinct trade-offs. Dedicated NMT systems typically achieve better performance on standard benchmarks, particularly for high-resource language pairs. They are more parameter-efficient, faster at inference, and easier to deploy in production environments. Their behavior is more predictable and controllable, making them suitable for applications requiring consistent quality.

Table 2. Comparison of Dedicated NMT Systems versus Large Language Models for Translation.

Criterion	Dedicated NMT	LLM-Based
Benchmark Performance	Excellent	Very Good
Parameter Efficiency	Very High	Low
Inference Speed	Fast	Slow
Context Window	Limited	Large
Knowledge Integration	Difficult	Natural
Domain Adaptation	Manual fine-tuning	Prompt engineering
Controllability	High	Moderate
Deployment Cost	Low	High

 doi:10.20944/preprints202601.0244.v1

LLMs excel in scenarios requiring world knowledge, cultural awareness, or handling of non-standard language. They can leverage context beyond the immediate sentence, understand implied meaning, and produce more natural-sounding translations for creative or informal text. Their ability to explain translation choices and handle ambiguity through interaction makes them valuable for human-in-the-loop translation workflows.

Hybrid approaches attempt to combine the strengths of both paradigms. Using LLMs to post-edit NMT output can improve fluency and naturalness while maintaining the efficiency of dedicated systems [84,85]. Ensemble methods that combine predictions from multiple models can achieve better accuracy than either approach alone [86,87]. Knowledge distillation from LLMs to smaller NMT models transfers some of the LLM's capabilities while maintaining deployment efficiency [86,88].

6.3. Challenges and Opportunities

The use of LLMs for translation presents several challenges. Computational cost remains a significant barrier, with LLM inference requiring substantially more resources than dedicated NMT systems. Ensuring consistency across long documents is difficult when each segment is translated independently with limited context. The tendency of LLMs to hallucinate or add information not present in the source text raises concerns for applications requiring faithful translation.

Privacy and security considerations are particularly important when using cloud-based LLM services for translation. Sensitive documents may need to be translated locally, requiring efficient deployment strategies for large models. Techniques like quantization, pruning, and knowledge distillation can reduce model size while preserving translation quality. Recent work on hardware-accelerated inference has made local deployment more feasible [89].

Future opportunities lie in developing specialized LLMs for translation that combine the advantages of both paradigms. Instruction-tuned models trained specifically for translation tasks show promise in achieving LLM-like capabilities with improved efficiency. Retrieval-augmented translation, where LLMs access external translation memories or terminology databases, can improve consistency and domain-specific accuracy.

7. Evaluation Metrics and Quality Assessment

7.1. Automatic Evaluation Metrics

Evaluating translation quality automatically enables rapid development and comparison of NMT systems. BLEU (Bilingual Evaluation Understudy) remains the most widely used metric, measuring n-gram overlap between machine translations and human references:

$$\text{BLEU} = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{11}$$

where p_n is the precision of n-grams, w_n are weights (typically uniform), and BP is a brevity penalty to discourage overly short translations.

Table 3. Comparison of Automatic Machine Translation Evaluation Metrics.

Metric	Type	Correlation	Coverage	Interpretation
BLEU	Surface-level	Moderate	N-grams	Word overlap
METEOR	Surface+Semantic	Good	Synonyms	Fluency aware
chrF	Character-level	Good	Characters	Script-robust
BERTScore	Neural embedding	Very Good	Contextual	Semantic similarity
COMET	Learned metric	Excellent	Trained	Human judgment
BLEURT	Fine-tuned BERT	Excellent	Quality dims	Nuanced quality

Alternative metrics address various limitations of BLEU. METEOR considers synonyms and paraphrases through stemming and WordNet lookups [90,91]. chrF operates at the character level,

 doi:10.20944/preprints202601.0244.v1

making it more suitable for morphologically rich languages [92]. BERTScore leverages contextual embeddings to capture semantic similarity beyond surface-level matches [93]. COMET uses neural models trained on human judgments to predict translation quality [94].

Recent neural metrics show stronger correlation with human judgments. BLEURT fine-tunes BERT on synthetic training data to predict human ratings [95]. Prism uses multilingual NMT models as zero-shot paraphrasers to score semantic equivalence. These learned metrics can capture nuances that overlap-based metrics miss, such as fluency, adequacy, and style preservation. Recent large-scale evaluation campaigns have refined assessment methodologies for diverse translation scenarios [96,97].

7.2. Human Evaluation

Human evaluation remains the gold standard for assessing translation quality, though it is expensive and time-consuming. Direct assessment asks evaluators to rate translations on a continuous scale for various quality dimensions. Ranking-based evaluation has annotators order multiple translation outputs by quality [98,99]. Error annotation identifies and categorizes specific translation errors, providing detailed feedback for system improvement.

Multidimensional quality metrics (MQM) provide a framework for comprehensive error annotation. Errors are categorized by type (accuracy, fluency, terminology, style) and severity (critical, major, minor). This structured approach enables detailed analysis of system strengths and weaknesses. The framework supports customization for different use cases, with domain-specific error categories and severity weightings.

Inter-annotator agreement is a crucial consideration in human evaluation. Translation quality is subjective, and evaluators may disagree on acceptable translations. Measuring agreement using metrics like Cohen's kappa or Krippendorff's alpha helps assess evaluation reliability. Using multiple annotators and aggregating their judgments improves robustness. Clear annotation guidelines and evaluator training are essential for consistent evaluation.

7.3. Quality Estimation

Quality estimation (QE) predicts translation quality without reference translations, enabling quality assessment in production settings where references are unavailable. This capability is valuable for identifying translations requiring human post-editing, routing requests to appropriate systems, and providing confidence scores to end users.

QE approaches span multiple granularities from word-level to document-level assessment. Word-level QE identifies potentially mistranslated words, helping post-editors focus on problematic segments. Sentence-level QE predicts overall translation quality scores or post-editing effort. Document-level QE assesses consistency and coherence across entire documents.

Modern QE systems use neural architectures similar to NMT models. Predictor-estimator architectures use one model to extract features and another to predict quality scores [100,101]. Direct estimation approaches train end-to-end models to predict quality from source and target text. Multi-task learning jointly trains QE with related tasks like translation or error detection, improving feature learning.

8. Low-Resource and Unsupervised Translation

8.1. Challenges in Low-Resource Settings

Low-resource translation faces fundamental challenges due to limited parallel data availability. Many of the world's 7,000+ languages have little to no parallel corpora, making traditional supervised approaches infeasible. Even languages with some resources may lack domain-specific parallel data needed for specialized applications. The long tail of language pairs means that most possible translation directions are low-resource.

Data sparsity leads to poor vocabulary coverage, with many words appearing rarely or not at all in training data. Models struggle to learn robust representations and generalize beyond seen examples. Morphologically rich languages are particularly affected, as their large vocabulary sizes

 doi:10.20944/preprints202601.0244.v1

exacerbate sparsity issues. The lack of standardized orthography in some languages adds another layer of complexity. Recent surveys comprehensively analyze these challenges and available solutions [15,102].

Language diversity presents additional challenges. Languages vary dramatically in their morphological complexity, word order, and semantic organization. Models trained on high-resource languages may encode biases that hurt performance on typologically different low-resource languages. The absence of linguistic resources like parsers and taggers for many languages limits the use of linguistically-informed approaches.

8.2. Data Augmentation Techniques

Data augmentation artificially expands limited parallel corpora through various techniques. Back-translation generates synthetic parallel data by translating monolingual target data back to the source language [103]. This approach is particularly effective when combined with techniques like sampling and noising to increase diversity:

$$D_{aug} = D_{parallel} \cup \{(back(y), y) | y \in D_{mono}^{tgt}\} \tag{12}$$

where $back(\cdot)$ represents the back-translation function.

Paraphrasing methods generate alternative translations by modifying existing parallel sentences. Techniques include synonym replacement, phrase substitution, and syntactic restructuring. Neural paraphrasing models trained on high-resource languages can be applied to generate diverse translations. These augmented examples help models learn invariance to surface form variations. Recent parameter-efficient approaches have proven particularly effective for adapting translation models with limited labeled data [104]. Moreover, data-centric approaches emphasizing careful data selection and filtering have shown comparable or superior results to large-scale augmentation strategies [79].

Cross-lingual data augmentation leverages related high-resource languages. Transliteration can convert between scripts for languages sharing vocabulary. Cognate detection identifies related words across languages. Pivot translation through high-resource intermediary languages creates additional training pairs. Code-switching augmentation mixes languages within sentences, improving robustness to real-world multilingual text.

8.3. Unsupervised and Semi-Supervised Approaches

Unsupervised neural machine translation learns to translate without any parallel data, relying only on monolingual corpora in both languages. The key insight is that languages share structural similarities that can be discovered through careful initialization and training objectives. Cross-lingual word embeddings provide initial alignment between languages, mapping words with similar distributions to nearby points in a shared space [42,105].

The training process alternates between denoising autoencoding and back-translation. Denoising autoencoding trains the model to reconstruct corrupted sentences in the same language:

$$\mathcal{L}_{DAE} = \mathbb{E}_{x \sim D_s}[-\log p(x|C(x), s)] \tag{13}$$

where $C(\cdot)$ is a corruption function (e.g., word dropping, shuffling) and s indicates the language.

Back-translation creates on-the-fly parallel data by translating in one direction and training in the reverse:

$$\mathcal{L}_{BT} = \mathbb{E}_{x \sim D_s}[-\log p(x|M(x, t), s)] \tag{14}$$

where $M(x, t)$ represents translation from language s to t.

Semi-supervised approaches combine limited parallel data with larger monolingual corpora. Pre-training on monolingual data through language modeling or denoising objectives provides better initialization. Iterative back-translation progressively improves translation quality by generating

 doi:10.20944/preprints202601.0244.v1

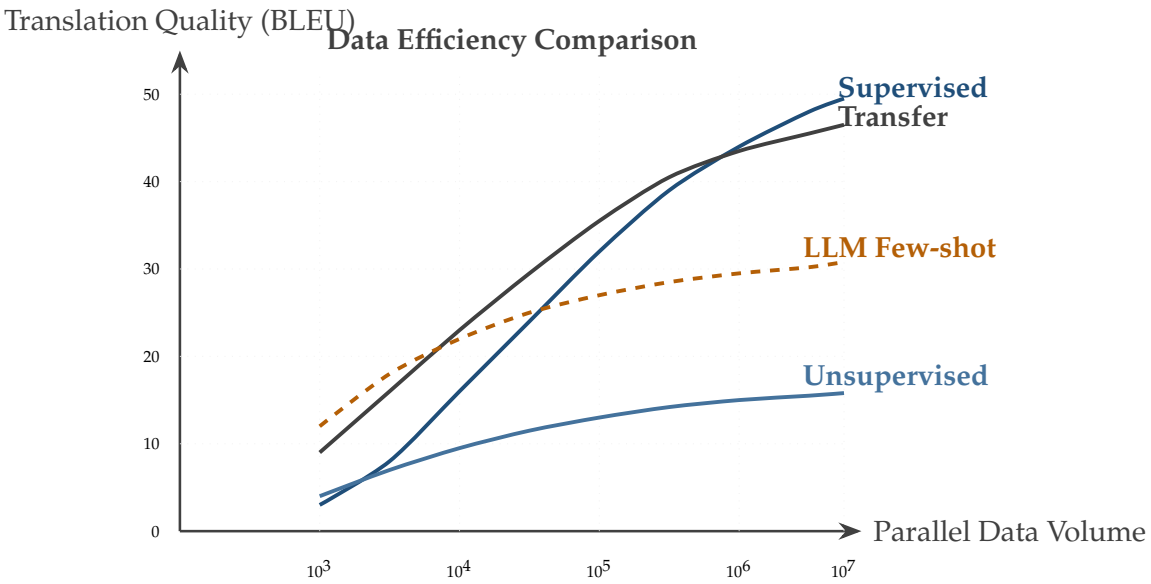

Figure 5. Comparative performance curves of different translation approaches across varying amounts of parallel training data. Supervised neural machine translation achieves the highest asymptotic performance but requires substantial data. Transfer learning from high-resource language pairs provides favorable data efficiency. Unsupervised methods and LLM few-shot approaches show promise in extremely low-resource settings where parallel data is scarce, though they plateau at lower absolute quality levels.

increasingly better synthetic parallel data. Multi-task learning jointly optimizes supervised translation and unsupervised auxiliary objectives. Recent research has shown how these techniques can be enhanced through advanced model architectures and training strategies, with attention to curriculum learning and data augmentation schedules that adaptively adjust difficulty [106,107]. Moreover, novel decoding strategies and ensemble approaches have demonstrated complementary improvements to training-time innovations [39].

9. Practical Implementation and Deployment

9.1. Training Strategies and Optimization

Training large-scale NMT models requires careful optimization strategies to achieve good performance within reasonable time and resource constraints. Learning rate scheduling plays a crucial role, with warmup phases helping stabilize early training and decay schedules preventing overfitting. The Noam schedule, popularized by the original Transformer, combines linear warmup with inverse square root decay:

$$lr = d_{model}^{-0.5} \cdot \min(step^{-0.5}, step \cdot warmup^{-1.5}) \tag{15}$$

Gradient accumulation enables training with larger effective batch sizes than GPU memory allows. Label smoothing regularizes the model by preventing overconfidence in predictions, replacing hard targets with soft distributions:

$$p_{smooth}(k) = (1 - \epsilon) \cdot \delta_{k,y} + \frac{\epsilon}{K} \tag{16}$$

where ϵ is the smoothing parameter, δ is the Kronecker delta, and K is the vocabulary size.

Mixed precision training accelerates computation and reduces memory usage by using float16 for most operations while maintaining float32 master weights for numerical stability. Gradient checkpointing trades computation for memory by recomputing activations during backpropagation rather than storing them [13]. Model parallelism distributes layers across GPUs, while data parallelism

 doi:10.20944/preprints202601.0244.v1

processes different batches on different devices. Advanced distributed training strategies have shown significant improvements in scalability to massive multilingual datasets [108].

9.2. Inference Optimization

Efficient inference is crucial for deploying NMT systems in production environments. Beam search remains the standard decoding algorithm, maintaining multiple hypotheses to find high-probability translations. The beam size trades off between translation quality and computational cost. Diverse beam search and sampling-based methods can produce more varied translations when needed.

Caching and batching strategies significantly improve throughput. Key-value caching in transformers avoids recomputing attention for previously generated tokens. Dynamic batching groups requests with similar lengths to minimize padding overhead. Continuous batching processes requests as they arrive while maintaining high GPU utilization.

Model compression techniques reduce deployment costs while maintaining quality. Quantization reduces numerical precision from float32 to int8 or even lower. Knowledge distillation trains smaller student models to mimic larger teacher models. Pruning removes less important weights based on magnitude or gradient information. These techniques can reduce model size by 10× or more with minimal quality loss.

9.3. Production Systems and Infrastructure

Production NMT systems must handle diverse requirements including low latency for interactive applications, high throughput for batch processing, reliability and fault tolerance, and scalability to handle varying loads. Modern serving frameworks like TensorRT and ONNX Runtime provide optimized inference engines. Container orchestration with Kubernetes enables automatic scaling and load balancing.

Quality assurance in production requires continuous monitoring and evaluation. A/B testing compares different model versions on live traffic. Online learning adapts models to changing data distributions. Fallback mechanisms handle model failures gracefully. Logging and analytics track performance metrics and identify issues.

Cost optimization balances quality requirements with resource constraints. Cascaded models use cheap models for easy translations and expensive models only when needed. Edge deployment brings translation closer to users, reducing latency and bandwidth costs. Serverless architectures provide automatic scaling without maintaining idle resources.

10. Applications and Use Cases

10.1. Commercial Applications

Commercial machine translation serves billions of users daily across diverse applications. Web page translation enables global access to information, with services like Google Translate processing over 100 billion words daily. E-commerce platforms use NMT to translate product descriptions, reviews, and customer communications, enabling cross-border trade. Social media platforms employ real-time translation to connect users across language barriers.

Enterprise solutions address specialized business needs. Document translation systems handle various formats while preserving layout and formatting. Email translation integrates with corporate communication tools. Customer support translation enables multilingual help desks and chatbots. These systems often require domain adaptation and terminology management to maintain consistency with corporate glossaries.

Content localization goes beyond literal translation to adapt content for different markets. Video game localization translates dialogue, user interfaces, and cultural references. Streaming services translate subtitles and dubbing scripts while maintaining timing constraints. Marketing content requires transcreation to preserve persuasive impact across cultures.

 doi:10.20944/preprints202601.0244.v1

10.2. Academic and Research Applications

Academic applications of NMT facilitate global research collaboration and knowledge dissemination. Scholarly article translation helps researchers access publications in other languages, though technical terminology and mathematical notation require special handling. Conference presentation translation enables broader participation in international events. Patent translation supports prior art searches and international filing.

Digital humanities projects use NMT to make historical texts accessible across languages. Literary translation experiments explore whether neural models can capture stylistic nuances and poetic devices. Endangered language documentation benefits from translation tools that work with limited resources. Cross-lingual information retrieval uses translation to search foreign language documents.

Interdisciplinary research leverages NMT in various fields. Biomedical translation handles clinical trials, medical records, and research papers. Legal translation deals with contracts, regulations, and court documents. Technical translation covers engineering specifications, user manuals, and safety documentation. Each domain requires specialized models and evaluation criteria.

10.3. Social Impact and Accessibility

Machine translation has profound social implications for global communication and information access. Crisis response benefits from rapid translation of emergency communications, social media posts for situational awareness, and coordination between international aid organizations. Real-time translation enables cross-language collaboration during natural disasters and humanitarian crises.

Educational access improves through translation of educational materials, online courses, and academic resources. Students can access knowledge regardless of the language it was created in. Language learning applications use translation to provide instant feedback and explanations. However, overreliance on machine translation may impact language learning motivation and proficiency.

Healthcare applications include translating patient information, medical histories, and discharge instructions. Telemedicine platforms use real-time translation for consultations across language barriers. Public health campaigns benefit from rapid translation of health information. However, medical translation errors can have serious consequences, requiring careful quality control.

11. Future Directions and Emerging Trends

11.1. Towards Universal Translation

The vision of universal translation—seamless communication across all human languages—drives ongoing research. Massively multilingual models continue to scale, with recent models handling 200+ languages in a single system. However, true universality requires addressing the long tail of low-resource languages, many of which lack digital resources entirely.

Language-agnostic representations that capture meaning independent of surface form show promise for zero-shot translation. Approaches like cross-lingual sentence embeddings and universal dependency parsing aim to find common structures across languages. Multilingual prompt learning enables rapid adaptation to new languages with minimal examples.

Speech-to-speech translation integrates automatic speech recognition, machine translation, and speech synthesis for seamless spoken communication. End-to-end models that directly map source speech to target speech avoid error propagation through pipeline components. Real-time translation with low latency remains challenging, requiring streaming models that can begin translation before the speaker finishes.

11.2. Integration with Large Language Models

The convergence of NMT and large language models represents a major trend. Instruction-following models can perform translation as one of many tasks, controlled through natural language instructions. This flexibility enables handling ambiguous requests, incorporating user preferences, and explaining translation decisions. State-of-the-art approaches have demonstrated that multimodal

 doi:10.20944/preprints202601.0244.v1

capabilities combined with translation-specific fine-tuning can achieve competitive results against dedicated systems [81].

Retrieval-augmented translation combines the generalization of neural models with the precision of translation memories. Models retrieve relevant examples from large databases and use them to guide translation. This approach is particularly valuable for maintaining consistency in technical translation and adapting to specific domains.

Multimodal large language models that process text, images, and audio within a unified framework open new possibilities. Visual context can disambiguate translations, while audio provides prosodic information lost in text. These models may eventually enable translation that preserves not just meaning but also emotion, tone, and cultural nuances. Recent surveys on multimodal transformers and vision-language models have clarified the architectural patterns necessary for effective multimodal integration [109].

11.3. Ethical Considerations and Challenges

The widespread deployment of machine translation raises important ethical considerations. Bias in translation systems can perpetuate stereotypes and discrimination. Gender bias in pronoun translation, cultural bias in metaphor interpretation, and representation bias favoring dominant languages all require attention. Fairness-aware training objectives and diverse evaluation sets help address these issues.

Privacy concerns arise when sensitive documents are translated through cloud services. Federated learning enables training on distributed data without centralization. Differential privacy techniques add noise to protect individual examples while maintaining overall quality. On-device translation keeps data local but requires efficient models.

The impact on human translators is complex. While NMT automates routine translation, it also creates new roles in post-editing, quality assurance, and specialized translation. The relationship between human and machine translation is evolving from replacement to collaboration, with humans handling creative and culturally sensitive content while machines handle volume. Recent developments in explainable AI for language models help build trust in automated translation systems [110].

12. Conclusions

Neural machine translation has fundamentally transformed how we bridge language barriers, evolving from early statistical methods to sophisticated neural architectures capable of translating between hundreds of languages. The progression from recurrent neural networks through attention mechanisms to transformers represents not just technological advancement but a deeper understanding of how to model language and meaning across linguistic boundaries.

The integration of large language models into the translation landscape introduces both opportunities and challenges. While these models demonstrate remarkable capabilities in handling context, ambiguity, and cultural nuances, they also raise questions about efficiency, controllability, and the future role of dedicated translation systems. The field is moving towards hybrid approaches that combine the strengths of different paradigms, leveraging the efficiency of specialized models with the flexibility and knowledge of large language models.

Multilingual and low-resource translation remain active areas of research with significant social impact. The ability to translate between any pair of the world's languages would democratize access to information and enable communication across all human communities. Progress in zero-shot translation, cross-lingual transfer, and unsupervised learning brings us closer to this goal, though significant challenges remain for truly low-resource languages.

The practical deployment of NMT systems continues to evolve with advances in model compression, inference optimization, and quality estimation. Production systems must balance multiple constraints including quality, latency, cost, and privacy. The development of better evaluation metrics

 doi:10.20944/preprints202601.0244.v1

that correlate with human judgments and capture different aspects of translation quality remains an important challenge.

Looking forward, the field of neural machine translation stands at an exciting juncture. The convergence of NMT with large language models, multimodal processing, and speech technology promises more natural and comprehensive translation capabilities. However, realizing this potential requires addressing fundamental challenges in efficiency, fairness, and evaluation. Recent surveys systematically document progress in transformer efficiency, parameter-efficient methods, and emergent architectures [66,111,112], while contemporary work on generative models and mixture-of-experts approaches demonstrates scalability to massively multilingual scenarios [108,113]. The continued collaboration between researchers, practitioners, and language communities will be essential in developing translation technology that serves all of humanity's linguistic diversity. Advanced training methodologies including masked language modeling and denoising pre-training have proven fundamental to multilingual model development [114,115]. Furthermore, specialized architectures for processing diverse modalities and robust handling of challenging linguistic phenomena continue to advance the field [116,117].

The social implications of advanced translation technology extend far beyond technical considerations. As these systems become more capable and ubiquitous, they will reshape global communication, education, commerce, and cultural exchange. Ensuring that these benefits are distributed equitably and that the technology respects linguistic diversity and cultural values remains a critical challenge for the field.

References

1. Bahdanau, D.; Cho, K.; Bengio, Y. Neural machine translation by jointly learning to align and translate. In Proceedings of the International Conference on Learning Representations, 2014.
2. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in neural information processing systems, 2017, pp. 5998–6008.
3. Sutskever, I.; Vinyals, O.; Le, Q.V. Sequence to sequence learning with neural networks. In Proceedings of the Advances in neural information processing systems, 2014, pp. 3104–3112.
4. Cho, K.; Van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078* **2014**.
5. Luong, T.; Pham, H.; Manning, C.D. Effective approaches to attention-based neural machine translation. In Proceedings of the Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015, pp. 1412–1421.
6. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. *Advances in neural information processing systems* **2020**, *33*, 1877–1901.
7. Zhu, W.; Liu, H.; Liu, Q.; Liu, J.; Zhou, J.; Zeng, J.; Sun, M. Multilingual machine translation with large language models: Empirical results and analysis. *arXiv preprint arXiv:2304.04675* **2023**.
8. Xu, H.; Kim, Y.J.; Sharaf, A.; Awadalla, H.H. A paradigm shift in machine translation: Boosting translation performance of large language models. In Proceedings of the arXiv preprint arXiv:2309.11674, 2023.
9. Wu, Y.; Schuster, M.; Chen, Z.; Le, Q.V.; Norouzi, M.; Macherey, W.; Krikun, M.; Cao, Y.; Gao, Q.; Macherey, K.; et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144* **2016**.
10. Arivazhagan, N.; Bapna, A.; Firat, O.; Lepikhin, D.; Johnson, M.; Krikun, M.; Chen, M.X.; Cao, Y.; Foster, G.; Cherry, C.; et al. Massively multilingual neural machine translation in the wild: Findings and challenges. *arXiv preprint arXiv:1907.05019* **2019**.
11. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the Proceedings of NAACL-HLT, 2019, pp. 4171–4186.
12. Conneau, A.; Khandelwal, K.; Goyal, N.; Chaudhary, V.; Wenzek, G.; Guzmán, F.; Grave, E.; Ott, M.; Zettlemoyer, L.; Stoyanov, V. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116* **2020**.

 doi:10.20944/preprints202601.0244.v1

13. Chen, K.; Bi, Z.; Niu, Q.; Liu, J.; Peng, B.; Zhang, S.; Liu, M.; Li, M.; Pan, X.; Xu, J.; et al. Deep learning and machine learning, advancing big data analytics and management: Tensorflow pretrained models. *arXiv preprint arXiv:2409.13566* **2024**.
14. Deutsch, D.; et al. Machine Translation Meta Evaluation through Translation Accuracy Challenge Sets. *Computational Linguistics* **2024**, *51*, 73–106.
15. Nzeyimana, A. Low-resource neural machine translation with morphological modeling **2024**. pp. 182–195.
16. Och, F.J. Minimum error rate training in statistical machine translation. In Proceedings of the Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, 2003, pp. 160–167.
17. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. BLEU: a method for automatic evaluation of machine translation. In Proceedings of the Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 2002, pp. 311–318.
18. Koehn, P.; Och, F.J.; Marcu, D. Statistical phrase-based translation. In Proceedings of the Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, 2003, pp. 48–54.
19. Tillmann, C.; Ney, H. Word Reordering and a Dynamic Programming Beam Search Algorithm for Statistical Machine Translation. *Computational Linguistics* **2003**, *29*, 97–133.
20. Zens, R.; Ney, H. Improvements in dynamic programming beam search for phrase-based statistical machine translation. In Proceedings of the International Workshop on Spoken Language Translation, 2008, pp. 198–205.
21. Chiang, D. Hierarchical phrase-based translation. In Proceedings of the computational linguistics, 2007, Vol. 33, pp. 201–228.
22. Schwenk, H. Continuous space language models. *Computer Speech and Language* **2007**, *21*, 492–518.
23. Bengio, Y.; Ducharme, R.; Vincent, P.; Jauvin, C. A Neural Probabilistic Language Model. *Journal of Machine Learning Research* **2003**, *3*, 1137–1155.
24. Auli, M.; Galley, M.; Quirk, C.; Zweig, G. Joint Language and Translation Modeling with Recurrent Neural Networks. In Proceedings of the Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, 2013, pp. 1044–1054.
25. Devlin, J.; Zbib, R.; Huang, Z.; Lamar, T.; Schwartz, R.; Makhoul, J. Fast and Robust Neural Network Joint Models for Statistical Machine Translation. In Proceedings of the Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, 2014, pp. 1370–1380.
26. Kalchbrenner, N.; Blunsom, P. Recurrent continuous translation models. In Proceedings of the Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, 2013, pp. 1700–1709.
27. Bentivogli, L.; Bisazza, A.; Cettolo, M.; Federico, M. Neural versus Phrase-Based Machine Translation Quality: a Case Study. In Proceedings of the Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 2016, pp. 257–267.
28. Toral, A.; Sánchez-Cartagena, V.M. A Multifaceted Evaluation of Neural versus Phrase-Based Machine Translation for 9 Language Directions. In Proceedings of the Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, 2017, pp. 1063–1073.
29. Hochreiter, S.; Schmidhuber, J. Long short-term memory. 1997, Vol. 9, pp. 1735–1780.
30. Chung, J.; Gulcehre, C.; Cho, K.; Bengio, Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555* **2014**.
31. Schuster, M.; Paliwal, K.K. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing* **1997**, *45*, 2673–2681.
32. Britz, D.; Goldie, A.; Luong, M.T.; Le, Q. Massive Exploration of Neural Machine Translation Architectures. In Proceedings of the Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017, pp. 1442–1451.
33. Feng, P.; Bi, Z.; Wen, Y.; Pan, X.; Peng, B.; Liu, M.; Xu, J.; Chen, K.; Liu, J.; Yin, C.H.; et al. Deep Learning and Machine Learning, Advancing Big Data Analytics and Management: Unveiling AI's Potential Through Tools, Techniques, and Applications. *arXiv preprint arXiv:2410.01268* **2024**.
34. Dai, Z.; Yang, Z.; Yang, Y.; Carbonell, J.; Le, Q.V.; Salakhutdinov, R. Transformer-xl: Attentive language models beyond a fixed-length context. In Proceedings of the Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 2978–2988.
35. Child, R.; Gray, S.; Radford, A.; Sutskever, I. Generating Long Sequences with Sparse Transformers. *arXiv preprint arXiv:1904.10509* **2019**.

 doi:10.20944/preprints202601.0244.v1

36. Beltagy, I.; Peters, M.E.; Cohan, A. Longformer: The Long-Document Transformer. *arXiv preprint arXiv:2004.05150* **2020**.
37. Parmar, N.; Vaswani, A.; Uszkoreit, J.; Kaiser, L.; Shazeer, N.; Ku, A.; Tran, D. Image Transformer. In Proceedings of the Proceedings of the 35th International Conference on Machine Learning, 2018, pp. 4055–4064.
38. Sukhbaatar, S.; Grave, E.; Bojanowski, P.; Joulin, A. Adaptive Attention Span in Transformers. In Proceedings of the Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 331–335.
39. Kitaev, N.; Kaiser, Ł.; Levskaya, A. Reformer: The efficient transformer. *arXiv preprint arXiv:2001.04451* **2020**.
40. Tang, Y.; Wang, Y.; Guo, J.; Tu, Z.; Han, K.; Hu, H.; Tao, D. A Survey on Transformer Compression. *arXiv preprint arXiv:2402.05964* **2024**.
41. Gehring, J.; Auli, M.; Grangier, D.; Yarats, D.; Dauphin, Y.N. Convolutional sequence to sequence learning. In Proceedings of the International conference on machine learning, 2017, pp. 1243–1252.
42. Jean, S.; Firat, O.; Cho, K.; Memisevic, R.; Bengio, Y. Montreal neural machine translation systems for WMT'15. In Proceedings of the Proceedings of the Tenth Workshop on Statistical Machine Translation, 2015, pp. 134–140.
43. Chen, H.; Peng, J.; Min, D.; Sun, C.; Chen, K.; Yan, Y.; Yang, X.; Cheng, L. MVI-Bench: A Comprehensive Benchmark for Evaluating Robustness to Misleading Visual Inputs in LVLMs. *arXiv preprint arXiv:2511.14159* **2025**.
44. Dehghani, M.; Gouws, S.; Vinyals, O.; Uszkoreit, J.; Kaiser, Ł. Universal transformers. In Proceedings of the International Conference on Learning Representations, 2018.
45. Liu, Y.; Gu, J.; Goyal, N.; Li, X.; Edunov, S.; Ghazvininejad, M.; Lewis, M.; Zettlemoyer, L. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics* **2020**, *8*, 726–742.
46. Conneau, A.; Lample, G. Cross-lingual language model pretraining. In Proceedings of the Advances in neural information processing systems, 2019, pp. 7059–7069.
47. Johnson, M.; Schuster, M.; Le, Q.V.; Krikun, M.; Wu, Y.; Chen, Z.; Thorat, N.; Viégas, F.; Wattenberg, M.; Corrado, G.; et al. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics* **2017**, *5*, 339–351.
48. Ha, T.L.; Niehues, J.; Waibel, A. Toward multilingual neural machine translation with universal encoder and decoder. *arXiv preprint arXiv:1611.04798* **2016**.
49. Firat, O.; Cho, K.; Bengio, Y. Multi-way, multilingual neural machine translation with a shared attention mechanism. In Proceedings of the Proceedings of NAACL-HLT, 2016, pp. 866–875.
50. Zhang, G.; Chen, K.; Wan, G.; Chang, H.; Cheng, H.; Wang, K.; Hu, S.; Bai, L. Evoflow: Evolving diverse agentic workflows on the fly. *arXiv preprint arXiv:2502.07373* **2025**.
51. Chen, K.; Lin, Z.; Xu, Z.; Shen, Y.; Yao, Y.; Rimchala, J.; Zhang, J.; Huang, L. R2I-Bench: Benchmarking Reasoning-Driven Text-to-Image Generation. *arXiv preprint arXiv:2505.23493* **2025**.
52. Wang, Z.; Lipton, Z.C.; Tsvetkov, Y. On Negative Interference in Multilingual Models: Findings and A Meta-Learning Treatment. In Proceedings of the Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, 2020, pp. 4438–4450.
53. Dufter, P.; Schütze, H. Identifying Elements Essential for BERT's Multilinguality. In Proceedings of the Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, 2020, pp. 4423–4437.
54. Wang, S.; Liu, Y.; Wang, C.; Luan, H.; Sun, M. Language models are good translators. In Proceedings of the arXiv preprint arXiv:2106.13627, 2021.
55. Sachan, D.; Neubig, G. Parameter Sharing Methods for Multilingual Self-Attentional Translation Models. In Proceedings of the Proceedings of the Third Conference on Machine Translation, 2018, pp. 261–271.
56. Platanios, E.A.; Sachan, M.; Neubig, G.; Mitchell, T. Contextual Parameter Generation for Universal Neural Machine Translation. In Proceedings of the Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018, pp. 425–435.
57. Wang, X.; Tsvetkov, Y.; Neubig, G. Balancing Training for Multilingual Neural Machine Translation. In Proceedings of the Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 8526–8537.

 doi:10.20944/preprints202601.0244.v1

58. Zhang, B.; Bapna, A.; Sennrich, R.; Firat, O. Share or Not? Learning to Schedule Language-Specific Capacity for Multilingual Translation. In Proceedings of the International Conference on Learning Representations, 2021.
59. Lepikhin, D.; Lee, H.; Xu, Y.; Chen, D.; Firat, O.; Huang, Y.; Krikun, M.; Shazeer, N.; Chen, Z. GShard: Scaling Giant Models with Conditional Computation and Automatic Sharding. In Proceedings of the International Conference on Learning Representations, 2021.
60. Kudugunta, S.; Huang, Y.; Bapna, A.; Krikun, M.; Lepikhin, D.; Luong, M.T.; Firat, O. Beyond Distillation: Task-level Mixture-of-Experts for Efficient Inference. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2021, 2021, pp. 3577–3599.
61. Arivazhagan, N.; Bapna, A.; Firat, O.; Aharoni, R.; Johnson, M.; Macherey, W. The Missing Ingredient in Zero-Shot Neural Machine Translation. *arXiv preprint arXiv:1903.07091* **2019**.
62. Pham, N.Q.; Niehues, J.; Ha, T.L.; Waibel, A. Improving Zero-shot Translation with Language-Independent Constraints. In Proceedings of the Proceedings of the Fourth Conference on Machine Translation, 2019, pp. 13–23.
63. Li, Z.; Hu, C.; Chen, J.; Chen, Z.; Guo, X.; Zhang, R. Improving Zero-Shot Cross-Lingual Transfer via Progressive Code-Switching. In Proceedings of the Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, 2024.
64. Vilar, D.; Freitag, M.; Cherry, C.; Luo, J.; Ratnakar, V.; Foster, G. Prompting PaLM for translation: Assessing strategies and performance. *arXiv preprint arXiv:2211.09102* **2022**.
65. Zhang, B.; Haddow, B.; Birch, A. Prompting large language model for machine translation: A case study. *arXiv preprint arXiv:2301.07069* **2023**.
66. Liang, C.X.; Bi, Z.; Wang, T.; Liu, M.; Song, X.; Zhang, Y.; Song, J.; Niu, Q.; Peng, B.; Chen, K.; et al. Low-Rank Adaptation for Scalable Large Language Models: A Comprehensive Survey **2025**.
67. Zoph, B.; Yuret, D.; May, J.; Knight, K. Transfer learning for low-resource neural machine translation. In Proceedings of the Proceedings of EMNLP, 2016, pp. 1568–1575.
68. Nguyen, T.Q.; Chiang, D. Transfer Learning across Low-Resource, Related Languages for Neural Machine Translation. In Proceedings of the Proceedings of the Eighth International Joint Conference on Natural Language Processing, 2017, pp. 296–301.
69. Dong, D.; Wu, H.; He, W.; Yu, D.; Wang, H. Multi-task learning for multiple language translation. In Proceedings of the Proceedings of ACL-IJCNLP, 2015, pp. 1723–1732.
70. Luong, M.T.; Le, Q.V.; Sutskever, I.; Vinyals, O.; Kaiser, L. Multi-task Sequence to Sequence Learning. In Proceedings of the International Conference on Learning Representations, 2016.
71. Bapna, A.; Firat, O. Simple, scalable adaptation for neural machine translation. *arXiv preprint arXiv:1909.08478* **2019**.
72. Philip, J.; Berard, A.; Huck, M.; Firat, O. Monolingual adapters for zero-shot neural machine translation. In Proceedings of the Proceedings of EMNLP, 2020, pp. 4465–4470.
73. Üstün, A.; Berard, A.; Besacier, L.; Gallé, M. Multilingual Unsupervised Neural Machine Translation with Denoising Adapters. In Proceedings of the Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021, pp. 6650–6662.
74. Peng, B.; Chen, K.; Li, M.; Feng, P.; Bi, Z.; Liu, J.; Niu, Q. Securing large language models: Addressing bias, misinformation, and prompt attacks. *arXiv preprint arXiv:2409.08087* **2024**.
75. Tang, Y.; Tran, C.; Li, X.; Chen, P.J.; Goyal, N.; Chaudhary, V.; Gu, J.; Fan, A. Multilingual translation from denoising pre-training. In Proceedings of the Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, 2021, pp. 3450–3466.
76. Aharoni, R.; Johnson, M.; Firat, O. Massively multilingual neural machine translation. In Proceedings of the Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, 2019, pp. 3874–3884.
77. Wu, M.; Wang, Y.; Foster, G.; Qu, L.; Haffari, G. Importance-Aware Data Augmentation for Document-Level Neural Machine Translation. In Proceedings of the Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics, 2024, pp. 740–752.
78. Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; Bengio, Y. Show, attend and tell: Neural image caption generation with visual attention. In Proceedings of the International conference on machine learning, 2015, pp. 2048–2057.
79. Tu, Z.; Lu, Z.; Liu, Y.; Liu, X.; Li, H. Modeling coverage for neural machine translation. In Proceedings of the Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 2016, pp. 76–85.

 doi:10.20944/preprints202601.0244.v1

80. Li, M.; Chen, K.; Bi, Z.; Liu, M.; Peng, B.; Niu, Q.; Liu, J.; Wang, J.; Zhang, S.; Pan, X.; et al. Surveying the mllm landscape: A meta-review of current surveys. *arXiv preprint arXiv:2409.18991* **2024**.
81. Shen, H.; et al. A Survey on Multi-modal Machine Translation: Tasks, Methods and Challenges. *arXiv preprint arXiv:2405.12669* **2024**.
82. Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Chi, E.; Le, Q.; Zhou, D. Chain-of-thought prompting elicits reasoning in large language models. In Proceedings of the NeurIPS, 2022.
83. Wang, X.; Wei, J.; Schuurmans, D.; Le, Q.; Chi, E.; Narang, S.; Chowdhery, A.; Zhou, D. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171* **2023**.
84. Freitag, M.; Caswell, I.; Roy, S. APE at Scale and Its Implications on MT Evaluation Biases. In Proceedings of the Proceedings of the Fourth Conference on Machine Translation, 2019, pp. 34–44.
85. Kepler, F.; Trénous, J.; Treviso, M.; Vera, M.; Góis, A.; Farajian, M.A.; Lopes, A.V.; Martins, A.F.T. Unbabel's Participation in the WMT19 Translation Quality Estimation Shared Task. In Proceedings of the Proceedings of the Fourth Conference on Machine Translation, 2019, pp. 78–84.
86. Freitag, M.; Al-Onaizan, Y.; Sankaran, B. Ensemble distillation for neural machine translation. In Proceedings of the arXiv preprint arXiv:1702.01802, 2017.
87. Imamura, K.; Sumita, E. Ensemble and Reranking: Using Multiple Models in the NICT-2 Neural Machine Translation System at WAT2017. In Proceedings of the Proceedings of the 4th Workshop on Asian Translation, 2017, pp. 127–134.
88. Kim, Y.; Rush, A.M. Sequence-level knowledge distillation. In Proceedings of the Proceedings of EMNLP, 2016, pp. 1317–1327.
89. Liu, Z.; Bi, Z.; Song, J.; Liang, C.X.; Wang, T.; Zhang, Y. Hardware Accelerated Foundations for Multimodal Medical AI Systems: A Comprehensive Survey **2025**.
90. Banerjee, S.; Lavie, A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. *Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures* **2005**, pp. 65–72.
91. Denkowski, M.; Lavie, A. Meteor Universal: Language Specific Translation Evaluation for Any Target Language. In Proceedings of the Proceedings of the Ninth Workshop on Statistical Machine Translation, 2014, pp. 376–380.
92. Popović, M. chrF: character n-gram F-score for automatic MT evaluation. *Proceedings of WMT* **2015**, pp. 392–395.
93. Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K.Q.; Artzi, Y. BERTScore: Evaluating text generation with BERT. *arXiv preprint arXiv:1904.09675* **2020**.
94. Rei, R.; Stewart, C.; Farinha, A.C.; Lavie, A. COMET: A neural framework for MT evaluation. In Proceedings of the Proceedings of EMNLP, 2020, pp. 2685–2702.
95. Sellam, T.; Das, D.; Parikh, A. BLEURT: Learning robust metrics for text generation. *Proceedings of ACL* **2020**, pp. 7881–7892.
96. Nakazawa, T.; et al. Overview of the 10th Workshop on Asian Translation. In Proceedings of the Proceedings of the 10th Workshop on Asian Translation, 2023.
97. Haque, R.; et al. Evaluating Machine Translation Quality. In Proceedings of the Proceedings of LREC-COLING 2024, 2024.
98. Bojar, O.; Chatterjee, R.; Federmann, C.; Graham, Y.; Haddow, B.; Huck, M.; Yepes, A.J.; Koehn, P.; Logacheva, V.; Monz, C.; et al. Findings of the 2016 conference on machine translation. In Proceedings of the Proceedings of WMT, 2016, pp. 131–198.
99. Bojar, O.; Chatterjee, R.; Federmann, C.; Graham, Y.; Haddow, B.; Huang, S.; Huck, M.; Koehn, P.; Liu, Q.; Logacheva, V.; et al. Findings of the 2017 Conference on Machine Translation. In Proceedings of the Proceedings of the Second Conference on Machine Translation, 2017, pp. 169–214.
100. Kim, H.; Lee, J.H.; Na, S.H. Predictor-estimator using multilevel task learning with stack propagation for neural quality estimation. *Proceedings of WMT* **2017**, pp. 562–568.
101. Kepler, F.; Trénous, J.; Treviso, M.; Vera, M.; Martins, A.F.T. OpenKiwi: An Open Source Framework for Quality Estimation. In Proceedings of the Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, 2019, pp. 117–122.
102. Xiao, Y.; et al. A Survey on Non-Autoregressive Generation for Neural Machine Translation and Beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2023**, *45*, 11407–11427.

 doi:10.20944/preprints202601.0244.v1

103. Sennrich, R.; Haddow, B.; Birch, A. Improving neural machine translation models with monolingual data. In Proceedings of the Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 2016, pp. 86–96.
104. Liu, Y.; et al. Communication Efficient Federated Learning for Multilingual Neural Machine Translation with Adapter. *arXiv preprint arXiv:2305.12449* **2023**.
105. Graves, A. Adaptive computation time for recurrent neural networks. *arXiv preprint arXiv:1603.08983* **2016**.
106. Fan, A.; Bhosale, S.; Schwenk, H.; Ma, Z.; El-Kishky, A.; Goyal, S.; Baines, M.; Celebi, O.; Wenzek, G.; Chaudhary, V.; et al. Beyond english-centric multilingual machine translation. In Proceedings of the Journal of Machine Learning Research, 2021, Vol. 22, pp. 1–48.
107. Niu, Q.; Liu, J.; Bi, Z.; Feng, P.; Peng, B.; Chen, K.; Li, M.; Yan, L.K.; Zhang, Y.; Yin, C.H.; et al. Large language models and cognitive science: A comprehensive review of similarities, differences, and challenges. *arXiv preprint arXiv:2409.02387* **2024**.
108. Zhang, o. Mixture of Experts for Multilingual Translation. *arXiv preprint* **2025**.
109. Han, K.; et al. A Survey on Vision Transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2023**, *45*, 87–110.
110. Hsieh, W.; Bi, Z.; Jiang, C.; Liu, J.; Peng, B.; Zhang, S.; Pan, X.; Xu, J.; Wang, J.; Chen, K.; et al. A comprehensive guide to explainable ai: From classical models to llms. *arXiv preprint arXiv:2412.00800* **2024**.
111. Song, J.; et al. Transformer Architecture Survey. *arXiv preprint* **2025**.
112. Huang, o. Methodologies for Neural Machine Translation. *arXiv preprint* **2025**.
113. Song, o. Generative Models for Machine Translation. *arXiv preprint* **2025**.
114. Li, M.; Bi, Z.; Wang, T.; Wen, Y.; Niu, Q.; Liu, J.; Peng, B.; Zhang, S.; Pan, X.; Xu, J.; et al. Deep learning and machine learning with gpgpu and cuda: Unlocking the power of parallel computing. *arXiv preprint arXiv:2410.05686* **2024**.
115. Ren, o. Deep Learning for NLP. *arXiv preprint* **2024**.
116. Jing, o. Semantic Processing in Neural Translation. *arXiv preprint* **2024**.
117. Peng, o. Noise Robust Neural Machine Translation. *arXiv preprint* **2024**.