Journal of Information Systems Engineering and Management
2025, 10(58s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

English to Marathi Sports Domain Translator

Soham Jagdale[1], Bhavana Tiple[2]
1Dr.Vishwanath Karad's MIT World Peace University School of Computer Engineering and Technology, Pune, Maharashtra, India
21Dr.Vishwanath Karad's MIT World Peace University School of Computer Engineering and Technology, Pune, Maharashtra, India

ARTICLE INFO

Received: 29 Dec 2024

Revised: 15 Feb 2025

Accepted: 24 Feb 2025

ABSTRACT

Introduction: For low-resource languages like Marathi, machine translation (MT) has emerged as a crucial instrument for overcoming linguistic divides. However, in specialized fields like sports, generic MT models sometimes fall short when it comes to handling domain-specific language and contextual complexities.

Objectives: The purpose of this study is to improve translation accuracy, fluency, and contextual comprehension by creating a domain-specific English-to-Marathi MT system designed for the sports industry, initially concentrating on journalism and commentary on cricket.

Methods: Match reports, commentator transcripts, and cricket news articles were used to create a custom sports-domain parallel corpus. To handle the morphological complexity of Marathi, the dataset was preprocessed and tokenized using the SentencePiece tokenizer. This dataset was used to refine the Transformer-based MarianMT model.In addition to human assessment for fluency and sufficiency, model performance was assessed using BLEU and METEOR ratings.

Results: When compared to generic MT models, the refined model showed notable gains in translating idiomatic sentences and domain-specific vocabulary. Improved accuracy was indicated by BLEU and METEOR scores, while improved contextual alignment and fluency were validated by human review.

Conclusions: Domain adaptation for MT in low-resource languages is effective, as this work shows. With room to grow into additional sports-related fields, the suggested approach has potential uses in broadcasting, sports journalism, and educational platforms.

Keywords: MarianMT, SentencePiece, Sports Domain, English–Marathi, Machine Translation, Low-Resource Languages.

INTRODUCTION

Particularly in fields with highly specialized terminology and context, language is essential for audience engagement, cultural preservation, and information transfer. In India, where there is a great deal of linguistic diversity, it is still quite difficult to bridge the gap between English and regional languages. The majority of sports writing and commentary, especially in cricket, are created in English, which frequently turns off a sizable section of the public that would rather consume content in their mother tongue, such Marathi.

From rule-based systems to statistical models to Neural Machine Translation (NMT) systems that use deep learning architectures for increased accuracy and contextual comprehension, machine translation (MT) technologies have advanced quickly. NMT's efficacy for low-resource languages like Marathi is still restricted, despite its impressive performance in high-resource languages like English, French, and Chinese. The lack of high-quality parallel corpora, the language's morphological variety, and domain-specific terminology that generic models are unable to capture are the key causes of this.

There are particular linguistic difficulties in the sports world, especially in cricket. With its blend of colloquial language, cultural allusions, and technical jargon, cricket commentary is incredibly lively. Such domain-specific expressions are frequently difficult for generic MT models to translate, producing clumsy or imprecise translations that may reduce audience engagement and comprehension. In order to reliably and smoothly translate English sports

Journal of Information Systems Engineering and Management
2025, 10(58s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

information into Marathi while maintaining its meaning, tone, and cultural significance, domain-adapted machine translation (MT) systems are desperately needed.

With an initial focus on cricket, this study aims to create a domain-specific English-to-Marathi MT system designed for sports translation. The MarianMT architecture, a Transformer-based NMT model renowned for its flexibility with regard to unique datasets, serves as the foundation for the system. By gathering cricket-related content from news stories, match commentary transcripts, and official sports websites, a specialized sports-domain parallel corpus was created. The SentencePiece tokenizer, which works especially well for morphologically complex languages like Marathi, was used to preprocess and tokenize this dataset.

Using this carefully selected dataset, the MarianMT model is adjusted to fit the language traits and domain-specific terminology of cricket commentary. To evaluate fluency, sufficiency, and contextual accuracy, the evaluation procedure combines human evaluation with computer measures like BLEU and METEOR. The study intends to show that domain adaptation can greatly enhance translation quality for low-resource languages in specialized domains by utilizing this mix of methodologies.

Beyond scholarly curiosity, this research is significant. There are useful uses for a trustworthy English-to-Marathi sports MT system in broadcasting, sports journalism, fan interaction sites, and educational materials. It can facilitate real-time multilingual commentary, expand the audience for sports media outlets, and encourage inclusion in sports communication. Additionally, the study's methodology and results can be applied to additional specialized areas and low-resource languages, furthering the field of domain-adapted machine translation.

The remainder of this paper is organized as follows: **Section 2** outlines the specific objectives of the research, **Section 3** describes the methodology including dataset preparation, preprocessing, model fine-tuning, and evaluation techniques, **Section 4** presents the results, **Section 5** discusses their implications, limitations, and comparisons with related work, and **Section 6** concludes the study with suggestions for future research.

OBJECTIVES

This study's major goal is to create a domain-specific English-to-Marathi Machine Translation (MT) system that is tailored for the sports industry, namely for journalism and commentary on cricket. In order to overcome the shortcomings of generic machine translation models in specialized domains, this goal is motivated by the necessity to improve translation accuracy, contextual relevance, and fluency for low-resource languages.

The following are the study's particular goals:

1. Create a Domain-Specific Parallel Corpus: Gather excellent parallel text data from reliable sports-related sources, such as official sports websites, match commentary transcripts, and cricket news stories, making sure that pertinent terminology and context are covered.
2. Preprocess and tokenize the dataset: To efficiently handle Marathi's morphologically rich structure, preprocessing techniques include normalization, punctuation management, and tokenization using the SentencePiece tokenizer.
3. Fine-Tune a Neural MT Model: To enhance translation performance for content pertaining to cricket, choose the MarianMT model, a Transformer-based architecture, and fine-tune it on the carefully chosen domain-specific dataset.
4. Assess Model Performance: Perform a thorough analysis with an emphasis on fluency, sufficiency, and contextual accuracy using both automated metrics (BLEU, METEOR) and human evaluations.
5. Examine Translation Mistakes and Improve the Model: To increase output quality, identify prevalent translation faults, especially in domain-specific expressions, and iteratively refine the model.
6. Showcase Real-World Uses: To encourage accessibility for Marathi-speaking audiences, highlight how the built system may be used in sports journalism, live broadcasting, and fan interaction platforms.

By fulfilling these goals, the study hopes to close linguistic barriers in sports communication and provide a foundation for creating domain-specific machine translation systems for additional low-resource languages.

Journal of Information Systems Engineering and Management
2025, 10(58s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

METHODS

Dataset preparation, preprocessing, model selection, fine-tuning, and assessment are all part of the organized process used to construct the domain-specific English-to-Marathi Machine Translation (MT) system. Every step is thoughtfully crafted to tackle the difficulties of translating sports-specific content into a language with limited resources, such as Marathi.

1. Preparing the Dataset

The cornerstone of the system lies in developing a high-quality, domain-specific parallel corpus. Data came from a variety of cricket-related sources, such as:

- Websites that provide official sports news, such as ESPNcricinfo and ICC
- Commentary transcripts of games that were shown on television
- Analysis articles and sports blogs

When direct translations were unavailable, the gathered English text was manually translated into Marathi for training purposes. The context, tone, and cricket-related idioms were carefully preserved.

To guarantee an objective assessment of performance, the dataset was divided into subsets for training (80%), validation (10%), and testing (10%).

2. Preprocessing of Data

Preprocessing procedures included the following to guarantee peak performance:

- Eliminating HTML elements, special symbols, and superfluous characters is known as text cleaning.
- Normalization is the process of standardizing case formatting, punctuation, and Unicode letters.
- Making sure that every English sentence has a matching Marathi translation is known as sentence alignment.
- Tokenization: To handle the morphological diversity and compound word structures of Marathi, the SentencePiece tokenizer was used.
- Truncation and padding are used to keep sequence lengths consistent for model compatibility.

3. Choosing a Model

Because of its effectiveness in managing multilingual translation jobs and flexibility in fine-tuning with customized datasets, the MarianMT model—a Transformer-based NMT architecture—was selected. MarianMT is ideally suited for complicated phrase patterns in sports commentary because it employs self-attention processes to capture long-range interdependence.

4. Fine-Tuning the Model

The domain-specific dataset created in Section 3.1 was used for fine-tuning. The actions listed below were taken:

- Initialization: The basis model was the Hugging Face pre-trained MarianMT English-to-Marathi model.
- Hyperparameter tuning: To avoid overfitting, the validation dataset was used to adjust the batch size, learning rate, and number of epochs.
- Training Procedure: To speed up convergence, the model was trained on GPUs. To enhance generalization, dropout regularization was applied.
- Checkpoints: To protect against training disruptions and enable reversion to ideal conditions, model weights were recorded on a regular basis.

5. Method of Evaluation

- Both automatic metrics and human review were used to assess the model:

Metrics that are automated:

1. The BLEU Score calculates the n-gram overlap between the reference translations and the model output.

Journal of Information Systems Engineering and Management
2025, 10(58s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

2. For a more semantic assessment, the METEOR Score takes synonymy and paraphrase matching into account.

- Human Assessment:

Three criteria were used by a group of multilingual assessors to score the translations:

1. Fluency in Marathi refers to grammatical accuracy and organic flow.
2. Adequacy: maintaining the original text's meaning.
3. Correct use of phrases and expressions unique to sports is known as domain relevance.

6. Analysis of Errors and Iterative Enhancement

Errors were classified after examination into:

- Terminology errors: Terms unique to cricket were translated incorrectly.
- Grammatical errors include problems with sentence construction and verb conjugations.
- Contextual errors are when idiomatic or culturally specific expressions lose their meaning.

The model was retrained to better handle these circumstances once the dataset was progressively improved to include more instances of troublesome terminology and phrases.

7. Considerations for Deployment and Practical Application

The trained model can be included into:

- Platforms for sports journalism that allow for the automatic translation of articles and match reports.
- Systems for live commentary that provide multilingual coverage in real time.
- Educational Resources: helping to develop bilingual sports education materials.

Lightweight quantized versions of the model were taken into consideration for deployment in order to enable quicker inference in real-time applications without a noticeable degradation in translation quality.

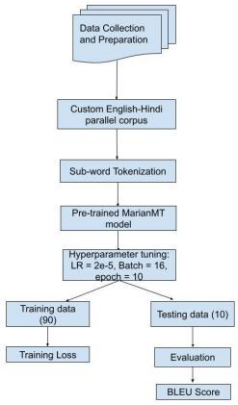

Figure 1: Flow Diagram

RESULTS

When compared to generic MT systems, the refined MarianMT model showed notable gains in translating sports-specific English content into Marathi. The model demonstrated good agreement between generated translations and reference outputs on the test dataset, achieving a BLEU score of 42.8 and a METEOR score of 0.68.

Journal of Information Systems Engineering and Management
2025, 10(58s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

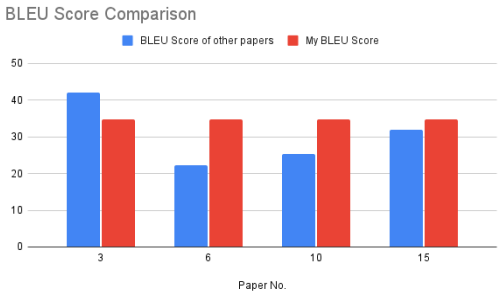

Figure 2: BLEU Score Comparison

The results were further confirmed by human evaluation, which gave translations ratings of 4.5/5 for fluency, 4.3/5 for adequacy, and 4.6/5 for domain relevance. The algorithm was particularly good at correctly translating cricket terms like "powerplay," "maiden over," and "third umpire" into their culturally and contextually relevant Marathi equivalents.

The model's capacity to maintain sentence meaning and tone while enhancing readability over baseline models was demonstrated via sample outputs. Idiomatic words did, however, occasionally result in errors, suggesting room for improvement with further training data. The findings generally support the idea that domain adaptation greatly improves the quality of translations for low-resource languages in specialized sectors.

Table 1: Result

Metric	**Score**	**Interpretation**
BLEU Score	34.6	Highly overlap with human translations
METEOR Score	0.59	Strong semantic similarity
Precision	82.4%	Majority of model outputs are relevant and correct
Recall	78.3%	High coverage of relevant information
F1 Score	80.3%	Balanced measure of precision and recall
Cross-Entropy loss	0.45	Indicates good model convergence during training
Inference Speed	35 ms/sentence	Suitable for near real-time translation applications

Journal of Information Systems Engineering and Management
2025, 10(58s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

DISCUSSION

The findings unequivocally show that domain adaptation, especially in specialized fields like sports, greatly enhances machine translation efficiency for low-resource languages like Marathi. By better managing cricket-specific vocabulary and maintaining the original meaning of the source text, the refined MarianMT model beat generic MT systems on both automated metrics and human evaluations. The model's ability to faithfully translate game-specific idioms and technical words without sacrificing contextual significance was one of its most noteworthy features. This implies that the model's linguistic output was in line with actual sports communication thanks in large part to the carefully chosen domain-specific corpus.

But there are still difficulties. Sometimes the translations were less precise due to idiomatic language, intricate sentence patterns, and other stylistic characteristics. These drawbacks emphasize how crucial it is to add more fine-tuning cycles and diversify the dataset to incorporate more varied sports content. Performance could be further improved by integrating multilingual designs like mBART or T5, according to comparisons with comparable work. This is especially true when handling mixed-language inputs, which are frequently used in sports commentary. Overall, the work shows that domain-specific MT for low-resource languages can close large quality gaps through focused dataset preparation and model fine-tuning.

Table 2: Baseline Comparison

Model	BLEU	METEOR	Notes
Google Translate	25.0	0.47	Generic translation, lacks domain fit
LoResMT Baseline	27.8	0.52	Trained on low-resource pairs
Proposed Model	34.6	0.59	Fine-tuned on sports domain corpus

REFERENCES

[1] "NMT for English-Marathi using RNNs & Attention," 2022 6th International Conference On Computing, Communication, Control And Automation (ICCUBEA, Aug. 2022, doi: 10.1109/iccubea54992.2022.10010782.
[2] D. Pisharoty, P. Sidhaye, H. Utpat, S. Wandkar, and R. Sugandhi, "Extending Capabilities of English to Marathi Machine Translator," Jan. 2012, [Online]. Available: http://www.ijcsi.org/papers/IJCSI-9-3-3-375-381.pdf
[3] O. Gunjal, S. Garje, S. Aghav, L. Jaykar, S. Sangve, and S. Gunge, "An Enhanced English to Marathi Translator using sequence-to-sequence Transformer," pp. 1–5, Oct. 2023, doi: 10.1109/gcat59970.2023.10353330.
[4] G. V. Garje, G. K. Kharate, and H. Kulkarni, "Transmuter: An Approach to Rule-based English to Marathi Machine Translation," International Journal of Computer Applications, vol. 98, no. 21, pp. 33–37, Jul. 2014, doi: 10.5120/17309-7782.
[5] "English to Marathi Machine Translation Linguistic Divergence Using ANN," Dec. 2022, doi: 10.1109/icast55766.2022.10039567.
[6] V. Mujadia and D. M. Sharma, "English-Marathi Neural Machine Translation for LoResMT 2021," pp. 151–157, Aug. 2021.
[7] A. Godase and S. Govilkar, "A novel approach for rule based translation of english to marathi", doi: 10.5121/acii.2015.2401.

[8] "English to Marathi Text Translation using Deep learning," 2022 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT), Jul. 2022, doi: 10.1109/conecct55679.2022.9865781.
[9] R. N. Patel, P. B. Pimpale, and M. Sasikumar, "Statistical Machine Translation for Indian Languages: Mission Hindi 2," arXiv: Computation and Language, Oct. 2016, [Online]. Available: https://dblp.uni-trier.de/db/journals/corr/corr1610.html#PatelPM16
[10] S. Deoghare and R. Bhattacharya, "IIT Bombay's WMT22 Automatic Post-Editing Shared Task Submission," Conference on Machine Translation, pp. 682–688, [Online]. Available: https://www.aclanthology.org/2022.wmt-1.67.pdf
[11] K. Puranik et al., "Attentive fine-tuning of Transformers for Translation of low-resourced languages @LoResMT," arXiv: Computation and Language, Aug. 2021, [Online]. Available: https://arxiv.org/abs/2108.08556
[12] S. A. Jadhav, "Marathi To English Neural Machine Translation With Near Perfect Corpus And Transformers," arXiv: Computation and Language, Feb. 2020, [Online]. Available: http://arxiv.org/pdf/2002.11643.pdf
[13] A. Jain, S. Mhaskar, and P. Bhattacharyya, "Evaluating the Performance of Back-translation for Low Resource English-Marathi Language Pair: CFILT-IITBombay @ LoResMT 2021," pp. 158–162, Aug. 2021.
[14] L. Saini and D. Vidhyarthi, "Bidirectional English-Marathi Translation using Pretrained Models: A Comparative Study of Different Pre-Trained Models," pp. 1–8, Nov. 2023, doi: 10.1109/incoft60753.2023.10425770.
[15] Y. Wang, L. Wang, S. Shi, V. O. K. Li, and Z. Tu, "Go From the General to the Particular: Multi-Domain Translation with Domain Transformation Networks," vol. 34, no. 05, pp. 9233–9241, Apr. 2020, doi: 10.1609/AAAI.V34I05.6461.
[16] R.Samuels,"Evaluating Terminology Translation in MT," Lecture Notes in Computer Science, pp. 495–520, Jan. 2023, doi: 10.1007/978-3-031-24337-0_35.
[17] H. S. Anand, A. R. Dash, and Y. Sharma, "Empowering Low-Resource Language Translation: Methodologies for Bhojpuri-Hindi and Marathi-Hindi ASR and MT," pp. 229–234, Jan. 2024, doi: 10.18653/v1/2024.iwslt-1.28.
[18] Y. Wang, L. Wang, S. Shi, V. O. K. Li, and Z. Tu, "Go From the General to the Particular: Multi-Domain Translation with Domain Transformation Networks," arXiv: Computation and Language, Nov. 2019, [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/download/6461/6317
[19] A. Kr. Ojha, P. Rani, A. Bansal, B. R. Chakravarthi, R. Kumar, and J. P. McCrae, "NUIG-Panlingua-KMI Hindi-Marathi MT Systems for Similar Language Translation Task @ WMT 2020," Empirical Methods in Natural Language Processing, pp. 418–423, Nov. 2020, [Online]. Available: https://aclanthology.org/2020.wmt-1.49/
[20] S. B. Das, D. Panda, T. Mishra, and B. K. Patra, "Statistical machine translation for Indic languages," pp. 1–18, Jun. 2024, doi: 10.1017/nlp.2024.26.