International Journal of Scientific Research & Engineering Trends
Volume 12, Issue 2, Mar-Apr-2026, ISSN (Online): 2395-566X

Bridging Linguistic and Structural Gaps in Marathi Government Document Translation: A Survey of Modern Approaches

Manasi Waghe, Danish Chandargi, Mohammad Aamir Rayyan, Raviraj Joshi, Dr. A.R. Deshpande
Infotmation Technology Smt.Kashibai Navale College of Engineering Vadgaon. Pune, India.

Abstract- The translation of government and legal documents from Marathi to English poses unique challenges due to linguistic complexity, domain-specific terminology, structural richness, and low-resource constraints. General-purpose machine translation systems often fail to maintain semantic fidelity, formatting, and terminological consistency required for administrative and legal texts. This survey explores recent advances in multilingual machine translation, domain adaptation techniques, OCR-driven document understanding, Marathi-specific NLP resources, and terminology- constrained translation methods. We examine the state-of-the-art in robust Marathi-to-English translation systems and highlight critical gaps, focusing on integrating layout-aware models and domain- specific constraints to improve translation quality and reliability for official government documentation.

Keywords- Document Translation, Layout Preservation, OCR, Neural Machine Translation, Document Understanding, E-Governance.

I. INTRODUCTION

Marathi, one of India's prominent regional languages, is extensively used in government communications, legal documentation, and administrative workflows. These documents often contain highly formalized language, legal jargon, and structured layouts including tables, forms, and annexures [9]. Translating such content accurately and efficiently is essential for multilingual governance, legal accessibility, and archival purposes [7].

Challenges in translating Marathi government documents include:

- **Low-resource scenario**: High-quality parallel corpora for Marathi-English translation are limited, especially for domain-specific texts like legal and administrative documents [17].
- **Complex linguistic structures:** Marathi exhibits rich morphology, postpositions, and formal syntactic structures that differ significantly from English.
- **Terminological consistency:** Government and legal texts contain fixed phrases, official terminology, and legal constructs that must be preserved during translation.
- **Structural preservation:**
- Official documents include ta- bles, numbered sections, stamps, and signatures, making layout-sensitive translation necessary.

While recent advances in neural machine translation have improved general-purpose multilingual translation, existing systems remain inadequate for handling the combined challenges of domain specificity, document structure, and terminology consistency required in government and legal contexts. In particular, most approaches treat translation as a sentence-level task, overlooking document-level coherence and layout preservation, which are critical for real-world administrative applications [3].

This paper surveys existing literature, tools, and methodologies relevant to robust domain-specific translation, identifies critical gaps in current approaches, and proposes directions for integrated systems that maintain both linguistic and structural fidelity.

A. Contributions

The key contributions of this paper are as follows:

- Comprehensive survey: We provide a structured review of recent advances in multilingual machine translation,

domain adaptation techniques, OCR-based document understanding, and Marathi-specific NLP resources relevant to government document translation.

- Gap analysis: We systematically identify key limitations in existing approaches, including lack of terminology enforcement, poor document-level consistency, and absence of layout-aware translation mechanisms.
- Comparative analysis: We present a comparative evaluation of different translation paradigms, highlighting trade-offs between domain accuracy, scalability, and structural awareness.
- Integrated pipeline proposal: We propose a unified translation pipeline that combines layout-aware document processing, terminology-constrained neural translation, and post-processing validation to address both linguistic and structural challenges.
- Future research directions: We outline key areas for future work, including multimodal translation architec- tures, domain-specific dataset creation, and explainable machine translation for legal and administrative applica- tions.

II. LITERATURE SURVEY

A. Multilingual Machine Translation Approaches

Recent advances in neural machine translation (NMT), particularly transformer-based architectures, have significantly improved translation quality for low-resource languages such as Marathi [3]. Models such as IndicTrans2 leverage multilingual transfer learning across Indian languages, enabling parameter sharing and improved generalization in low-resource settings [1]. However, their performance remains constrained in domain-specific contexts due to training on predominantly general-domain corpora.

Large-scale multilingual models such as M2M-100 and NLLB eliminate pivot-based translation, reducing error prop- agation across intermediate languages [4], [5]. While these models demonstrate strong cross-lingual capabilities, they ex- hibit several limitations in the context of government and legal document translation. First, they often fail to preserve domain-specific terminology, leading to semantic drift in legally sensitive phrases. Second, they lack mechanisms for enforcing consistency across long documents, resulting in variability in the translation of repeated terms.

A key limitation across multilingual models is their optimization objective, which prioritizes overall fluency and adequacy rather than domain fidelity. As a result, translations may be grammatically correct but legally inaccurate. Furthermore, these models are not inherently designed to handle structured inputs such as tabular data or multi-section documents, leading to degradation in performance when applied to real-world government files.

To address these limitations, domain adaptation techniques have been proposed, including fine-tuning on curated in-domain corpora and the incorporation of terminology constraints [8]. While fine-tuning improves lexical accuracy, it is often limited by the scarcity of high-quality domain- specific parallel datasets. Consequently, hybrid approaches that combine neural models with rule-based or constraint-driven mechanisms offer a more reliable solution for domain-sensitive translation tasks [6].

B. Translation of Legal and Government Texts

Legal and administrative documents present unique challenges that extend beyond conventional translation tasks [7], [9]. These documents are characterized by rigid syn- tactic structures, domain-specific terminology, and context-dependent semantics, where even minor deviations can lead to significant interpretational errors.

A primary challenge lies in lexical ambiguity, where a single term may correspond to multiple legal interpretations depending on context. Standard NMT systems, which rely on probabilistic token prediction, often fail to resolve such ambiguities accurately without explicit contextual grounding. Additionally, government documents require strict adherence to predefined terminology, including official designations, procedural phrases, and statutory expressions. Inconsistent translation of such terms undermines both legal validity and administrative usability [10].

Another critical aspect is document-level consistency. Un- like sentence-level translation benchmarks, real-world docu- ments require uniform translation of recurring entities and phrases across multiple sections. Existing models lack mech- anisms for maintaining such global consistency, resulting in intra-document variability.

Moreover, structural elements such as sections, clauses, annexures, and enumerations carry semantic significance. Cur-

rent text-only translation systems ignore these structural cues, leading to loss of logical organization and reduced interpretability.

These challenges highlight the need for translation systems that integrate contextual disambiguation, terminology enforcement, and document-level reasoning, rather than relying solely on sentence-level neural translation.

C. Marathi NLP Resources

The effectiveness of domain-specific translation systems depends heavily on the availability of high-quality linguis- tic resources [2]. Monolingual corpora such as L3Cube-MahaCorpus provide extensive datasets for training language models, improving fluency and grammatical accuracy in Marathi [16].

Parallel corpora, particularly Marathi-English datasets, are essential for supervised machine translation [17]. These datasets enable models to learn direct mappings between source and target languages, improving translation accuracy in bilingual contexts. However, domain-specific parallel corpora remain limited, posing a significant challenge.

Lexical resources such as Marathi WordNet provide se- mantic relationships between words, aiding in disambiguation and context-aware translation [18]. Named entity recognition datasets further enhance the system's ability to correctly identify and preserve proper nouns, place names, and official titles. Together, these resources enable more robust handling of rare words, idiomatic expressions, and formal administrative language.

D. Terminology-Constrained and Fact-Consistent Translation

Ensuring semantic fidelity in government document translation requires mechanisms to enforce consistent use of domain-specific terminology. Terminology-constrained neural machine translation introduces constraints during decoding to ensure that predefined terms are used consistently across the translation.

Glossary injection techniques further enhance factual consistency by incorporating domain-specific dictionaries into the translation process. This prevents semantic drift and ensures that critical terms retain their intended meaning across different contexts.

These approaches are particularly important in legal and administrative domains, where even minor inconsistencies can lead to significant interpretational errors. Integrating terminology constraints with neural models results in translations that are both fluent and legally compliant [20], [21].

E. Large Language Models for Document Translation

Recent advancements in large language models (LLMs), such as :contentReference[oaicite:0]index=0 and :contentReference[oaicite:1]index=1, have introduced new paradigms for machine translation, particularly in low-resource and domain-specific settings. Unlike traditional neural machine translation systems, LLMs are trained on large-scale multilingual cor- pora and exhibit strong contextual understanding, enabling improved handling of long-range dependencies and document-level coherence.

LLMs demonstrate several advantages for government and legal document translation. First, their ability to process extended context windows allows them to maintain consistency across long documents, addressing a key limitation of sentence-level NMT systems. Second, LLMs can incorpo- rate implicit world knowledge, improving disambiguation of context-dependent legal terminology. Third, they can be leveraged for post-editing and refinement, enhancing fluency and correcting grammatical inconsistencies in machine-generated translations.

However, despite these advantages, LLM-based approaches present notable limitations. They lack explicit mechanisms for enforcing strict terminology constraints, which is critical in legal and administrative contexts. Additionally, their probabilistic generation process may introduce hallucinations or unintended variations in phrasing, potentially compromising legal accuracy. Furthermore, LLMs are not inherently designed to process structured document layouts, limiting their effectiveness in handling tables, forms, and multi-column formats. Recent research has explored hybrid approaches that com- bine LLMs with traditional NMT systems, where LLMs are used for post-editing, validation, and consistency enforce- ment. Such approaches aim to leverage the strengths of both paradigms, combining the reliability of constrained translation with the contextual reasoning capabilities of LLMs. These developments indicate a promising direction toward document-level, context-aware translation systems for complex administrative data.

F. Proposed Integrated Translation Pipeline

Based on the identified limitations of existing approaches, we propose an integrated pipeline for translating Marathi government documents that jointly addresses linguistic, structural, and domain-specific challenges.

The pipeline begins with document acquisition, where inputs may include scanned images or digitally generated PDFs. A layout-aware document understanding module is employed to extract both textual content and spatial structure, preserving elements such as tables, headers, and section boundaries [11]–[13], [15].

In the preprocessing stage, extracted text is normalized and segmented into semantically coherent units. Named entities and domain-specific terms are identified using lexical resources such as Marathi WordNet and domain glossaries [18], [19].

The core translation module combines a fine-tuned multilingual NMT model with terminology-constrained decoding. This ensures that predefined domain-specific terms are consistently preserved during translation. Additionally, document-level context is incorporated to maintain consistency across sections.

A post-processing layer performs validation using rule-based checks and large language model-based refinement to correct grammatical inconsistencies and enforce domain constraints. This stage also resolves residual ambiguities and improves fluency without compromising semantic fidelity.

Finally, a layout reconstruction module restores the translated content into the original document structure using positional metadata [14]. This ensures that the output document retains both visual and logical organization, making it suitable for official and legal use.

Unlike prior approaches that address individual aspects of the problem in isolation, the proposed pipeline integrates translation, structure preservation, and terminology enforcement into a unified framework, thereby improving both reliability and usability.

G. Comparative Analysis of Existing Approaches

A comparative evaluation of existing approaches reveals critical gaps in their applicability to Marathi government document translation.

Multilingual models provide strong baseline performance but fail to enforce domain constraints [4], [5]. Hybrid systems improve accuracy and consistency but lack scalability and structural awareness [6]. Conversely, layout-aware models effectively capture document structure but do not address translation fidelity [11]–[13], [15].

This fragmentation indicates that no single approach sufficiently addresses all critical requirements. Therefore, an integrated pipeline combining multilingual translation, domain adaptation, terminology constraints, and layout-aware processing is necessary for robust performance.

Table I
Comparison Of Translation Approaches For Government Documents

Approach	Domain Accuracy	Terminology Consistency	Document Consistency	Layout Awareness	Scalability
Multilingual NMT (NLLB, M2M)	Medium	Low	Low	None	High
Indic Models (IndicTrans2)	Medium–High	Medium	Low–Medium	None	High
Hybrid MT Systems	High	High	Medium–High	Low	Medium
Layout-Aware Models	Low	Low	Low	High	Medium

H. Error Analysis in Marathi Government Document Translation

An analysis of existing literature reveals several recurring error patterns in machine translation systems when applied to government and legal documents.

Terminology inconsistency: One of the most critical issues is the inconsistent translation of domain-specific terms across a document. Multilingual NMT models often generate different translations for the same term in different contexts, leading to ambiguity and reduced legal reliability [20].

Semantic drift: Models frequently produce translations that are fluent but semantically inaccurate, particularly for legally sensitive phrases. This occurs due to the optimization of fluency over domain fidelity, resulting in subtle but significant deviations in meaning [21].

Lexical ambiguity errors: Marathi words with multiple meanings are often incorrectly translated when contextual cues are insufficient. In legal documents, such ambiguity can lead to incorrect interpretation of clauses and provisions.

Named entity errors: Improper handling of named entities, such as government bodies, legal acts, and geographical locations, leads to mistranslations or inconsistent transliteration. This reduces both accuracy and usability in official contexts. Structural and layout loss: Standard text-based translation systems fail to preserve document structure, including tables, enumerations, and section hierarchies. This results in outputs that are difficult to interpret and unsuitable for administrative use.

Document-level inconsistency: Existing models lack mechanisms for maintaining coherence across long documents, leading to variation in tone, terminology, and phrasing across sections.

Hallucination and over-generation: Particularly in LLM-based systems, there is a tendency to introduce additional content or rephrase information beyond the source text, which is unacceptable in legal and administrative settings.

These error patterns highlight the limitations of current approaches and reinforce the need for integrated systems that incorporate terminology constraints, document-level context, and layout-aware processing to ensure both linguistic accuracy and structural fidelity.

I. Challenges and Future Directions

Despite advancements in multilingual translation and document understanding, several challenges remain unresolved. The scarcity of domain-specific parallel corpora continues to limit the effectiveness of supervised learning approaches, particularly for specialized administrative language [17], [18]. Another major challenge is terminology drift, where models fail to consistently translate domain-specific terms across long documents [20]. This issue is exacerbated in multilingual settings, where subtle variations in phrasing can alter legal meaning.

Structural complexity further complicates translation, as existing systems struggle to handle multi-column layouts, nested tables, and embedded elements. While layout-aware models address structural representation, they do not integrate seamlessly with translation systems.

Future research should focus on developing end-to-end multimodal architectures that jointly model text, layout, and semantics. The integration of large language models for post-editing and validation presents a promising direction, particularly for improving factual consistency and fluency.

Additionally, the creation of large-scale, domain-specific datasets through institutional collaboration can significantly enhance model performance. Explainable machine translation is another critical area, enabling transparency in translation decisions, which is essential for legal compliance and auditing. Addressing these challenges requires a shift from isolated model improvements to holistic system design that integrates linguistic, structural, and domain-specific considerations.

III. CONCLUSION

This survey highlights the limitations of existing machine translation approaches in handling Marathi government documents, which require not only linguistic accuracy but also structural preservation and domain-specific consistency. While multilingual NMT models provide strong generalization capabilities, they fall short in enforcing terminology constraints and maintaining document-level coherence.

Through a detailed analysis of multilingual models, hybrid approaches, layout-aware systems, and linguistic resources, we identify the need for integrated solutions that combine these complementary strengths. The proposed pipeline addresses this gap by incorporating domain adaptation, terminology constraints, and layout-aware processing into a unified framework. Future work should focus on building scalable, end-to-end systems that jointly model text and document structure while ensuring explainability and legal reliability. Such systems have the potential to significantly enhance multilingual governance, improve accessibility, and enable accurate digitization of administrative processes.

REFERENCES

1. A. Bapna et al., “IndicTrans2: Towards High-Quality and Accessible Machine Translation Models for All 22 Scheduled Indian Languages,” AI4Bharat, 2023.
2. Kunchukuttan et al., “AI4Bharat IndicNLP Suite: Monolingual Corpora, Embeddings and Language Models for Indian Languages,” 2021.
3. G. Kunchukuttan and Pratyush Kumar, “Multilingual Machine Transla- tion for Low Resource Indian Languages,” 2020.
4. Meta AI, “NLLB: No Language Left Behind – Scaling Human-Centered Machine Translation,” 2022.
5. Fan et al., “M2M-100: Multilingual Machine Translation Across 100 Languages without English,” Facebook AI, 2020.
6. S. R. Lilhare and S. A. Katkar, “A Hybrid Marathi-to-English Machine Translation System Using Rule-Based and Statistical Approach,” 2019.
7. Kumar and Bhattacharyya, “Machine Translation for Indian Adminis- trative Language: A Survey,” 2021.
8. Chalkidis et al., “Domain Adaptation for Low Resource Legal Machine Translation,” 2020.
9. P. Deshmukh et al., “Translation of Indian Government Documents: Challenges and Approaches,” 2022.
10. V. Kulkarni, “Handling Ambiguity in Legal Marathi-English Transla- tion,” 2020.
11. Xu et al., “LayoutXLM: Multimodal Pre-training for Multilingual Visually-rich Document Understanding,” 2021.
12. Kim et al., “Donut: Document Understanding Transformer without OCR,” 2022.
13. Huang et al., “LayoutLMv3: Pre-training for Document AI with Unified Text and Image Masking,” 2022.
14. Li et al., “Document Structure Extraction using PyMuPDF and Deep Layout Parsing,” 2021.
15. Appalaraju et al., “DocFormer: End-to-End Transformer for Document Understanding,” 2022.
16. L3Cube Pune, “L3Cube-MahaCorpus: Marathi Monolingual Datasets for NLP,” 2022.
17. IIT Bombay, “Marathi-English Parallel Corpus for Machine Translation,” 2018.
18. IIIT Hyderabad, “Marathi WordNet: A Lexical Database for Marathi,” 2010.
19. Jadhav et al., “Marathi Named Entity Recognition Using Deep Learn- ing,” 2021.
20. Dinu et al., “Terminology-Constrained Neural Machine Translation,” 2019.
21. Song et al., “Fact-Consistent Translation via Glossary Injection,” 2022.