PROBLEM STATEMENT

NLP-Based Multilingual Machine Translation System

For Automobile Domain Technical Documentation

Document Version	1.0
Date	26 November 2025
Project Domain	Natural Language Processing / Machine Translation
Target Industry	Automotive Aftermarket / Commercial Vehicles
Project Owner	Dr Karthikeyan Saminathan, 9791306877, karthikeyans@ggsinc.com

1. OBJECTIVE

The primary objective of this initiative is to develop a robust, domain-specific Neural Machine Translation (NMT) system capable of accurately translating technical automotive documentation from English to multiple Indian vernacular languages while preserving domain-specific terminology, technical accuracy, and contextual meaning.

1.1 Specific Objectives

1. Enable accurate and fluent translation of automobile-related content across multiple Indian languages including Hindi, Tamil, Telugu, Malayalam, and Kannada.
2. Preserve domain-specific terminology such as "torque converter", "ABS", "drivetrain", "ECU", and other automotive technical terms with consistent translations.
3. Support both real-time and batch translation workflows for various content types including service manuals, pamphlets, brochures, and diagnostic procedures.
4. Implement quality estimation mechanisms to flag low-confidence translations for human review, reducing post-editing effort.
5. Create a continuous learning pipeline that incorporates human feedback to iteratively improve translation quality.
6. Achieve translation quality comparable to or exceeding commercial MT systems (Google Translate, Microsoft Azure) as measured by BLEU, COMET, and human evaluation metrics.

1.2. Purpose

India's automotive aftermarket sector serves a vast, multilingual customer base across diverse regional markets. Technical documentation—including service manuals, parts catalogs, diagnostic guides, and training materials—is predominantly created in English, creating significant accessibility barriers for mechanics, technicians, and end-users who operate primarily in their native languages.

1.3 *Existing Problems in Multilingual Translation*:

- Inconsistent accuracy across languages
- Low performance for low-resource Indian languages
- Poor handling of idioms, proverbs, and cultural expressions
- Lack of proper context understanding
- Errors in long-sentence translations
- Code-mixing and code-switching not handled well

- Morphological complexity in Indic languages
- Ambiguity and polysemy causing wrong translations
- Named entities mistranslated or wrongly transliterated
- No domain-specific customization (healthcare, legal, automotive, etc.)
- Weak speech-to-text and speech-to-speech translation accuracy
- Accent, dialect, and noise issues in ASR
- Unicode/script rendering issues across Indic languages
- Limited parallel corpora and training datasets
- Confusion in gender, plurality, and formality levels
- Tone, politeness, and emotion not preserved
- High GPU/compute cost for NMT models
- No automated quality control or terminology enforcement
- Data privacy and security concerns during translation

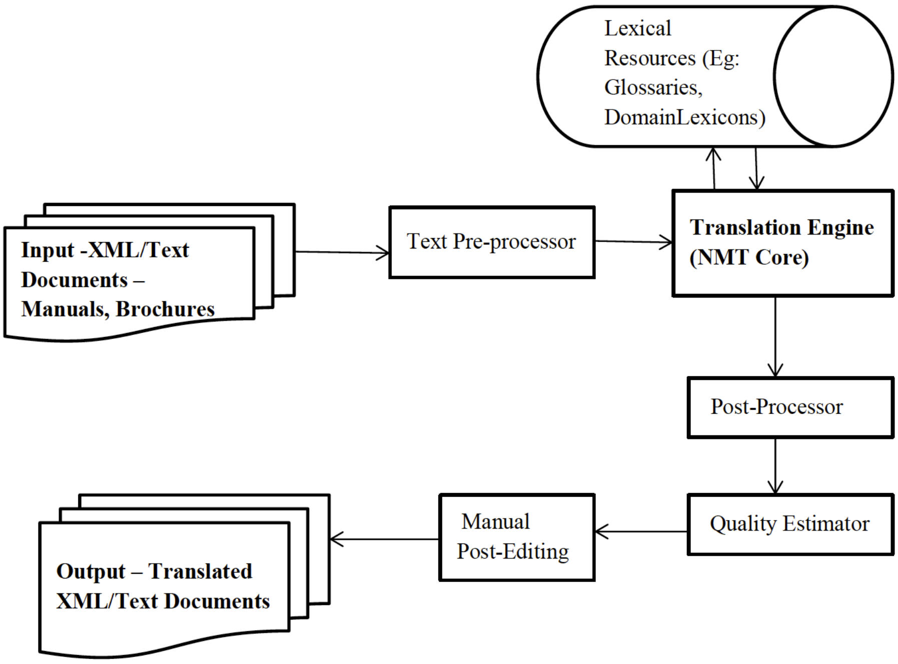
2. PROPOSED SYSTEM

The proposed NLP-Based Multilingual Machine Translation System adopts a modular, pipeline architecture specifically designed for automotive domain translation. The system leverages state-of-the-art transformer-based neural machine translation models fine-tuned on automotive parallel corpora.

2.1 System Architecture Overview

The architecture consists of six interconnected modules working in a sequential pipeline with feedback loops for continuous improvement:

2. System Architecture


Lexical Resources (Eg: Glossaries, DomainLexicons)
Input -XML/Text Documents – Manuals, Brochures
Text Pre-processor
Translation Engine (NMT Core)
Post-Processor
Quality Estimator
Manual Post-Editing
Output – Translated XML/Text Documents

Module Description

a. Text Pre-processor

- Sentence segmentation
- Named entity recognition (NER)
- Glossary enforcement (e.g., “ECU” always translated consistently)

b. Translation Engine

- Fine-tuned Transformer-based Neural MT
- Trained on parallel corpora from automotive manuals, brochures etc
- Supports multiple language pairs (e.g., English →Hindi, Tamil, Malayalam, Telugu, Kannada

c. post-processor

- Domain Terminology consistency check
- Formatting preservation (e.g., Translated text to XML population)
- Compare Named Entity similarity between Source and Target sentences.

d. Quality Estimator

- Confidence scoring per sentence
- Flags low-confidence translations for human review

e. Manual Post-Editing

- Low scored sentences are corrected by manual human editing

4. Lexical Resources

- Parallel corpora from automotive OEMs
- Bilingual (or Multilingual) Domain Term Lexicons
- Named Entity Lexicons

5. Feedback & Continuous Learning

- Human reviewers validate flagged translations
- Active learning loop to retrain models
- Glossary/Lexicons updates

2.2 Key Features of Proposed System

1. **Domain-Adapted NMT Engine:** Leveraging IndicTrans2 or NLLB-200 or any other as base models, fine-tuned on automotive parallel corpora comprising service manuals, parts catalogs, and technical bulletins.
2. **Terminology Enforcement Layer:** Pre-translation glossary lookup ensures critical automotive terms are translated consistently using approved terminology database.
3. **Named Entity Preservation:** NER module identifies part numbers, model names, and specifications, protecting them from translation while enabling appropriate transliteration where needed.
4. **Quality-Aware Pipeline:** Real-time confidence scoring using COMET-QE enables automatic routing of uncertain translations to human reviewers.
5. **Active Learning Loop:** Human corrections are captured and used for periodic model retraining, continuously improving translation quality.
6. **Format Preservation:** XML/DITA structure maintained throughout translation pipeline, eliminating manual reformatting effort.

2.3 Data Flow Diagram

The system processes documents through the following data flow:

```
Input (XML/doc/xlx) → Pre-processor → NMT Engine → Post-processor → Quality Estimator
→ Output/Review
                   ↑                                                                ↓
              Lexical Resources ←←←←←←←← Human Post-Editing (Feedback Loop) ←←←←←←←←
```

3. EXPECTED RESULTS

3.1 Input Specifications

Input Type	Description
Source Documents	XML, PDF manuals, Text documents in English
Target Languages	Hindi, Tamil, Telugu, Malayalam, Kannada (extendable)
Domain Lexicons	Bilingual automotive terminology glossaries (CSV/JSON)
Named Entity Lists	Part numbers, model names, brand names for preservation
Parallel Corpora	Previously translated manuals for model fine-tuning

Phase-I	Phase-I	**Phase-II**
1. Hindi **2. Tamil**	**3. Telugu** **4. Kannada** **5. Malayalam** **6. Bengali** **7. Punjabi** **8. Gujarati** **9. Odia** 10. **Marathi** 11. Assamese 12. Kashmiri 13. Urdu **14.** Sanskrit **15. Native French** **16. European Spanish** **17. Bahasa (Indonesian)** **18. Bahasa (Malay)**	1. US English 2. Canadian French 3. Belgian French 4. Swiss French 5. German 6. Italian 7. Dutch 8. Portuguese 9. Swedish 10. Polish 11. Danish 12. Arabic 13. Japanese 14. Chinese 15. Korean 16. Thai 17. Vietnamese 18. Latin Spanish

3.2 Expected Output & Quality Metrics

PROJECT KPI'S

Performance metrics will be evaluated based on:
- System Accuracy and Response Time
- Successful Retrieval Rate
- Document Conversion Efficiency
- User Adoption and Engagement Metrics

Metric	Target
Translation Accuracy	≥ 90% for domestic, ≥ 85% for foreign languages
Terminology Consistency	≥ 95%
XML Structural Integrity	100% validated
Processing Speed	≤ 60 seconds per file
Retraining Cycle	Every 2 weeks post-feedback integration

4. TENTATIVE TECHNOLOGY STACK (For Sample)

Category	Technology	Purpose
Base NMT Model	IndicTrans2 / NLLB-200	Multilingual translation for 22 Indic languages
Deep Learning Framework	PyTorch / Hugging Face Transformers	Model fine-tuning and inference
NER Component	spaCy / Stanza / Custom BERT-NER	Named entity extraction and preservation
Quality Estimation	COMET / COMET-QE	Translation quality scoring
Evaluation Metrics	SacreBLEU, TER, chrF	Standard MT evaluation metrics
Backend API	FastAPI / Flask	REST API for translation services
Cloud Platform	Google Cloud Platform (Vertex AI)	Model training, deployment, MLOps
Document Processing	lxml, BeautifulSoup, PDFPlumber	XML/PDF parsing and reconstruction
Data Storage	PostgreSQL / Cloud SQL	Glossaries, translation memory, logs
MLOps	MLflow / Kubeflow / Vertex AI Pipelines	Experiment tracking, model versioning

5. PROJECT TIMELINE

The project is planned across four phases over a 3-month duration:

Phase	Duration	Key Deliverables
Phase 1	Month 1	**Data Collection & Preparation:** Parallel corpora curation, glossary development, NER lexicon creation
Phase 2	Month 2	**Model Development:** Base model selection, fine-tuning on automotive domain, NER integration
Phase 3	Month 3	**System Integration:** Pipeline development, API creation, quality estimation integration, UI development
Phase 4	Month 4	**Testing & Deployment:** UAT, performance optimization, production deployment, documentation

6. REFERENCES & RESOURCES

6.1 Research Papers & Publications

- Gala, J., et al. (2023). "IndicTrans2: Towards High-Quality and Accessible Machine Translation Models for all 22 Scheduled Indian Languages." Transactions on Machine Learning Research.
- Costa-jussà, M.R., et al. (2022). "No Language Left Behind: Scaling Human-Centered Machine Translation." Meta AI Research.

- Koehn, P. & Knowles, R. (2017). "Six Challenges for Neural Machine Translation." ACL Workshop on Neural Machine Translation.
- Rei, R., et al. (2020). "COMET: A Neural Framework for MT Evaluation." EMNLP.

7. CONCLUSION

The proposed NLP-Based Multilingual Machine Translation System addresses critical gaps in automotive technical documentation translation by combining state-of-the-art neural machine translation with domain-specific customization. By leveraging open-source models like IndicTrans2, implementing robust terminology enforcement, and establishing continuous learning loops, the system aims to deliver translation quality that meets or exceeds commercial alternatives while significantly reducing time-to-market and operational costs.

The modular architecture ensures extensibility to additional languages and domains, while the quality estimation component provides transparency and control over translation output. This initiative represents a strategic investment in AI-driven automation that aligns with organizational digital transformation objectives and positions the organization as a leader in multilingual technical communication within the automotive sector.

Prepared By: Dr Karthikeyan S ______________________ Date: _______________	**Approved By:** ______________________ Date: _______________