Document-level Machine Translation Through Discourse Modelling: A Survey

Himanshu Dutta **Pushpak Bhattacharyya** **Preethi Jyothi**
Computation for Indian Language Technology
Department of Computer Science and Engineering
Indian Institute of Technology Bombay, India
{himanshud, pb, pjyothi}@cse.iitb.ac.in

Abstract

Machine translation has achieved remarkable success at the sentence level, yet significant challenges persist in generating coherent and consistent translations across entire documents. Document-level machine translation (DocMT) addresses these limitations by incorporating inter-sentential context, enabling improvements in discourse coherence, coreference resolution, and lexical consistency. This survey presents a comprehensive overview of the field, exploring the motivations and challenges of DocMT, state-of-the-art modeling approaches, and evaluation methodologies. We discuss recent advances, including hierarchical attention networks, memory-augmented systems, and pretrained language models, alongside specialized datasets and discourse-sensitive evaluation metrics. The paper also identifies open problems, such as data scarcity, evaluation limitations, and handling long-range dependencies, and outlines promising future directions, including multimodal DocMT and real-time translation systems. By synthesizing key research contributions, this survey provides a roadmap for advancing DocMT toward achieving human-level translation quality.

1 Introduction

Machine Translation (MT) has long been a cornerstone of natural language processing (NLP), facilitating cross-lingual communication in an increasingly interconnected world. Over the years, advancements in neural machine translation (NMT) have propelled the field, with state-of-the-art systems achieving near-human performance in sentence-level translation tasks for high-resource languages. Despite these achievements, a significant limitation persists: the inability of sentence-level models to consider the broader context of a document, leading to inconsistencies and inadequacies in translation.

Sentence-level MT systems operate under the assumption that each sentence can be translated independently. While this approach simplifies the translation process and yields impressive results for isolated sentences, it disregards the contextual dependencies critical for maintaining coherence and consistency across documents. This shortcoming becomes evident in discourse-level phenomena such as:

- **Coreference Resolution:** Correctly translating pronouns (e.g., *he*, *she*, *it*) often requires understanding the antecedents in preceding sentences.
- **Lexical Cohesion:** Ensuring consistent terminology usage throughout a document to maintain fluency and readability.
- **Deixis and Ellipsis:** Interpreting context-dependent expressions and implied meanings that rely on prior text.
- **Discourse Coherence:** Aligning sentences to form a logically structured and cohesive narrative.

Document-level machine translation (DocMT) seeks to address these challenges by integrating inter-sentential context into the translation process. Unlike sentence-level MT, which is prone to local inconsistencies and ambiguity, DocMT focuses on global coherence and consistency, making it better suited for translating long-form content such as legal documents, technical manuals, and literary works.

The motivation for DocMT is both theoretical and practical. From a theoretical perspective, linguistic elements like discourse markers, topic continuity, and pragmatic nuances demand contextual understanding. Practically, human translators naturally consider document context, producing translations that are coherent and consistent at the macro

level. Bridging this gap between human and machine translation is essential for applications where context fidelity is critical.

Despite its potential, DocMT presents significant challenges:

- **Data Scarcity:** The availability of parallel corpora with document-level alignment is limited, especially for low-resource languages.
- **Model Complexity:** Capturing long-range dependencies requires sophisticated architectures and efficient handling of computational overhead.
- **Evaluation Limitations:** Traditional metrics like BLEU, designed for sentence-level evaluation, fail to capture document-level phenomena such as coherence and consistency.

Recent advances in the field have started addressing these issues. Context-aware NMT architectures, pretrained models, and targeted evaluation metrics have shown promise in improving DocMT performance. However, many questions remain unanswered, providing fertile ground for future research.

This paper provides a comprehensive survey of document-level machine translation, synthesizing the state-of-the-art research and highlighting critical gaps and future directions. The remainder of this paper is organized as follows:

- Section 2 delves into the problem statement and motivations for transitioning from sentence-level to document-level MT.
- Section 3 discusses the key challenges unique to DocMT, including discourse phenomena, data scarcity, and computational constraints.
- Section 4 reviews modeling approaches, ranging from context-aware architectures to hierarchical and retrieval-augmented models.
- Section 5 provides an overview of available datasets, benchmarks, and evaluation protocols, emphasizing the need for discourse-sensitive metrics.
- Section 6 highlights recent advances and notable systems in DocMT, showcasing the progress made and current state of the field.
- Section 7 identifies open problems and future directions, exploring avenues for research and innovation.
- Finally, Section 8 concludes the paper, summarizing findings and emphasizing the significance of DocMT in bridging the gap between human and machine translation.

2 Motivation and Problem Statement

The need for document-level machine translation (DocMT) arises from the inherent limitations of sentence-level translation systems. While sentence-level MT excels in processing individual sentences, it fails to account for inter-sentential dependencies, leading to errors in coherence, consistency, and adequacy when translating longer texts.

2.1 Limitations of Sentence-Level MT

Sentence-level MT models operate under the assumption that each sentence is independent, an approach that simplifies modeling but disregards the interconnected nature of language in documents. Some key limitations include:

- **Ambiguity in Pronoun Resolution:** Pronouns such as *he*, *she*, *it*, and *they* often refer to entities mentioned in previous sentences. Without context, determining the correct antecedent can lead to mistranslation.
- **Lexical Cohesion and Consistency:** Sentence-level models may translate the same term differently across sentences, resulting in a lack of cohesion.
- **Ellipsis and Deixis:** Phrases with implied meaning or context-dependent references are often misinterpreted.
- **Discourse Coherence:** Translations that lack logical flow and narrative structure detract from the readability of the text.

For example, consider the following text:

> *The company announced its quarterly earnings yesterday. It reported a significant increase in profits.*

A sentence-level model might fail to link *it* to *the company*, resulting in an incoherent or inaccurate translation.

2.2 Motivations for Document-Level MT

Document-level MT aims to bridge the gap by incorporating inter-sentential context, addressing the limitations of sentence-level models. The key motivations include:

- **Improved Accuracy:** Incorporating context enables more accurate translations of ambiguous terms and references.
- **Enhanced Coherence:** Capturing document-level discourse improves the logical flow and readability of translations.
- **Consistency in Terminology:** Context-aware models ensure uniformity in translating recurring terms and phrases.
- **Alignment with Human Translation:** Human translators naturally consider document-level context, making DocMT a step toward replicating human-like translation quality.

2.3 Formalizing the Problem

The task of document-level machine translation can be formally defined as follows:

$$\hat{D} = \arg\max_{D} P(D|S, C; \theta), \tag{1}$$

where S represents the source document, C denotes the contextual information (e.g., preceding and following sentences), D is the translated document, and θ are the model parameters. The objective is to generate a target document $\hat{D}$ that maximizes the conditional probability given the source document S and context C.

Table 1 summarizes the key differences between sentence-level and document-level MT.

In summary, DocMT represents a paradigm shift in machine translation, addressing critical linguistic and practical issues that arise in real-world applications. The next section explores the challenges associated with implementing effective document-level translation systems.

3 Key Challenges in Document-Level MT

Document-level machine translation (DocMT) introduces several unique challenges that must be addressed to achieve effective translation across entire documents. These challenges stem from the need to handle complex linguistic phenomena, model long-range dependencies, and evaluate outputs in a way that captures document-level quality.

3.1 Discourse Phenomena

One of the most significant challenges in DocMT is accurately translating discourse-level phenomena. These include:

- **Coreference Resolution:** Correctly linking pronouns or noun phrases to their antecedents is crucial for coherence. For instance, translating *"The team won their match. They celebrated with a dinner."* requires associating *"they"* with *"the team"*.
- **Lexical Cohesion:** Maintaining consistent terminology across a document is essential, especially in technical or formal texts.
- **Ellipsis and Deixis:** Handling omitted information (ellipsis) or context-dependent expressions (deixis) demands an understanding of surrounding sentences.
- **Discourse Markers:** Translating conjunctions and transitional phrases (e.g., *however, therefore*) correctly ensures logical flow between sentences.

3.2 Sparse Training Data for Discourse

Document-aligned parallel corpora are scarce compared to sentence-aligned datasets, particularly for low-resource languages. Many available corpora are noisy or domain-specific, limiting their generalizability. This scarcity hinders the ability of models to learn document-level dependencies effectively.

3.3 Modeling Long-Range Dependencies

Capturing dependencies that span multiple sentences or paragraphs requires sophisticated architectures. Traditional NMT models struggle with long-range context due to:

- **Memory Constraints:** Transformer models have quadratic complexity with respect to input length, making them inefficient for long documents.
- **Error Propagation:** Errors in earlier translations can propagate, compounding inconsistencies throughout the document.
- **Attention Diffusion:** As context length increases, attention mechanisms may lose focus on relevant parts of the input.

Aspect	Sentence-Level MT	Document-Level MT
Context Consideration	None	Incorporates inter-sentential context
Coherence	Limited	Enhanced coherence across sentences
Pronoun Resolution	Ambiguous	Resolves based on document context
Lexical Consistency	Inconsistent	Consistent terminology usage

Table 1: Comparison of Sentence-Level and Document-Level MT

3.4 Evaluation Metric Limitations

Traditional evaluation metrics like BLEU and METEOR focus on sentence-level n-gram overlaps, failing to capture:

- **Coherence:** Logical flow and narrative consistency across sentences.
- **Pronoun Resolution and Consistency:** Accurate translation of references and uniform terminology.
- **Fluency:** Human-like readability and stylistic appropriateness across an entire document.

Developing metrics tailored to document-level phenomena is an ongoing area of research.

3.5 Computational and Deployment Constraints

Integrating document-level context in real-time translation systems introduces significant computational challenges, including:

- **Latency:** Increased input length leads to higher processing times, which may not be feasible for real-time applications.
- **Scalability:** Managing large-scale document-level translations for multiple language pairs requires efficient use of resources.

In conclusion, while DocMT offers the potential for significant improvements in translation quality, overcoming these challenges requires innovative modeling approaches, enhanced datasets, and refined evaluation metrics. The next section explores the modeling strategies developed to address these challenges.

4 Modeling Approaches

Advancing document-level machine translation (DocMT) requires innovative modeling approaches that address the challenges of incorporating inter-sentential context, managing long-range dependencies, and preserving computational efficiency. This section categorizes and details various modeling paradigms for DocMT, highlighting their architectures, mechanisms, and contributions to the field.

4.1 Concatenation-Based Methods

One of the simplest approaches to introduce context in DocMT is concatenating multiple sentences before feeding them to the model. This approach treats a sequence of sentences as a single, long input.

4.1.1 Single-Stream Concatenation

In this method, the current sentence and its context (e.g., previous and subsequent sentences) are concatenated with delimiters and passed to the encoder. Mathematically, the input can be represented as:

$$\mathbf{X} = [\mathbf{x}_{t-1}; \texttt{[SEP]}; \mathbf{x}_t; \texttt{[SEP]}; \mathbf{x}_{t+1}]$$

where $\mathbf{x}_t$ represents the tokens of the current sentence, and `[SEP]` is a delimiter token.

Although straightforward, this method often results in performance degradation for longer contexts due to diluted attention over extended input lengths. Studies like Tiedemann and Scherrer (2017) have shown modest improvements using this approach for tasks involving limited context.

4.1.2 Segmented Concatenation with Hierarchical Encoding

To mitigate the issues of single-stream concatenation, hierarchical encoding first processes sentences individually and then aggregates their representations. This is often implemented using hierarchical attention mechanisms.

4.2 Multi-Encoder Architectures

Multi-encoder architectures leverage separate encoders for the current sentence and its context, merging their outputs for translation.

4.2.1 Parallel Encoding

Each sentence in the context is encoded independently using separate encoders:

$$\mathbf{H}_t = \text{Encoder}(\mathbf{x}_t), \quad \mathbf{H}_{t-1} = \text{Encoder}(\mathbf{x}_{t-1})$$

Feature	Single-Stream	Hierarchical
Context Handling	Linear	Structured
Scalability to Long Contexts	Limited	Better
Computational Complexity	Low	Moderate

Table 2: Comparison of Single-Stream and Hierarchical Concatenation Methods

These representations are combined using attention or gating mechanisms during decoding:

$$\mathbf{z} = \alpha \cdot \mathbf{H}_t + (1 - \alpha) \cdot \mathbf{H}_{t-1}$$

where α is a learned weight.

4.2.2 Sequential Encoding

In sequential encoding, context is processed iteratively to capture dependencies:

$$\mathbf{H}_t = \text{Encoder}(\mathbf{H}_{t-1}, \mathbf{x}_t)$$

This approach captures cumulative context but at the cost of increased computational complexity.

4.3 Hierarchical Attention Networks (HANs)

Hierarchical Attention Networks (HANs) extend multi-encoder architectures by employing multi-level attention. Sentence-level representations are aggregated into a document-level representation using:

$$\mathbf{H}_{doc} = \sum_i \beta_i \cdot \mathbf{H}_i, \quad \beta_i = \frac{\exp(\mathbf{u}_i^\top \mathbf{v})}{\sum_j \exp(\mathbf{u}_j^\top \mathbf{v})}$$

where β_i is the attention weight for sentence i, and $\mathbf{v}$ is a learned context vector.

HANs have been particularly effective for capturing inter-sentential dependencies in tasks like anaphora resolution and lexical cohesion (Miculicich and et al., 2018).

4.4 Graph-Based Models

Graph-based models represent documents as graphs, where nodes correspond to sentences or phrases and edges capture relationships such as coreference, discourse relations, or lexical connections. Graph Convolutional Networks (GCNs) are commonly employed in these models.

4.4.1 Discourse Graph Construction

Nodes are initialized with sentence embeddings, while edges represent discourse links derived from rhetorical structure theory (RST) or linguistic annotations. The graph is updated iteratively:

$$\mathbf{H}_i^{(l+1)} = \sigma\left(\mathbf{W} \sum_{j \in \mathcal{N}(i)} \mathbf{H}_j^{(l)} + \mathbf{b}\right)$$

where $\mathcal{N}(i)$ denotes the neighbors of node i, and $\mathbf{W}, \mathbf{b}$ are trainable parameters.

4.4.2 Graph Attention Networks (GATs)

Attention mechanisms can be integrated into GCNs to assign importance to different nodes:

$$\alpha_{ij} = \frac{\exp(\text{LeakyReLU}(\mathbf{a}^\top [\mathbf{H}_i \| \mathbf{H}_j]))}{\sum_{k \in \mathcal{N}(i)} \exp(\text{LeakyReLU}(\mathbf{a}^\top [\mathbf{H}_i \| \mathbf{H}_k]))}$$

4.5 Memory-Augmented Models

Memory-augmented models maintain an external memory to store representations of previous translations, dynamically retrieving relevant information during decoding. For instance:

$$\mathbf{m}_t = \text{RetrieveMemory}(\mathbf{M}_{t-1}, \mathbf{q}_t)$$

where $\mathbf{M}_{t-1}$ is the memory at time $t-1$, and $\mathbf{q}_t$ is the query vector derived from the decoder state.

4.5.1 Cache-Based Approaches

Cache-based methods update memory in real-time to ensure consistent terminology usage:

$$\mathbf{M}_t = \mathbf{M}_{t-1} \cup \{\mathbf{H}_t\}$$

4.6 Pretrained Language Models

Pretrained models like BERT, mBART, and GPT have been fine-tuned for DocMT, leveraging their contextual embeddings to capture document-level dependencies.

4.6.1 Fine-Tuning for DocMT

Pretrained models are adapted by adding task-specific heads or objectives:

$$\mathcal{L} = \mathcal{L}_{\text{MT}} + \lambda \cdot \mathcal{L}_{\text{context}}$$

where $\mathcal{L}_{\text{context}}$ enforces consistency across sentences.

Approach	Strengths	Weaknesses	Complexity
Concatenation-Based	Simple, low cost	Limited to short contexts	Low Multi-Encoder
Flexible, modular	Requires multiple encoders	Moderate	
Hierarchical Attention	Effective for structured dependencies	High computational cost	High
Graph-Based	Captures discourse relations	Dependency on graph structure	High
Memory-Augmented	Consistent terminology usage	Requires efficient memory management	Moderate
Pretrained Models	Rich contextual embeddings	High resource requirements	High

Table 3: Comparison of Modeling Approaches for DocMT

4.7 Comparison of Approaches

5 Datasets and Benchmarks

The development and evaluation of document-level machine translation (DocMT) systems rely heavily on the availability of high-quality datasets and robust benchmarks. These resources enable researchers to train context-aware models, evaluate their performance, and address specific document-level challenges such as coherence, coreference, and consistency.

5.1 Document-Aligned Parallel Corpora

Document-aligned parallel corpora form the backbone of DocMT research, providing sentence pairs aligned at the document level. Some widely used datasets include:

- **Europarl:** A collection of European Parliament proceedings, available in multiple languages. While initially sentence-aligned, efforts have restructured it into document-aligned versions to support DocMT tasks.
- **OpenSubtitles:** Contains subtitles from movies and TV shows, offering informal and conversational text across various languages. Its short, context-rich sentences make it suitable for studying intra-document phenomena.
- **News Commentary:** A dataset of aligned news articles, ideal for analyzing document-level features in formal text.
- **TED Talks:** Contains transcripts of TED Talks in multiple languages, offering diverse topics and consistent document structure.
- **JRC-Acquis:** A legal corpus from European Union documents, valuable for studying consistency and terminology in formal contexts.

Despite their utility, these datasets often have limitations, such as noise, domain specificity, and varying lengths of documents, which can affect their suitability for certain DocMT tasks.

5.2 Specialized Datasets for Document-Level Phenomena

Several datasets have been developed to address specific document-level challenges, including:

- **ContraPro:** A dataset designed for evaluating pronoun translation in context. It provides contrastive test sets where one translation is correct and others contain deliberate errors.
- **MuST-Cinema:** Focuses on subtitle translation, emphasizing coherence and speaker consistency.
- **BWB:** The Bilingual Book Worms dataset, which aligns literary texts across languages, is annotated for coreference and discourse relations.
- **M3T:** A multi-modal dataset combining textual and visual content, enabling studies on multi-modal DocMT.

These specialized datasets allow targeted evaluation of discourse phenomena, offering insights into model capabilities beyond BLEU scores.

5.3 Evaluation Benchmarks and Metrics

Evaluation metrics play a crucial role in assessing the quality of DocMT systems. While traditional metrics like BLEU, METEOR, and TER remain popular, they have limitations in capturing document-level phenomena. Recent efforts have introduced tailored benchmarks and metrics:

- **BLEU and Document-Level BLEU:** Traditional BLEU scores are computed at the sentence level, but document-level BLEU aggregates n-gram overlaps across entire documents, offering a more holistic view.
- **COMET:** A model-based metric that incorporates linguistic features, capable of evaluating coherence and fluency.

- **Contrastive Test Suites:** Datasets like ContraPro use contrastive examples to evaluate specific phenomena, such as pronoun resolution or lexical consistency.
- **Human Evaluation:** Human judgments remain the gold standard, often conducted using Multi-dimensional Quality Metrics (MQM) frameworks to assess fluency, coherence, and adequacy.

5.4 Challenges in Dataset Development

Creating datasets for DocMT involves unique challenges:

- **Data Scarcity:** Document-aligned corpora are relatively scarce compared to sentence-aligned datasets, especially for low-resource languages.
- **Noisy Alignments:** Many document-level corpora are automatically aligned, introducing errors that can affect model performance.
- **Diversity:** Ensuring diversity in topics, domains, and languages is critical to building robust DocMT systems.

Efforts to address these challenges include mining document-level alignments from web resources, as demonstrated by ParaCrawl, and constructing synthetic datasets through back-translation or data augmentation.

In summary, datasets and benchmarks are foundational to advancing DocMT research. While existing resources provide a strong starting point, ongoing efforts to create high-quality, diverse, and specialized datasets will be essential for tackling the complexities of document-level translation. The next section explores the recent advances and notable systems that have leveraged these resources to push the boundaries of DocMT.

6 Recent Advances and Notable Systems

Prior work on document-level MT may be grouped into several strands. Early methods extended sentence models with context from adjacent sentences via multi-encoder or concatenation strategies (Jean et al., 2017; Wang et al., 2017a). Hierarchical attention networks condition translation on both word- and sentence-level encodings to capture structured context (Miculicich et al., 2018). Cache-based approaches store recently translated words or topical words in dynamic and topic caches to model coherence (Kuang et al., 2017; Tong et al., 2020). Continuous cache methods leverage a light-weight history memory to adapt translations on the fly (Tu et al., 2018). More recently, unified context models explicitly encode both local sentence context and global document context in Transformer architectures (Ohtani et al., 2019). Large-scale surveys have summarized these approaches and highlighted persistent gaps in modelling and evaluation (Maruf et al., 2021).

Wang et al. (2025) focuses on treating document-level MT as a sequential translation task. They translate each sentence individually in a sequence, while maintaining a persistent memory, which is updated after the translation of each sentence. Similarly, Hu and Wan (2023) explore the inherent discourse structure present in documents by utilizing the paragraph as the discourse boundary. Both these approaches use heuristic rules to decide discourse boundaries, which are difficult to align with the idea of discourse segmentation for the task of translation.

Ideally, a discourse segmentation strategy should ensure that each discourse unit is self-contained for handling intra-discourse phenomena, such as pronoun resolution and verb tense consistency, utilizing its own context. Simultaneously, it should facilitate the handling of inter-discourse phenomena, like entity co-reference and lexical cohesion, by leveraging context from related discourse units. Further, most memory or cache-based document-level MT systems discussed above treat context as a flat or heuristic aggregation of preceding sentences, failing to model fine-grained dependencies between discourse units (Bawden et al., 2018; Voita et al., 2018, 2019).

Document-Level MT Approaches: Document-to-Sentence (Doc2Sent) methods (Wang et al., 2017b; Miculicich et al., 2018; Guo and Nguyen, 2020) incorporate contextual signals from neighboring sentences to enhance translation quality but often treat sentences as isolated units during generation. This results in fragmented discourse and missed target-side cues, as highlighted by Mino et al. (2020); Jin et al. (2023). On the other hand, Document-to-Document (Doc2Doc) approaches (Wu et al., 2023; Wang et al., 2023; Pang et al., 2025) jointly model multiple sentences, capturing long-range dependencies and improving discourse coherence. However, these approaches often face challenges with ultra-long documents, such as con-

tent omissions and scalability limitations. Recent advances leverage large language models (LLMs) for document-level MT, as demonstrated by Wang et al. (2023); Wu et al. (2024); Li et al. (2025). These models process long contexts to generate more coherent translations and address discourse-level phenomena.

Agentic Frameworks with LLMs: Agentic systems utilize autonomous LLMs to decompose complex tasks into specialized subtasks. Multi-agent architectures, such as ExpeL (Zhao et al., 2024) and DELTA (Wang et al., 2025), employ mechanisms like retrieval, iterative refinement, and multi-level memory to enhance task performance and ensure consistency. Related work (Park et al., 2023; Zhang et al., 2024; Qian et al., 2025; Madaan et al., 2023; Koneru et al., 2024; Guo et al., 2024) explores agentic paradigms for maintaining long-context memory, refining outputs, and addressing discourse-level challenges. These frameworks often draw upon discourse theories Grosz and Sidner (1986); Mann and Thompson (1988) for segmenting and maintaining text coherence.

The field of document-level machine translation (DocMT) has witnessed significant progress in recent years, driven by advancements in neural architectures, the integration of pretraining strategies, enhanced evaluation methodologies, and the availability of specialized datasets. These developments address key challenges in capturing discourse-level phenomena, modeling long-range dependencies, and achieving computational efficiency while maintaining translation quality. This section delves into recent innovations and highlights notable systems that have shaped the current state of DocMT.

6.1 Advancements in Neural Architectures

Traditional sentence-level models struggle to incorporate contextual information effectively, motivating the development of specialized architectures for DocMT. Recent innovations in neural architectures include:

6.1.1 Hierarchical Attention Networks (HANs)

Hierarchical Attention Networks (HANs) operate on two levels: they encode individual sentences and aggregate these sentence representations using higher-level attention mechanisms. By attending to relevant context at both the sentence and document levels, HANs effectively model inter-sentential dependencies, improving coherence and cohesion. For example, Miculicich and et al. (2018) demonstrated that HANs significantly improve pronoun resolution and lexical consistency in multi-sentence documents.

6.1.2 Graph Neural Networks (GNNs)

Graph-based methods represent documents as graphs, where nodes correspond to sentences or phrases and edges capture relationships such as coreference or discourse connections. Systems like Disco2NMT (Zhang and et al., 2022) utilize Graph Convolutional Networks (GCNs) to propagate contextual information across the graph structure. This approach enhances the model's ability to understand complex discourse phenomena, such as rhetorical relations and topic continuity.

6.1.3 Gated Architectures

G-Transformers incorporate gating mechanisms to regulate the flow of contextual information during encoding and decoding. This selective attention to context ensures that irrelevant information does not overwhelm the translation process, leading to improved focus and computational efficiency. Recent experiments with gated architectures have shown significant gains in tasks requiring long-range dependency modeling.

6.2 Integration of Pretrained Language Models

Pretrained language models (PLMs) have been instrumental in advancing DocMT, leveraging their ability to encode rich contextual information. Key contributions include:

6.2.1 mBART and mT5

Multilingual pretrained models such as mBART and mT5 incorporate document-level context during pretraining. These models use denoising objectives that extend across sentences, enabling them to generalize effectively across languages and domains. Fine-tuning these PLMs for DocMT tasks has demonstrated improvements in discourse-sensitive phenomena, including coreference resolution and lexical cohesion.

6.2.2 GPT-Based Approaches

GPT models, especially GPT-3 and GPT-4, have been fine-tuned for DocMT, leveraging their generative capabilities and contextual understanding. Few-shot approaches with GPT have been particularly effective, where in-context learning allows

the model to adapt to document-level nuances without extensive retraining. Studies by Cui and et al. (2024) highlight the potential of GPT models in maintaining coherence across translations.

6.2.3 ParaFormer

ParaFormer is a specialized pretrained model designed for paragraph-level context. By training on larger contextual windows, ParaFormer effectively models long-range dependencies, making it particularly suitable for translating formal documents such as legal texts and research papers.

6.3 Innovative Evaluation Methodologies

The limitations of traditional metrics like BLEU have spurred the development of novel evaluation methodologies tailored for DocMT:

6.3.1 Contrastive Evaluation Sets

Datasets such as ContraPro evaluate models' ability to handle specific discourse phenomena, including pronoun resolution and lexical consistency. These datasets present contrastive examples where models must distinguish between correct and incorrect translations, providing targeted insights into their contextual understanding.

6.3.2 Model-Based Metrics

Metrics like COMET and BLEURT incorporate features from pretrained models, offering a more nuanced evaluation of fluency, coherence, and adequacy. These metrics have shown better correlation with human judgments, making them valuable tools for assessing document-level translation quality.

6.3.3 Human Evaluation Frameworks

Human evaluations remain indispensable for DocMT, particularly for assessing document-level fluency and consistency. Protocols such as the Multi-dimensional Quality Metrics (MQM) framework have been extended to include document-specific criteria, offering a comprehensive assessment of translation quality.

6.4 Notable DocMT Systems

Several state-of-the-art systems have emerged, integrating advancements in architecture, pretraining, and evaluation:

6.4.1 DiscoTransformer

DiscoTransformer combines hierarchical attention with discourse parsing to enhance inter-sentential context modeling. By leveraging discourse relations explicitly, this system achieves improvements in coherence and fluency.

6.4.2 Doc2NMT

Doc2NMT employs a multi-encoder architecture that processes sentences and their surrounding context independently before merging the representations using gating mechanisms. This modular approach ensures flexibility and robustness across diverse document types.

6.4.3 Dynamic Memory Models

Inspired by traditional translation memories, dynamic memory models maintain a cache of previously translated sentences, which can be retrieved during decoding. This mechanism supports consistent terminology usage and resolves ambiguities across documents.

6.4.4 Hierarchical Transformers

Hierarchical Transformers extend standard Transformers with an additional layer of attention dedicated to inter-sentential context. These models have been particularly effective in tasks requiring global document understanding, such as literary translation and technical documentation.

6.5 Emerging Trends and Future Directions

Recent research has also explored emerging areas such as:

- **Multi-Modal Translation:** Systems like M3T integrate textual and visual data, enabling richer contextual understanding for tasks involving images and videos.
- **Low-Resource DocMT:** Techniques such as unsupervised pretraining and synthetic data generation address the scarcity of document-aligned corpora for low-resource languages.
- **Real-Time DocMT:** Advances in computational efficiency, including sparse attention mechanisms and on-the-fly context retrieval, are paving the way for real-time document translation systems.

7 Open Problems and Future Directions

Despite significant advancements in document-level machine translation (DocMT), several open problems persist. Addressing these challenges requires interdisciplinary innovation, integration

of novel methodologies, and development of new benchmarks. This section outlines key open problems and identifies promising future directions for DocMT.

7.1 Discourse Representation and Modeling

Effective document translation requires accurate representation and modeling of discourse phenomena, yet many challenges remain:

7.1.1 Capturing Implicit Discourse Relations

Current models struggle with implicit discourse relations (e.g., causal or temporal connections) that are not explicitly signaled by discourse markers. Advanced models incorporating explicit discourse parsers or leveraging discourse theories (e.g., Rhetorical Structure Theory) may help bridge this gap.

7.1.2 Fine-Grained Contextual Dependencies

Modeling fine-grained dependencies, such as long-range anaphora or topic shifts, remains difficult. Future research could explore hierarchical representations combined with fine-grained attention mechanisms to capture nuanced inter-sentential relationships.

7.2 Evaluation Limitations and New Metrics

Traditional evaluation metrics, such as BLEU, fail to capture document-level phenomena, underscoring the need for robust evaluation frameworks:

7.2.1 Discourse-Sensitive Metrics

Metrics that directly measure coherence, cohesion, and consistency at the document level are still underdeveloped. For example, embedding-based metrics like COMET or BLEURT could be extended with discourse-specific objectives.

7.2.2 Human-Centric Evaluation Protocols

Human evaluations are the gold standard for translation quality but are often time-intensive and subjective. Developing semi-automated evaluation frameworks that combine human judgments with model-based metrics could enhance reliability and scalability.

7.3 Data Scarcity and Low-Resource Scenarios

The scarcity of document-aligned parallel corpora, especially for low-resource languages, poses significant challenges:

7.3.1 Synthetic Data Generation

Techniques such as back-translation, paraphrasing, and monolingual data augmentation offer potential solutions for expanding document-level datasets. Leveraging pretrained language models to generate pseudo-parallel corpora can help address data scarcity.

7.3.2 Unsupervised and Semi-Supervised Learning

Unsupervised and semi-supervised learning paradigms hold promise for low-resource DocMT. Future research could explore cross-lingual transfer learning, where high-resource language pairs inform the training of low-resource models.

7.4 Handling Long-Range Dependencies

Capturing long-range dependencies remains a critical bottleneck in DocMT:

7.4.1 Efficient Architectures for Long Contexts

Transformers with full attention mechanisms struggle with computational efficiency for long documents. Sparse attention models, such as Longformer or BigBird, and segment-level approaches like Transformer-XL, could offer practical solutions.

7.4.2 Memory-Augmented Models

Dynamic memory models that store representations of previously translated sentences or phrases can help manage long-range dependencies. Further research into efficient memory retrieval mechanisms is needed to optimize these systems.

7.5 Incorporating Multi-Modality

Translation tasks often involve multimodal inputs, such as text paired with images or audio. Integrating multimodal context into DocMT could improve translation quality:

7.5.1 Text-Image Translation

Incorporating visual context, such as diagrams or images accompanying text, could provide additional cues for ambiguous phrases. Benchmarks like M3T could guide the development of text-image DocMT systems.

7.5.2 Speech and Video Translation

Expanding DocMT to include speech or video data introduces new challenges in modeling temporal and contextual information across modalities. Joint

embeddings for audio, text, and visual content could enable richer contextual understanding.

7.6 Real-Time Document Translation

Real-time document translation is essential for applications like live subtitling or dynamic web content translation:

7.6.1 Balancing Speed and Quality

Optimizing computational efficiency while maintaining high translation quality is a critical challenge. Lightweight architectures, approximate inference techniques, and incremental translation methods could be explored.

7.6.2 Progressive Document Translation

Progressive translation, where translations are refined iteratively as more context becomes available, could strike a balance between real-time processing and global coherence.

7.7 Addressing Ethical Concerns and Biases

As DocMT systems become more integrated into real-world applications, ethical considerations and biases need to be addressed:

7.7.1 Mitigating Linguistic Biases

Biases in training data, such as gendered translations or cultural insensitivity, can propagate into machine-generated translations. Future research should focus on creating balanced datasets and implementing fairness-aware training objectives.

7.7.2 Transparency and Explainability

Providing interpretable outputs, such as visualizing attention weights or explaining model decisions, could enhance trust in DocMT systems. Explainability is particularly critical in applications involving sensitive or high-stakes content.

7.8 Future Directions

Based on the identified challenges, the following directions offer promising avenues for advancing DocMT:

- **Cross-Disciplinary Integration:** Leveraging insights from computational linguistics, cognitive science, and multimodal learning to enhance discourse understanding and context modeling.
- **Robust Low-Resource Strategies:** Extending transfer learning, data augmentation, and unsupervised approaches to handle diverse and low-resource languages effectively.
- **Personalized Document Translation:** Developing systems that adapt translations based on user preferences, domain knowledge, or stylistic requirements.
- **End-to-End Multimodal DocMT:** Building unified models that jointly process textual, visual, and auditory inputs, paving the way for more holistic translation systems.

8 Conclusion

Document-level machine translation (DocMT) represents a significant leap forward in bridging the gap between machine-generated translations and human-quality outputs. By incorporating inter-sentential context, DocMT addresses fundamental challenges in discourse coherence, coreference resolution, lexical consistency, and long-range dependency modeling. Recent advancements, including hierarchical architectures, memory-augmented systems, and pretrained language models, have demonstrated the potential to achieve higher-quality translations that align closely with human expectations.

Despite these advancements, substantial challenges remain. The scarcity of document-aligned parallel corpora, the limitations of existing evaluation metrics, and the computational overhead of handling long-range dependencies are ongoing barriers to progress. Additionally, expanding DocMT to encompass low-resource languages, real-time applications, and multimodal inputs requires innovative solutions and interdisciplinary collaboration.

This survey highlights the current state of DocMT, synthesizing research across datasets, benchmarks, architectures, and evaluation methods. By identifying open problems and proposing future directions, we aim to provide a roadmap for advancing the field. As DocMT continues to evolve, its success will not only enhance machine translation quality but also contribute to broader advancements in natural language processing, cross-lingual understanding, and global communication.

References

Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2018. Evaluating discourse phenomena in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics:*

Human Language Technologies, Volume 1 (Long Papers), pages 1304–1313, New Orleans, Louisiana. Association for Computational Linguistics.

Jiawei Cui and et al. 2024. Context-aware prompting for document-level machine translation. *Transactions of the ACL*.

Barbara J Grosz and Candace L Sidner. 1986. Attention, intentions, and the structure of discourse. In *Computational Models of Discourse*, pages 31–51. MIT Press.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. Large language model based multi-agents: A survey of progress and challenges. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 8048–8057. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Zhiyu Guo and Minh Le Nguyen. 2020. Document-level neural machine translation using BERT as context encoder. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 101–107, Suzhou, China. Association for Computational Linguistics.

Xinyu Hu and Xiaojun Wan. 2023. Exploring discourse structure in document-level machine translation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13889–13902, Singapore. Association for Computational Linguistics.

Sebastien Jean, Stanislas Lauly, Orhan Firat, and Kyunghyun Cho. 2017. Does neural machine translation benefit from larger context? *arXiv preprint arXiv:1704.05135*.

Linghao Jin, Jacqueline He, Jonathan May, and Xuezhe Ma. 2023. Challenges in context-aware neural machine translation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15246–15263, Singapore. Association for Computational Linguistics.

Sai Koneru, Miriam Exel, Matthias Huck, and Jan Niehues. 2024. Contextual refinement of translations: Large language models for sentence and document-level post-editing. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2711–2725, Mexico City, Mexico. Association for Computational Linguistics.

Shaohui Kuang, Deyi Xiong, Weihua Luo, and Guodong Zhou. 2017. Cache-based document-level neural machine translation. *ArXiv*, abs/1711.11221.

Zongyao Li, Zhiqiang Rao, Hengchao Shang, Jiaxin Guo, Shaojun Li, Daimeng Wei, and Hao Yang. 2025. Enhancing large language models for document-level translation post-editing using monolingual data. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8830–8840, Abu Dhabi, UAE. Association for Computational Linguistics.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, and 1 others. 2023. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36:46534–46594.

William Mann and Sandra Thompson. 1988. Rethorical structure theory: Toward a functional theory of text organization. *Text*, 8:243–281.

Sameen Maruf, Fahimeh Saleh, and Gholamreza Haffari. 2021. A survey on document-level neural machine translation: Methods and evaluation. *ACM Comput. Surv.*, 54(2).

Lesly Miculicich and et al. 2018. Document-level neural machine translation with hierarchical attention networks. *Proceedings of EMNLP*.

Lesly Miculicich, Dhananjay Ram, Nikolaos Pappas, and James Henderson. 2018. Document-level neural machine translation with hierarchical attention networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2947–2954, Brussels, Belgium. Association for Computational Linguistics.

Hideya Mino, Hitoshi Ito, Isao Goto, Ichiro Yamada, and Takenobu Tokunaga. 2020. Effective use of target-side context for neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4483–4494, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Takumi Ohtani, Hidetaka Kamigaito, Masaaki Nagata, and Manabu Okumura. 2019. Context-aware neural machine translation with coreference information. In *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*, pages 45–50, Hong Kong, China. Association for Computational Linguistics.

Jianhui Pang, Fanghua Ye, Derek Fai Wong, Dian Yu, Shuming Shi, Zhaopeng Tu, and Longyue Wang. 2025. Salute the classic: Revisiting challenges of machine translation in the age of large language models. *Transactions of the Association for Computational Linguistics*, 13:73–95.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22.

Hongjin Qian, Zheng Liu, Peitian Zhang, Kelong Mao, Defu Lian, Zhicheng Dou, and Tiejun Huang. 2025. Memorag: Boosting long context processing with global memory-enhanced retrieval augmentation. In *Proceedings of the ACM on Web Conference 2025*, WWW '25, page 2366–2377, New York, NY, USA. Association for Computing Machinery.

Jörg Tiedemann and Yves Scherrer. 2017. Context-aware neural machine translation using synthetic context. *Proceedings of the Second Conference on Machine Translation*, pages 192–202.

Yiqi Tong, Jiangbin Zheng, Hongkang Zhu, Yidong Chen, and Xiaodong Shi. 2020. A document-level neural machine translation model with dynamic caching guided by theme-rheme information. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4385–4395, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Zhaopeng Tu, Yang Liu, Shuming Shi, and Tong Zhang. 2018. Learning to remember translation history with a continuous cache. *Transactions of the Association for Computational Linguistics*, 6:407–420.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019. When a good translation is wrong in context: Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1198–1212, Florence, Italy. Association for Computational Linguistics.

Elena Voita, Pavel Serdyukov, Rico Sennrich, and Ivan Titov. 2018. Context-aware neural machine translation learns anaphora resolution. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1264–1274, Melbourne, Australia. Association for Computational Linguistics.

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023. Document-level machine translation with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16646–16661, Singapore. Association for Computational Linguistics.

Longyue Wang, Zhaopeng Tu, Andy Way, and Qun Liu. 2017a. Exploiting cross-sentence context for neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2826–2831, Copenhagen, Denmark. Association for Computational Linguistics.

Longyue Wang, Zhaopeng Tu, Andy Way, and Qun Liu. 2017b. Exploiting cross-sentence context for neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2826–2831, Copenhagen, Denmark. Association for Computational Linguistics.

Yutong Wang, Jiali Zeng, Xuebo Liu, Derek F. Wong, Fandong Meng, Jie Zhou, and Min Zhang. 2025. DelTA: An online document-level translation agent based on multi-level memory. In *The Thirteenth International Conference on Learning Representations*.

Minghao Wu, George Foster, Lizhen Qu, and Gholamreza Haffari. 2023. Document flattening: Beyond concatenating context for document-level neural machine translation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 448–462, Dubrovnik, Croatia. Association for Computational Linguistics.

Minghao Wu, Thuy-Trang Vu, Lizhen Qu, George Foster, and Gholamreza Haffari. 2024. Adapting large language models for document-level machine translation. *arXiv preprint arXiv:2401.06468*.

Hao Zhang and et al. 2022. Discourse-aware neural machine translation with graph convolutional networks. *Proceedings of ACL*.

Yusen Zhang, Ruoxi Sun, Yanfei Chen, Tomas Pfister, Rui Zhang, and Sercan Arik. 2024. Chain of agents: Large language models collaborating on long-context tasks. *Advances in Neural Information Processing Systems*, 37:132208–132237.

Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. 2024. Expel: Llm agents are experiential learners. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'24/IAAI'24/EAAI'24. AAAI Press.