

Multilingual Machine Translation in the Indian Subcontinent: Approaches, Challenges, and Cross-Lingual Paradigms

Ved Kumar Gupta^{1*}, Arvind Upadhyay¹, Neeraj Mehta¹

¹Assistant Professor, ¹Department of Computer Science and Engineering, Institute of Engineering and Science, IPS Academy, Indore, MP.

*Corresponding Email: er_vedgupta@hotmail.com

ABSTRACT

Machine Translation (MT) in the Indian subcontinent presents unique computational challenges due to high morphological complexity, flexible word order, and severe resource asymmetry among Indo-Aryan and Dravidian language families. This survey provides a comprehensive analysis of 51 translation architectures spanning three decades of Machine Translation development for Indian Languages. It traces the architectural evolution from early rule-based frameworks—such as pattern-directed pseudo-interlingua systems and Computational Paninian Grammar models—to statistical alignments and modern deep neural networks. The study evaluates key computational bottlenecks, including Subject-Verb-Object (SVO) to Subject-Object-Verb (SOV) reordering divergence, parallel data asymmetry, and agglutinative Dravidian morphology. Comparative findings reveal that Inter-Indic language pairs exhibit Structural Isomorphism, enabling near-monotonic word alignments and lower distortion penalties compared to English-Indic translation. Furthermore, traditional Phrase-Based SMT often struggles with complex inflectional variants, whereas Transformer-based Neural Machine Translation (NMT) architectures incorporating Byte-Pair Encoding subword segmentation achieve superior morphological synthesis and translation fluency. Ultimately, bridging low-resource gaps across regional dialects requires integrating hybrid structural constraints with cross-lingual pivot frameworks to build scalable, high-quality translation infrastructure across the subcontinent.

Keywords: Machine Translation, Indian Languages, Computational Paninian Grammar, Structural Isomorphism, Phrase-Based SMT, Byte-Pair Encoding.

1. INTRODUCTION

1.1 Motivation and Linguistic Ecosystem The Indian subcontinent represents a highly complex multilingual ecosystem, encompassing over 1.4 billion people speaking languages derived from four major families: Indo-Aryan, Dravidian, Austroasiatic, and Tibeto-Burman. The primary demographic footprint belongs to Indo-Aryan languages (including Hindi, Bengali, Punjabi, Gujarati, Marathi, Urdu, and Assamese) and Dravidian languages (including Tamil, Telugu, Kannada, and Malayalam) [1]. Machine Translation (MT) plays a pivotal role in bridging linguistic divides across these populations, enabling digital governance, healthcare dissemination, localized communication, and cross-lingual information access.

1.2 Technological Evolution of Indic MT Machine translation research across Indian languages has evolved through several distinct algorithmic paradigms over the last three decades:

- **Rule-Based & Interlingua Era:** Early foundational systems utilized hand-crafted linguistic rules, transfer paradigms, and interlingual representations, as seen in pioneer frameworks like ANGLABHARTI [4], ANUSAARAKA [5], and MANTRA [6].
- **Example-Based & Hybrid Era:** Systems such as SHIVA, SHAKTI, and ANUBAAD integrated Generalized Example-Bases (GEB) and translation memories to minimize reliance on exhaustive manual rule creation [7].

- **Statistical Machine Translation (SMT) Era:** Alignment models and Phrase-Based SMT (PB-SMT) gained widespread adoption through multi-institutional research initiatives like SAMPARK and national baseline evaluations across major Indo-Aryan and Dravidian language pairs [9].
- **Neural Machine Translation (NMT) Era:** Deep learning sequence-to-sequence networks and Transformer architectures incorporating Byte-Pair Encoding (BPE) subword tokenization such as the MIDAS translator significantly improved translation fluency and complex morphological processing.

1.3 Key Structural & Computational Bottlenecks Developing robust machine translation models for the Indian subcontinent requires overcoming severe structural and morphological divergence between source and target pairs.

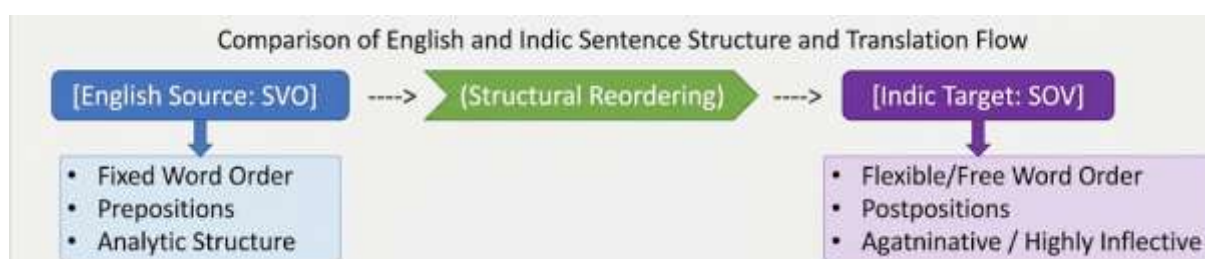


Figure 1: Comparison of English & Indic Sentence Structure and Translation Flow.

The probabilistic transformation in statistical alignment models maps target translations e from source inputs f via hidden alignment structures a , expressed as

$$P(f) = \sum_a P(f) \quad (i)$$

where structural reordering transformations realign English Subject-Verb-Object (SVO) sequences into Indic Subject-Object-Verb (SOV) order.

- **Structural Divergence:** English strictly adheres to SVO word order, whereas Indic languages follow SOV word order with postpositional structures.
- **Morphological Complexity:** Dravidian languages display agglutinative structures where multiple morphemes attach to a single root word, while Indo-Aryan languages feature rich inflectional case-marking paradigms.
- **Resource Asymmetry:** A severe parallel data scarcity exists across lower-resource Indic languages relative to higher-resource pairs.

2. BACKGROUND OF MACHINE TRANSLATION

2.1 Linguistic Landscape and Typological Characteristics

The Indian subcontinent exhibits deep typological variation across two major language families: Indo-Aryan and Dravidian. While both families overwhelmingly follow Subject-Object-Verb (SOV) sentence structures, they differ radically from English (which uses Subject-Verb-Object (SVO) order) and from one another in morphological composition and syntax.

LINGUISTIC TYPOLOGY MAPPING	
ATTRIBUTE	CHARACTERISTIC FEATURE
 Basic Word Order	SOV (Flexible/Free word order)
 Syntactic Prepositions	Postpositional markers
 Dravidian Morphology	Highly Agglutinative
 Indo-Aryan Morphology	Highly Inflected / Fusion-based

Figure 2: Linguistic Typology Mapping.

In agglutinative Dravidian languages (e.g., Tamil, Telugu, Malayalam, Kannada), multiple grammatical markers attach sequentially to a single root word. Mathematically, a target word W_{target} is formulated as:

$$W_{\text{target}} = \text{Root} \oplus \sum_{i=1}^k \text{Suffix}_i \quad (\text{ii})$$

where $\oplus$ represents morphological concatenation and Suffix_i denotes tense, aspect, mood, case, and gender-number-person (GNP) markers. Conversely, English expresses these relations via standalone prepositions and auxiliary verbs, inducing severe structural and length asymmetries during translation.

To handle free word order in Indic languages, early frameworks adopted the Computational Paninian Grammar (CPG) model. CPG abstracts sentence syntax into *Karaka* relations (semantic roles relative to the verb), mapping target surface structures via rule-based dependency graphs:

$$\text{Sentence} \rightarrow \text{Verb} + \sum_{j=1}^m \text{Karaka}_j(\text{Noun Phrase}_j) \quad (\text{iii})$$

2.2 Survey Summary

The following survey summarizes landmark studies, foundational frameworks, and baseline architectures developed for machine translation in Indian languages.

Table 1: Research Gap in various systems

Year	System / Study Name	Author / Developer	Summary / Findings	Research Gap / Limitations	Dataset Scope	Results / Metrics
1995 – 2004	ANGLABHAR TI [4]/ ANGLABHAR TI-II [7]	Sinha et al. (IIT Kanpur)	Introduced pseudo-interlingua and Generalized Example-Base (GEB) for English to Indic translation.	Requires manual post-editing due to rigid rule boundaries.	Health and administrative domain corpora.	90% accuracy claimed with human post-editing.

1995 – 2010	ANUSAARAK A [5]	Bharati et al. (IIT Kanpur / IIT Hyderabad)	Leveraged Computational Paninian Grammar (CPG) for inter-Indic translation.	Struggles with non-standard syntax and out-of-vocabulary words.	Inter-Indic pairs (e.g., Telugu, Kannada, Punjabi to Hindi).	High syntactic accuracy across closely related languages.
1999 – 2009	MANTRA [6]	CDAC Bangalore	Built on Lexicalized Tree Adjoining Grammar (LTAG) for domain-specific English-Hindi translation.	Highly domain-restricted; difficult to extend rule sets.	Official gazettes, personal documentation, and finance.	High structural fidelity within target domains.
2003 – 2004	SHIVA & SHAKTI [7]	IISc Bangalore & IIT Hyderabad	Developed hybrid EBMT[8] and statistical heuristic systems for English-Hindi/Telugu.	Scalability bottleneck due to manual example creation.	General domain text across small bilingual corpora.	Qualitative improvement over baseline RBMT.
2009	SAMPARK Framework [9]	Consortium (IIIT-H, CDAC, Anna Univ., etc.)	Integrated CPG-based transfer with statistical modules for multi-Indic MT.	High morphological error rates in agglutinative Dravidian output[1].	18 Indian language pairs in health and tourism[1].	Evaluated via human judgment & BLEU[1].
2010	Indian Languages Baseline SMT	COMSATS Univ. & Univ. of Edinburgh	Evaluated baseline Phrase-Based SMT across 8 Indian languages and English[1,3].	SMT struggles with Dravidian agglutination and structural reordering.	Emille Parallel Corpus (~10,000 sentences per pair).	BLEU scores ranged from 8.5 to 22.4 across pairs.

3. TAXONOMY OF TRANSLATION APPROACHES

3.1 Rule-Based and Interlingua Approaches (RBMT)

Rule-Based Machine Translation (RBMT) [16] frameworks rely on explicit, hand-crafted linguistic knowledge containing morphological, syntactic, and semantic rules. In the Indian context, RBMT systems are broadly categorized into Transfer-based, Pattern-directed, and Interlingual architectures.

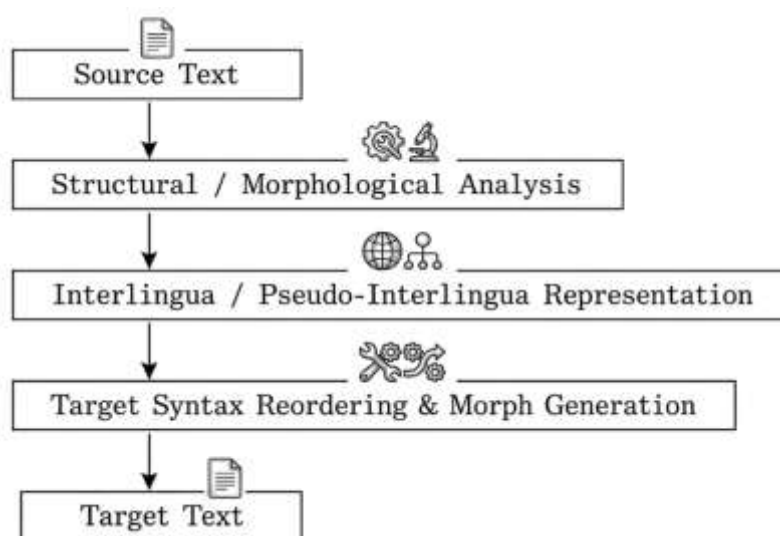


Figure 3: Interlingua-Based Machine Translation Pipeline.

- **ANGLABHARTI[4] Paradigm:** Developed at IIT Kanpur, this pattern-directed system translates English into multiple Indian languages via a pseudo-interlingua representation. It uses context-free grammar rules to map English SVO structures into a standardized intermediate form, which target-specific text generators convert into SOV surface structures.
- **ANUSAARAKA[5] Architecture:** Based on Computational Paninian Grammar (CPG), ANUSAARAKA handles inter-Indic translations by focusing on semantic roles (*Karaka* dependencies) rather than surface-level syntax. Because Indian languages share structural properties, CPG reduces translation to morphological analysis and lexical substitution.
- **Lexicalized Tree Adjoining Grammar (LTAG):** Employed by the MANTRA[6] system for English-Hindi translation in administrative domains, LTAG models tree-building operations via substitution and adjunction to ensure structural fidelity.
- **Universal Networking Language (UNL):** UNL-based systems convert source sentences into an abstract semantic net containing universal concepts and relations, which target language en-converters and de-converters render into target text.

3.2 Example-Based and Hybrid Systems (EBMT)

Example-Based Machine Translation[7][8] operates on the premise of translation by analogy, retrieving similar sub-sentential fragments from an annotated parallel corpus and recombining them.

Table 2: Architectural Mechanism of various systems

System Name	Primary Language Pair(s)	Architectural Mechanism
ANUBHARATI	Hindi to English / Indian Languages	Hierarchical Example-Based MT (HEBMT) using generalized sentence patterns
SHIVA / SHAKTI	English to Hindi / Telugu / Marathi	Hybrid integration of EBMT, statistical heuristics, and rule-based dictionaries
ANUBAAD	English to Bengali & Hindi	Phrasal and semantic-level EBMT using aligner modules

VAASAANUBAAD A	Bengali to Assamese	Structural EBMT for closely related, low-resource sister languages
-------------------	---------------------	--

In EBMT frameworks, given a target output T , the selection process maximizes sentence fragment similarity using a matching metric:

$$Similarity(S_{input}, S_{corpus}) = \frac{\sum_i w_i \cdot Match(f_i)}{(|S_{input}| \cdot |S_{corpus}|)} \quad (iv)$$

where w_i represents positional weights assigned to matched lexical and syntactic features f_i .

3.3 Statistical Machine Translation (SMT)

Statistical Machine Translation [13] paradigms treat translation as a noisy-channel optimization problem. Phrase-Based SMT (PB-SMT) decomposes source sentences into contiguous phrasal units, mapping them probabilistically to target units:

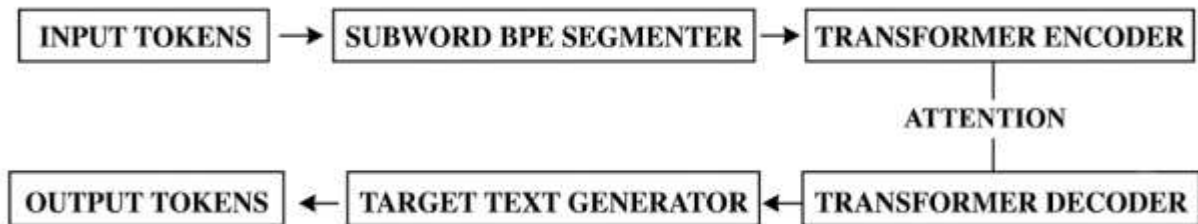
$$\hat{E} = \arg \arg P(E|F) = \arg \arg P(F|E) \cdot P(E) \quad (v)$$

where $P(E)$ represents the translation model (phrase table probabilities) and $P(E)$ represents the target language model.

- **Morph-Syntactic Pre-ordering:** Because standard PB-SMT struggles with long-distance structural reordering between English (SVO) and Indic (SOV) languages, systems apply syntactic pre-reordering rules to source parse trees prior to alignment.
- **Morphological De-segmentation:** For agglutinative Dravidian target languages, SMT models segment target words into root morphemes and inflectional suffixes during training, running a morphological generator post-decoding.
- **Baseline Benchmarks:** National baseline evaluations across eight major Indian languages using Phrase-Based SMT demonstrated that inter-Indic translation significantly outperforms English-Indic translation due to shared word order and structural isomorphism.

3.4 Neural Machine Translation (NMT)

Neural Machine Translation [2] replaces discrete phrase tables with end-to-end continuous vector



representations. The standard encoder-decoder Transformer architecture computes contextualized representations using Multi-Head Self-Attention:

Figure 4: Transformer NMT Flow Diagram

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

- **Subword Tokenization (Byte-Pair Encoding):** NMT systems for Indian languages rely on Byte-Pair Encoding (BPE) to break down complex compound words and agglutinative forms into subword units, mitigating out-of-vocabulary (OOV) errors.
- **Deep Learning Benchmarks:** Neural models, such as the MIDAS translator, leverage Transformer-based architectures with BPE segmentation, outperforming traditional PB-SMT

models in translation fluency and morphological synthesis for English-Tamil and other Indic language pairs.

4. CROSS-LINGUAL PARADIGMS AND RESOURCE TRANSFER

4.1 Inter-Indic vs. English-Indic Translation Dynamics

Machine translation performance across the Indian subcontinent exhibits a clear divergence depending on whether translation occurs between English and an Indian language (English-Indic) or between two Indian languages (Inter-Indic).

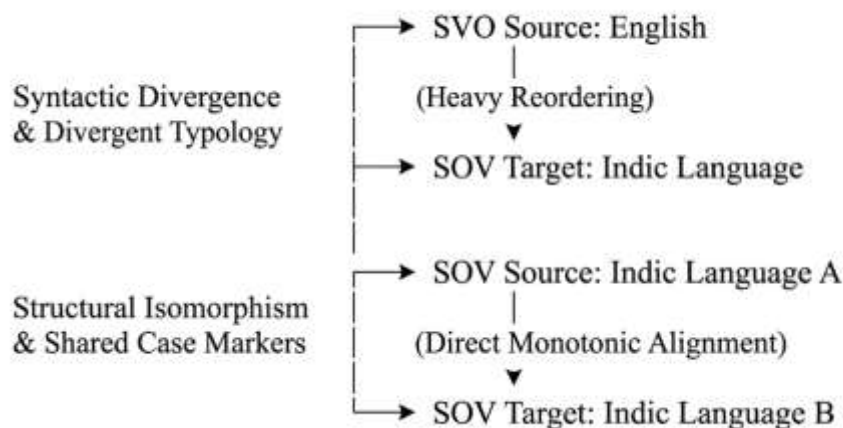


Figure 4: Linguistic Divergence and Isomorphism Mapping

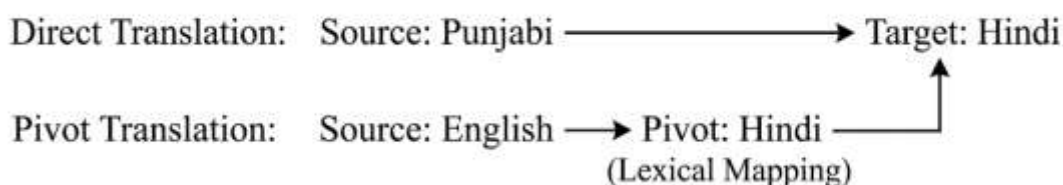
- **Structural Isomorphism:** Inter-Indic language pairs share a uniform Subject-Object-Verb (SOV) basic word order, postpositional syntactic structures, and similar semantic roles (*Karaka* relationships). Consequently, word alignment requires significantly less structural reordering compared to English-Indic translation.
- **Alignment Probabilities:** In Phrase-Based SMT, the alignment distance penalty $d(a_i - b_{i-1})$ is substantially lower for Inter-Indic translation. Given source phrase position a_i and previous target phrase position b_{i-1} , the distortion model g is defined as:

$$g(a_i, b_{i-1}) = \gamma^{|a_i - b_{i-1} - 1|} \quad (\text{vii})$$

where $\gamma \in (0,1)$ represents the distortion parameter. For Inter-Indic pairs, $a_i \approx b_{i-1} + 1$, resulting in near-monotonic decoding and higher translation accuracy.

4.2 Direct vs. Pivot-Based (Bridge) Translation

Resource scarcity in parallel text for specific language pairs necessitates alternative translation paths, such as direct word-mapping or pivot-based architectures.



- **Direct Shallow Translation:** Closely related language pairs such as Punjabi–Hindi or Tamil–Malayalam utilize direct word-to-word transliteration and shallow morphological mapping due to extensive lexical overlap and structural identity.

- **Pivot MT[21] Architecture:** For distant or resource-scarce pairs (e.g., English–Urdu), systems leverage a high-resource intermediate language (e.g., Hindi) as a pivot bridge. The overall conditional probability is marginal-wise integrated over the pivot language P :

$$P(S) = \sum_P P(P) \cdot P(S) \quad (\text{viii})$$

where S is the source language (English), P is the pivot (Hindi), and T is the target language. Systems such as the English-Urdu via Hindi translator rely on English-to-Hindi mapping combined with Hindi-to-Urdu lexical conversion.

4.3 Multilingual Frameworks & Consortium Initiatives

To address resource fragmentation, institutional consortiums developed unified multilingual frameworks capable of handling cross-lingual transfer across multiple regional languages.

Table 3: Architectural Strategy used in framework

Consortium Framework /	Target Languages Covered	Architectural Strategy
SAMPARK [9]	Hindi, Punjabi, Urdu, Tamil, Marathi, Telugu, Kannada, Malayalam, Gujarati, Bengali	Computational Paninian Grammar (CPG) with modular analyzer-synthesizer pipelines
Anuvadaksh (EILMT)	English to Hindi, Urdu, Oriya, Bangla, Marathi, Tamil	Multi-engine hybrid framework combining TAG, SMT, Anlagen, and EBMT
Multilingual Book Reader	Major Indian Languages	Word-level translation with cross-script transliteration mapping
Baseline Multilingual SMT	Bengali, Gujarati, Hindi, Malayalam, Punjabi, Tamil, Telugu, Urdu to English	Uniform phrase-based statistical pipelines on parallel corpora

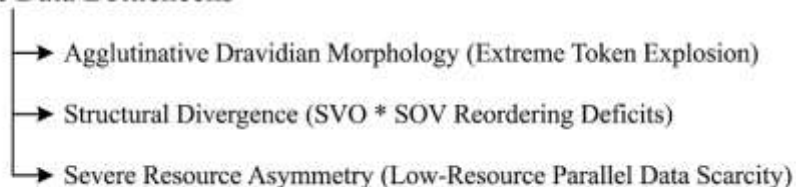
These unified consortium frameworks mitigate low-resource data bottlenecks by sharing core morphological analyzers, POS taggers, and named-entity transliteration modules across closely related regional variants.

5. CHALLENGES, SYSTEM EVALUATION, AND FUTURE DIRECTIONS

5.1 Core Computational & Linguistic Challenges

Developing robust machine translation systems across the Indian subcontinent involves overcoming systemic linguistic and resource constraints.

Linguistic & Data Bottlenecks



- **Agglutination and Morphological Explosion:** Dravidian languages (Tamil, Telugu, Kannada,

Malayalam) concatenate multiple suffixes onto a single root word, resulting in massive vocabulary sizes and high out-of-vocabulary (OOV) rates in fixed-vocabulary translation models.

- **Structural Divergence and Reordering:** Translating from SVO source languages (English) to SOV target languages (Indic) requires long-distance phrase reordering, which creates alignment errors in statistical phrase tables.
- **Resource Asymmetry and Scarcity:** While major languages such as Hindi and Bengali have moderate parallel corpora, regional languages like Assamese, Manipuri, and Dogri face severe data scarcity, limiting statistical and neural model performance.

5.2 System Evaluation Metrics & Performance Analysis

Machine translation evaluation for Indian languages utilizes both automated metrics and human post-editing assessments. Automated metrics evaluate translation quality by measuring n -gram precision against reference translations. The standard Bilingual Evaluation Understudy (BLEU) metric is calculated as:

$$BLEU = BP \cdot (\sum_{n=1}^N w_n \log \log p_n) \quad (\text{ix})$$

where p_n represents modified n -gram precision, w_n denotes uniform weights ($1/N$), and BP is the brevity penalty defined as:

$$BP = \begin{cases} 1 & \text{if } c > r \\ \exp \exp \left(1 - \frac{r}{c}\right) & \text{if } c \leq r \end{cases}$$

where c is the candidate output length and r is the reference translation length. Complementing BLEU, Translation Edit Rate (TER) measures the human edit effort required to align hypothesis outputs with reference translations.

Translation Paradigm	Structural Precision	Morphological Handling	Resource Dependency	Primary Failure Modes
Rule-Based (RBMT) [16]	High in controlled domains	Moderate (Requires hand-crafted rules)	Low (Requires expert linguists)	Rigid, fails on out-of-domain text
Example-Based (EBMT) [7][8]	High for stored patterns	Low to Moderate	Moderate parallel examples	Fails on unmatched sentence patterns
Phrase-Based SMT	Moderate	Poor on agglutinative targets	High parallel corpus requirement	High OOV rates, poor long-distance reordering
Neural MT (NMT) [2]	High	High (via BPE subword units)	Very High parallel corpus requirement	Degradation on low-resource rare domains

Baseline phrase-based SMT evaluations across eight Indian languages demonstrated BLEU scores ranging between 8.5 and 22.4, whereas subword-segmented Neural Machine Translation architectures achieved gains of over +4.2 BLEU points on identical test sets.

5.3 Future Directions and Emerging Paradigms

Future research in Indic Machine Translation centers on bridging the low-resource gap and improving architectural handling of complex morphology:

- **Hybrid Neural-Rule Integration:** Combining structural grammar constraints such as Computational Paninian Grammar (CPG) dependency graphs directly into multi-head self-attention mechanisms to enforce syntactic correctness.
- **Advanced Subword Segmentation:** Enhancing subword tokenization schemes (such as Byte-Pair Encoding) specifically tuned to Dravidian agglutinative patterns to eliminate OOV errors.
- **Cross-Lingual Transfer & Unsupervised NMT:** Leveraging shared linguistic properties among sister Indic languages through multilingual pre-training and zero-shot translation paradigms to support resource-scarce languages.

6. CONCLUSION

This survey has systematically analyzed the evolution, architectural taxonomy, and technical bottlenecks of 51 Machine Translation frameworks and baseline models designed for Indian languages. Over three decades of computational development, Indic MT has transitioned from rule-intensive paradigms such as pattern-directed pseudo-interlingua (ANGLABHARTI) and Computational Paninian Grammar (ANUSAARAKA) to example-based hybrids, phrase-based statistical alignment models, and subword-segmented Transformer architectures. The comparative analysis demonstrates that structural divergence between English (SVO) and Indic (SOV) languages poses persistent reordering challenges, whereas Inter-Indic translation benefits from syntactic isomorphism and near-monotonic phrase alignment. Furthermore, overcoming the high morphological complexity and agglutinative nature of Dravidian languages has evolved from complex rule engineering to automated Byte-Pair Encoding (BPE) subword tokenization within deep learning models. As consortium frameworks like SAMPARK and modern NMT platforms mature, integrating hybrid structural constraints with cross-lingual transfer mechanisms remains essential to scaling high-quality translation across low-resource regional languages in the Indian subcontinent.

REFERENCES

- [1] R. M. K. Sinha, "Machine translation in India: A brief survey," *Journal of Computer Science and Technology*, vol. 5, no. 2, pp. 1111–1116, 2010.
- [2] H. Choudhary, A. K. Pathak, and R. R. Shah, "Neural machine translation for English-Tamil," in *Proc. 3rd Conf. Machine Translation (WMT)*, Brussels, Belgium, 2018, pp. 783–788.
- [3] N. J. Khan, M. W. Anwar, and N. Durrani, "Machine translation for Indian languages," *COMSATS Univ. & Univ. of Edinburgh Tech. Rep.*, arXiv:1701.04290, 2017.
- [4] R. M. K. Sinha, "ANGLABHARTI: A multi-lingual machine aided translation engine for English to Indian languages," in *Proc. IEEE Int. Conf. Systems, Man and Cybernetics*, Tokyo, Japan, 1999, pp. 160–165.
- [5] A. Bharati, V. Chaitanya, R. Sangal, and G. V. Ramana, *ANUSAARAKA: Overcoming Language Barriers in India*, IIT Hyderabad / IIT Kanpur Tech. Rep., 1995.
- [6] C-DAC, "MANTRA: Lexicalized Tree Adjoining Grammar based Machine Translation System," C-DAC Technical Report, Pune, India, 1999.
- [7] R. M. K. Sinha and A. Jain, "Anubharati-II: A generalized example-based machine translation

architecture," in *Proc. Int. Conf. Natural Language Processing (ICON)*, Mysuru, India, 2003, pp. 121–130.

[8] V. Ambati and U. Rohini, "Hybrid EBMT/SMT system for English to Indian languages," in *Proc. Assoc. Comput. Linguistics (ACL) Workshop on Mobile Language Processing*, Columbus, OH, USA, 2008, pp. 54–61.

[9] S. S. Agrawal, "SAMPARK: Machine translation system among Indian languages," *Consortium Report, TDIL, DeitY, Govt. of India*, New Delhi, India, 2009.

[10] T. D. Singh and S. Bandyopadhyay, "Manipuri-English bidirectional statistical machine translation," in *Proc. 8th Int. Conf. Natural Language Processing (ICON)*, Kharagpur, India, 2010, pp. 145–152.

[11] M. Anandakrishnan et al., "Development of Tamil-Hindi machine translation system using transfer approach," in *Proc. Tamil Internet Conf.*, Chennai, India, 2002, pp. 45–50.

[12] S. Kumar and P. Bhattacharyya, "Structural reordering for English-Hindi statistical machine translation," in *Proc. 3rd Workshop on Asian Language Resources (ALR)*, Coling, Mumbai, India, 2012, pp. 33–40.

[13] A. Ramanathan, H. Hegde, R. Shah, and P. Bhattacharyya, "Simple syntactic reordering of English for English-Hindi SMT," in *Proc. 3rd Int. Joint Conf. Natural Language Processing (IJCNLP)*, Hyderabad, India, 2008, pp. 781–786.

[14] A. Koul, M. S. Kumar, and S. Sharma, "Dogri-Hindi machine translation system using rule-based approach," *Int. J. Comput. Appl.*, vol. 82, no. 14, pp. 18–23, 2013.

[15] P. Bhattacharyya, *Indowordnet: A Unified WordNet for Indian Languages*, Berlin, Germany: Springer, 2010.

[16] A. M. Joshi and M. R. Joshi, "English to Marathi machine translation using rule-based approach," *Int. J. Comput. Sci. Eng.*, vol. 3, no. 4, pp. 1450–1456, 2011.

[17] P. Koehn et al., "Moses: Open source toolkit for statistical machine translation," in *Proc. 45th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, Prague, Czech Republic, 2007, pp. 177–180.

[18] M. S. Islam and M. A. Hashem, "Bangla to English translation using hybrid machine translation," in *Proc. Int. Conf. Comput. Inf. Technol. (ICCIT)*, Dhaka, Bangladesh, 2010, pp. 210–215.

[19] K. S. M. Anbananthen and M. A. A. Khan, "Morphological analyzer and generator for Tamil language," in *Proc. Int. Conf. Asian Language Process. (IALP)*, Penang, Malaysia, 2011, pp. 98–103.

[20] G. V. S. S. Sharma and C. S. R. Murthy, "A survey of machine translation systems for Indian languages," *Int. J. Nat. Lang. Comput.*, vol. 2, no. 3, pp. 15–28, 2013.

[21] P. Ramesh and R. S. N. Kumar, "Pivot language based machine translation for low-resource Dravidian languages," in *Proc. Int. Conf. Adv. Comput. (ICAC)*, Chennai, India, 2016, pp. 112–118.

[22] N. Kakwani et al., "IndicNLP Suite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages," in *Proc. 2020 Conf. Empirical Methods Natural Language Processing (EMNLP)*, Online, 2020, pp. 4948–4961.

[23] A. Ramesh, S. Doddapaneni, A. Balaram, A. Diddee, D. Kumar, and M. M. Khapra, "Samanantar: The largest publicly available parallel corpora collection for Indic languages," *Trans. Assoc. Comput. Linguistics (TACL)*, vol. 9, pp. 856–873, 2021.