MahaParaphrase: A Marathi Paraphrase Detection Corpus and BERT-based Models

Suramya Jadhav, Abhay Shanbhag, Amogh Thakurdesai,
Ridhima Sinare, Ananya Joshi, Raviraj Joshi

Proceedings of the 39th Pacific Asia Conference on
Language, Information and Computation (PACLIC 39)

Emmanuele Chersoni, Jong-Bok Kim (eds.)

2025

MahaParaphrase: A Marathi Paraphrase Detection Corpus and BERT-based Models

Suramya Jadhav[1,4], **Abhay Shanbhag**[1,4], **Amogh Thakurdesai**[1,4],
Ridhima Sinare[1,4], **Ananya Joshi**[2,4], and **Raviraj Joshi**[*3,4]
[1]Pune Institute of Computer Technology, Pune
[2]MKSSS' Cummins College of Engineering for Women, Pune
[3]Indian Institute of Technology Madras, Chennai
[4]L3Cube Labs, Pune

Abstract

Paraphrases are a vital tool to assist language understanding tasks such as question answering, style transfer, semantic parsing, and data augmentation tasks. Indic languages are complex in natural language processing (NLP) due to their rich morphological and syntactic variations, diverse scripts, and limited availability of annotated data. In this work, we present the L3Cube-MahaParaphrase Dataset, a high-quality paraphrase corpus for Marathi, a low resource Indic language, consisting of 8,000 sentence pairs, each annotated by human experts as either Paraphrase (P) or Non-paraphrase (NP). We also present the results of standard transformer-based BERT models on these datasets. The dataset and model are publicly shared at https://github.com/l3cube-pune/MarathiNLP.

1 Introduction

Paraphrasing is the task of generating semantically equivalent sentences with different wording or structure. (Bhagat and Hovy, 2013) defines paraphrases as "different surface realizations of the same semantic content, while Barzilay and McKeown (2001) describes paraphrases as "textual expressions that share the same meaning but differ in form". It plays a crucial role in various natural language processing (NLP) applications and is inherently familiar to speakers of all languages (Madnani and Dorr, 2010). Paraphrasing can be a vital tool to assist language understanding tasks such as question answering, style transfer (Krishna et al., 2020), semantic parsing (Cao et al., 2020), and data augmentation tasks (Gao et al., 2020). Interestingly, paraphrase identification can also be effectively implemented for plagiarism detection (Hunt et al., 2019).

The MRPC [1], an English paraphrase corpus, is one such dataset that set a benchmark in creating paraphrase datasets. Since then, a wide variety of techniques, as mentioned in Zhou and Bhat (2021), Madnani and Dorr (2010), Gadag and Sagar (2016) have been developed. However, many of such developments have been around the English language, which for a long time now has been a high-resource language. With plenty of corpora spanning multiple domains like news, sentiment analysis, etc., the preliminary source of sentences becomes rich in diversity, making paraphrase data generation easier. Moreover, models used for detecting semantic and lexical relations between sentence pairs are extensively being developed and put into use, as in Khairova et al. (2022). Developments have also been around languages like Vietnamese (Phan et al., 2022) and Finnish (The Turku Paraphrase (Kanerva et al., 2024)). The ParaCotta corpus (Aji et al., 2022) consists of a paraphrase dataset for around 17 languages and also illustrates how Sentence Transformers like S-BERT can be effectively used for generation as well as evaluation.

While the most important thing to build any model or task-specific dataset (a paraphrase dataset in this case) is having a diverse corpus of scraped and manually verified data, this is severely lacking in the case of Indic languages. This is because of the complexity of Indic languages due to their rich morphological and syntactic variations, diverse scripts, and limited availability of annotated data. However, there has been significant progress in Indic NLP research due to the AI4Bharat-IndicNLP project and IndicNLPSuite (Kakwani et al., 2020), who provide corpora and resources like pretrained models for 10 Indian languages across tasks like sentiment analysis and news headline classification. The Amritha corpus is a paraphrase dataset focused on 4 languages: Hindi, Malayalam, Punjabi, and Tamil (Anand Kumar et al., 2016). The BanglaParaphrase (Akil et al., 2022) focuses on using IndicBART for curating the Bangla paraphrase

[*] Correspondence: ravirajoshi@gmail.com

[1] Microsoft Research Paraphrase Corpus

corpus.

Very few research groups such as L3Cube [2] are focusing on regional, low resource Indic languages like Marathi. They have also demonstrated that using LLMs for dataset curation (for annotations) has not shown promising results (Jadhav et al., 2024). Moreover, handling paraphrases in Marathi is tricky due to its lexical syntax, complex linguistic features, and the influence of various dialects (Lahoti et al., 2022), (Dani and Sathe, 2024). L3Cube's MahaNLP project (Joshi, 2022b), focused specifically on the Marathi language by developing a Marathi corpus across multiple domains, which helps Marathi NLP.

Contributing to the same project, in this work, we present the L3Cube-MahaParaphrase[3] Dataset, a robust Marathi paraphrase corpus with each sentence pair annotated and manually curated as Paraphrase (P) or Non-paraphrase (NP) with a total of 8K sentence pairs. We further divide the dataset into 5 buckets based on the increasing degree of paraphrase with word overlap and semantic accuracy as factors, giving future research a chance to explore based on varying degrees of paraphrase. The 2-label annotation approach employed is thoroughly described. Furthermore, we also present the results of standard transformer-based BERT models on these datasets. Our key contributions are as follows:

- Created a gold standard 8K Paraphrase corpus for Marathi with labelled sentences pairs as P or NP (4K each for P and NP).

- We divide the MahaParaphrase corpus into multiple buckets based on lexical (word-level) overlap and semantic similarity, thereby capturing varying degrees of paraphrastic and non-paraphrastic relationships between sentence pairs.

- We evaluate existing models like Muril, mBERT, IndicBERT as well as L3Cube's MahaBERT for benchmarking. Additionally, we release MahaParaphrase-BERT[4], a fine-tuned version of MahaBERT trained on the MahaParaphrase corpus.

2 Literature Review

Research on paraphrase detection and generation has been extensively explored in high-resource languages like English. Different techniques have emerged for paraphrase generation and detection, ranging from using Bi-LSTM with pretrained GLoVe word vectors (Shahmohammadi et al., 2021) to fine-tuning T5 models (Kubal and Palivela, 2021; Palivela, 2021), and using advanced transformer models like GPT and BERT (Natsir et al., 2023). Combined techniques like variational sampling with hashing sampling, an unsupervised method, have been used for phrase-level and sentence-level paraphrase detection (Hejazizo, 2021). Gangadharan et al. (2020) demonstrated how word vectorization can convert textual data into numerical representations for paraphrase detection and analysis, exploring Count Vectorizer, Hashing Vectorizer, TF-IDF Vectorizer, FastText, ELMo, GloVe, and BERT.

It is important to note that much of this research primarily used English paraphrase corpora for experimentation. Experimentation for other languages is limited due to the lack of quality datasets.

As far as low-resource paraphrase datasets are concerned, Kanerva et al. (2024) introduced a comprehensive dataset, 'Turku Paraphrase,' for the Finnish language. The OpenParcus Dataset consists of paraphrases for six European languages. The ParaCotta Corpus (Aji et al., 2022), which includes around 17 languages, including Hindi, is one of the most diverse datasets spanning a wide variety of languages.

Talking about Indic languages, generation of paraphrases becomes difficult because of rich morphological and syntactic variations and diverse scripts. Moreover, all Indic languages fall under the low-resource category due to the lack of annotated data. The Bangla Paraphrase (Akil et al., 2022) uses IndicBART to synthetically generate paraphrases. In Anand Kumar et al. (2016), a significant milestone was achieved with the release of the Amritha paraphrase corpus for four Indic languages: Hindi, Malayalam, Punjabi, and Tamil, as part of the DPIL@FIRE2016 Shared Task, enabling participants to experiment further.

Another notable effort is the IndicParaphrase Dataset by AI4Bharat(Kumar et al., 2022), which includes 11 Indic languages: Assamese (as), Bengali (bn), Gujarati (gu), Kannada (kn), Hindi (hi), Malayalam (ml), Marathi (mr), Oriya (or), Pun-

[2] https://github.com/l3cube-pune/MarathiNLP

[3] https://huggingface.co/datasets/l3cube-pune/MahaParaphrase

[4] https://huggingface.co/l3cube-pune/marathi-paraphrase-detection-bert

jabi (pa), Tamil (ta), and Telugu (te). This dataset provides input and target sentences, as well as a reference list of five sentences with different levels of lexical correlation.

While the Marathi subset in IndicParaphrase is huge for Marathi, it is important to note that it consists only of paraphrased sentence pairs. The same applies to many of the above-mentioned corpora, like BanglaParaphrase, Turku Paraphrase, and OpusParcus. However, the Amritha corpus for the DPIL@FIRE2016 shared task includes labeled sentence pairs as P (Paraphrase) and NP (Non-paraphrase) but does not include Marathi. To date, there is no Marathi paraphrase dataset that consists of both P and NP sentence pairs with 5 varying paraphrastic levels.

3 Dataset

This section provides information on how the dataset was collected. We created the paraphrase dataset in three phases: gathering sentences from MahaCorpus, categorizing them into P and NP using both cosine similarity and back-translation approaches, and then these sentences were manually verified for errors by four native Marathi human annotators. Finally, the sentences were divided into five equally distributed buckets based on word overlaps. Each of these steps is discussed in the following subsections and represented in Figure 1.

3.1 Collection

The required Marathi sentences were taken from the MahaCorpus dataset by L3 Cube, which spans a wide range of topics, including news, sentiment, and hate speech. These sentences were collected from various news sources from the Maharashtra region.

We randomly selected 1 million sentences from this corpus as our primary dataset. In this section, we elaborate on the annotation process for labeling sentences as Paraphrase (P) and Non-Paraphrase(NP).

3.2 Annotations

The collected sentence pairs were annotated using 2 approaches so as to get a mixture of both real and synthetic data. We now explain the two approaches used to categorize sentences as P or NP.

3.2.1 Approach 1: Cosine Similarity

In this approach, we calculated the cosine similarity for every pair of sentences from the 1 million collected sentences. Since contextualized token embeddings have been shown to be effective for paraphrase detection (Devlin et al., 2019), we use BERTScore (Zhang et al., 2020) to ensure semantic similarity between the source and candidates. To do this, we used the sentence transformer MahaSBERT (Joshi et al., 2023) to generate sentence embeddings. Then, we calculated the cosine similarity scores between the embeddings of each pair of sentences.

Based on the scores, we categorized the sentences as follows:

- If the cosine similarity (C.S) score was less than 0.8, the sentences were labeled as NP.
- If the cosine similarity score was between 0.8 and 0.99, the sentences were labeled as P.

3.2.2 Approach 2: Back Translation

In this approach, we used the back-translation technique to generate paraphrase sentences. The process involves:

- Translating a Marathi sentence (S1) into English (S2) using Google Translator.
- Translating it back from English (S2) to Marathi (S3) using Google Translator.

This gives us a pair of sentences: the original sentence (S1) and the back-translated sentence (S3), which we consider as paraphrases.

Filter: To ensure that S3 is not identical to S1, we applied a filter after translation. We used a sentence transformer to calculate the cosine similarity between the sentences and enforced the following rules:

- If the cosine similarity (C.S) score was less than 0.8, we discarded the pair (indicating the meaning might have changed).
- If the cosine similarity score was greater than 0.99, we discarded the pair (indicating the sentences were too similar, likely identical, and not valid paraphrases).

For Marathi-to-English translations (i.e. S1 to S2), we used IndicSBERT (Deode et al., 2023), and for Marathi-to-Marathi comparisons, we used MahaSBERT (Joshi et al., 2023) to compute the cosine similarity score between the two sentences (i.e S1 and S3) in the filter.

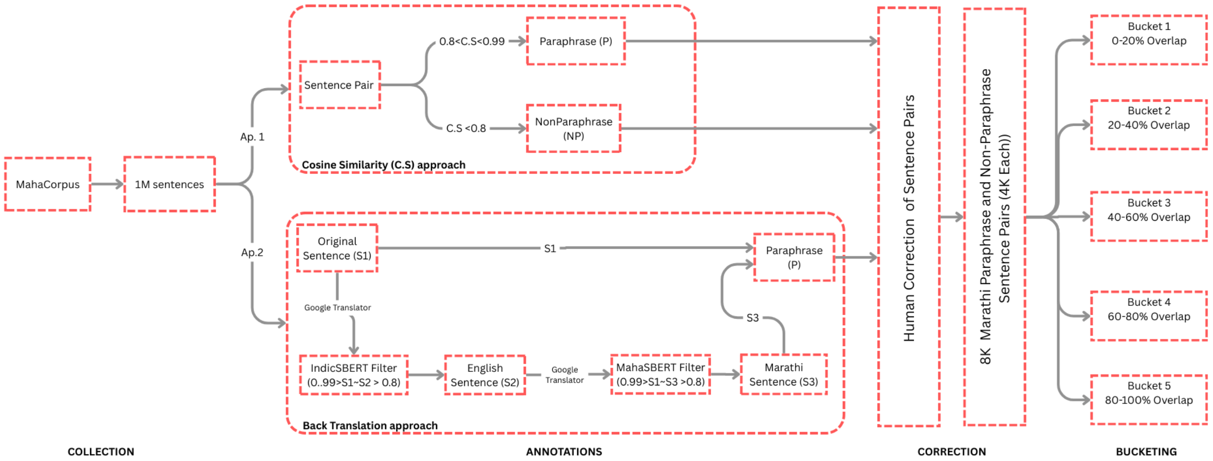

Figure 1: MahaParaphrase Dataset Curation Workflow.

Combining Approaches: Approach 1 provides real data, as both sentences are directly taken from the MahaCorpus (Joshi, 2022a). On the other hand, Approach 2 generates new sentences, which are synthetic. To maintain balance, we used an equal number of sentences from both approaches.

3.3 Human Correction

To ensure that all sentences were correctly classified, the entire dataset was manually verified by native Marathi speakers proficient in reading and writing Marathi.

Any errors were corrected by manually modifying the sentences to ensure accuracy and consistency.

3.4 Bucketing

We further categorized the P and NP data into five buckets for each category, based on word overlap.

- Bucket B5: 80-100% word overlap
- Bucket B4: 60-80% word overlap
- Bucket B3: 40-60% word overlap
- Bucket B2: 20-40% word overlap
- Bucket B1: 0-20% word overlap

Word overlap is calculated as:

$$\text{Word Overlap}(A, B) = \frac{|W(A) \cap W(B)|}{|W(A) \cup W(B)|} \quad (1)$$

where $W(A)$ and $W(B)$ denote the sets of words in sentences A and B respectively.

For example, a pair of sentences in B5 of the NP dataset will be semantically different (and hence categorized as NP), even though they have around 80% or more word overlap. This is significant because it highlights that even with high word overlap, sentences can have different meanings, emphasizing the importance of word order when considering paraphrase pairs.

Now consider another example: a pair of sentences in B1 of the P dataset. These sentences have low word overlap but are still considered paraphrases. This shows that sentences with different words (such as synonyms) can also form valid paraphrase pairs.

This categorization into buckets makes the dataset robust and versatile for evaluation across different scenarios, such as high-overlap non-paraphrase sentence pairs or low-overlap paraphrase sentence pairs. Refer table 1 for bucket-wise examples from the dataset.

4 Dataset Statistics

The dataset contains 4000 rows of sentence pairs labeled as paraphrase (P), and 4000 rows of sentence pairs labeled as non-paraphrase (NP). Each row in the dataset contains a sentence along with a paraphrase or non-paraphrase label.

The average word count for sentences in the dataset, as well as the difference between the averages, is given in Table 2.

Figure 2 shows the distribution of sentence lengths for both Paraphrase (P) and Non-paraphrase (NP) classes in our dataset. Both distributions follow a similar pattern, with the highest frequency occurring for sentence lengths between

Bucket	P	NP
0–20	तर बहिणीला वस्त्राभुषणाची भेट देऊन तिला सुख, शांतता, सौभाग्य, समृद्धी प्राप्त व्हावी, यासाठी भाऊराया प्रार्थना करतो (So that the sister receives happiness, peace, good fortune, and prosperity, the brother prays while giving her clothes and ornaments.) रक्षाबंधन आटोपल्यावर भावाने वस्त्र, आभुषणे किंवा इच्छित भेटवस्तू बहिणीला देऊन तिच्या सुखी जीवनासाठी प्रार्थना करावी (After Raksha Bandhan, the brother should give clothes, ornaments, or a desired gift to the sister and pray for her happy life.) **(19.35%)**	विरोधी भाजपताराराणी आघाडीकडून महापौरपदासाठी अर्ज दाखल न झाल्याने महापौरपद निवडीची केवळ औपचारिकता राहिली आहे (As no nomination was filed for mayoral post from the opposition BJP alliance, the selection has become a mere formality.) काम पूर्ण न झाल्याने मंत्रालय पातळीवर नाराजी असल्याने फिआफच्या परिषदेसाठी कुणाला पाठवले गेले नसल्याचे बोलले जात आहे (Due to incomplete work and dissatisfaction at the ministry level, no one was sent for the FIAF conference, as per reports.) **(19.35%)**
20–40	या प्रकरणाची केस डायरी, घटना स्थळाचे फोटो, ऑटोप्सी रिपोर्ट, मुंबई पोलिसांचा फॉरेन्सिक रिपोर्ट आणि नोंदवलेल्या साक्षीदारांचे जबाब लवकरात लवकर पाठवण्याची विनंती सीबीआयकडून करण्यात येणार आहे (CBI will request to send the case diary, crime scene photos, autopsy report, forensic report from Mumbai Police, and recorded witness statements as soon as possible.) याशिवाय सीबीआय मुंबई पोलिसांकडून केसची डायरी, ऑटोप्सी रिपोर्ट, क्राइम सीनचे फोटो, मुंबई पोलिसांचे फॉरेन्सिक रिपोर्ट, पोस्टमॉर्टेम रिपोर्ट, साक्षीदारांचे नोंदवलेले जबाब यांच्या प्रती घेणार आहे (Additionally, CBI will collect copies of the case diary, autopsy report, crime scene photos, forensic report from Mumbai Police, postmortem report, and witness statements.) **(39.13%)**	गेल्या सलग दोन वर्षांत ठाण्यात लायसन्ससाठी अर्ज केलेल्या उमेदवारांपैकी १७ टक्के अर्जदारांना वाहतूक चिन्हांबाबत माहितीच नसल्याचे स्पष्ट झाले आहे (In the past two years, 17% of license applicants in Thane were found to lack knowledge of traffic signs.) या पार्श्वभूमीवर, गेल्या २ वर्षभरातील अर्जदारांच्या चाचणीच्या निकालाची माहिती मिळवली असता, तब्बल १७ टक्के उमेदवारांना वाहतूक चिन्हांची माहितीच नसल्याचे उघड झाले आहे (Data from applicants' tests over the past 2 years revealed that 17% of candidates had no knowledge of traffic signs.) **(39.02%)**
40–60	म्हणून लवकरच अर्थव्यवस्था पूर्वपदावर येतील आणि इंधन मागणी वाढेल, अशी अपेक्षा सौदी अरामकोचे मुख्य कार्यकारी अधिकारी अमीन नासिर यांनी सांगितले (The economy is expected to recover soon and fuel demand will rise, said Amin Nasser, CEO of Saudi Aramco.) म्हणून अर्थव्यवस्था लवकरच पूर्वपदावर येईल आणि इंधनाची मागणी वाढेल, असे सौदी आरामकोचे सीईओ अमीन नासेर यांनी सांगितले (The economy will soon return to normal, and fuel demand will rise, said Saudi Aramco CEO Amin Nasser.) **(59.46%)**	या उपोषणात कार्यकर्त्यांनी सहभागी व्हावे, असे आवाहन भाजपचे शहर जिल्हाध्यक्ष भगवान घडमोडे व ग्रामीण जिल्हाध्यक्ष एकनाथराव जाधव यांनी केले आहे (Workers should participate in this protest, appealed BJP's city district president Bhagwan Ghadmode and rural president Eknathrao Jadhav.) भाजपचे पदाधिकारी, कार्यकर्ते व नागरिकांनी या आंदोलनात सहभागी व्हावे, असे आवाहन भगवान घडमोडे, विजय साळवे यांनी केले आहे (BJP officials, workers, and citizens should join the protest, appealed Bhagwan Ghadmode and Vijay Salve.) **(59.46%)**
60–80	तसेच, जीवे मारण्याची धमकी देऊन त्यांच्याकडील ७०० रुपयांची रोख रक्कम आणि दुचाकी असा ४० हजार ७०० रुपयांचा ऐवज चोरून नेला (Threatening to kill, he stole □700 in cash and a bike totaling □40,700.) तसेच, जीवे मारण्याची धमकी देत त्यांच्याकडील ७०० रुपयांची रोख रक्कम आणि दुचाकी असा ४० हजार ७०० रुपयांचा ऐवजाची चोरी केली (He threatened and stole □700 in cash and a bike worth □40,700.) **(78.95%)**	पिंपरीतील शंभर टक्के निकालस्टेर्लिंग हायस्कूल, प्रियदर्शनी, स्वामी समर्थ (भोसरी), कमलनयन बजाज, गीतामाता, (चिंचवड), एसएनबीपी, अल्फोन्सा, निर्मल बेथनी, रॉजर्स इंग्लिश, डी (100% results in schools like Sterling High School, Priyadarshani, etc. in Pimpri) शंभर टक्के निकाल लागलेल्या शाळा स्टेर्लिंग हायस्कूल, प्रियदर्शनी, स्वामी समर्थ (भोसरी), कमलनयन बजाज, गीता माता, (चिंचवड), एसएनबीपी, अल्फोन्सा, निर्मल बेथनी, रॉजर्स इंग्लिश, डी (Schools with 100% result: Sterling High School, Priyadarshani, etc.) **(79.07%)**
80–100	भाद्रपद महिन्यातील शुद्ध अष्टमीला राधादेवीचा जन्म झाला (Radhadevi was born on Shuddha Ashtami in the month of Bhadrapada.) राधादेवीचा जन्म भाद्रपद महिन्यातील शुद्ध अष्टमीला झाला (Radhadevi's birth occurred on Shuddha Ashtami in the month of Bhadrapada.) **(100.00%)**	मी खरी भविष्यवाणी केली नव्हती काय? (Didn't I make the correct prediction?) मी खरी भविष्यवाणी केली नव्हती काय? (Didn't I make the correct prediction?) **(100.00%)**

Table 1: P and NP Sentence Pairs with Overlap Percentages by Bucket. Every cell consists of S1 and S2 along with their english translations followed by the word overlap percentages (in **bold**). The examples choosen are pairs with max word overlap in that particular bucket.

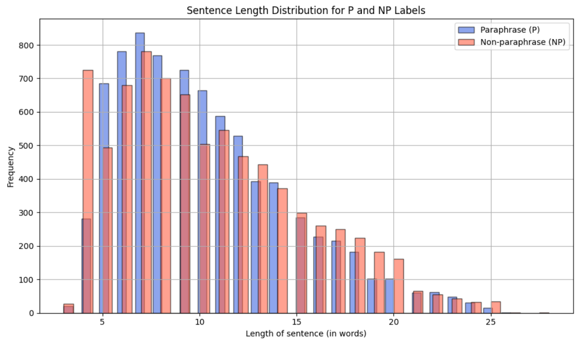

Figure 2: Sentence length distribution for Paraphrase (P) and Non-paraphrase (NP) classes. The x-axis shows sentence length in words, and the y-axis indicates the frequency of those lengths.

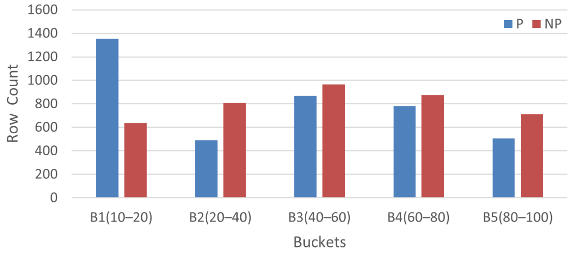

Figure 3: Bucket Wise Distribution. The values in brackets are the word overlap percentages for each bucket.

5 and 15 words.

Figure 3 shows the bucket wise row count distribution for both Paraphrase (P) and Non-paraphrase (NP) classes. While the total count of P and NP are same (i.e 4000 each), their distribution across buckets is varied.

Dataset	Sentence 1 Avg.	Sentence 2 Avg.	Avg. Diff.
Paraphrase	10.43	9.99	1.45
Non-Paraphrase	9.52	11.22	3.36

Table 2: Average word counts and average difference in sentence lengths per record for Paraphrase and Non-Paraphrase datasets.

5 Baseline Models

5.1 Muril

MuRIL is a language model built especially for Indian languages and trained entirely on a large volume of Indian language text (Khanuja et al., 2021). The dataset contains both translated and transliterated document pairings in order to introduce supervised cross-lingual learning during training.

5.2 MBERT

A BERT-based model called Multilingual BERT (mBERT) was trained using text in 104 distinct languages (Devlin et al., 2019). It is trained with masked language modeling (MLM) and next sentence prediction (NSP) objectives, and it supports a variety of downstream applications, including sentiment analysis.

5.3 IndicBERT

Based on the ALBERT architecture (Lan et al., 2020), IndicBERT is a language model that was trained on a huge corpus of 12 major Indian languages, including Bengali, English, Gujarati, Hindi, Kannada, Malayalam, Marathi, Oriya, Punjabi, Tamil, Telugu, and Assamese. It employs a combined training technique and makes use of data from the IndicCorp dataset (Kunchukuttan et al., 2020) in order to better accommodate low-resource languages. There are two versions of the model: IndicBERT (MLM+TLM) and IndicBERT (MLM alone). IndicBERT-MLM is trained by masking random tokens in monolingual text and predicting them using context. IndicBERT-TLM uses parallel sentences in different languages, masking tokens and predicting them using both languages.

5.4 MahaBERT

A multilingual BERT model called MahaBERT was refined using the L3Cube-MahaCorpus and additional publically accessible Marathi monolingual datasets (Joshi, 2022a).

6 Result

The baseline models described above were fine-tuned and evaluated on our dataset, and the results are presented in Table 3. Among the models that were evaluated, MahaBERT was the most accurate model, followed by IndicBERT (MLM + TLM), Muril, IndicBERT (MLM only) and MBERT.

Model	Score
MahaBERT	88.7
IndicBERT (MLM+TLM)	87.1
Muril	86.9
IndicBERT (MLM only)	85.9
MBERT	84.59

Table 3: Model Performance Comparison. MLM stands for Masked Language Modeling and TLM stands for Translation Language Modeling.

7 Conclusion

In this paper, we present the MahaParaphrase dataset, comprising of 8,000 labeled pairs of both paraphrase and non-paraphrase sentences. The entire dataset was manually verified by four native Marathi speakers and is further divided into five buckets based on word overlap. These bucketed subsets capture varying degrees of paraphrasing intensity, which can support more nuanced research in this domain.

Furthermore, we evaluate the MahaParaphrase dataset using five baseline models, with MahaBERT achieving the highest performance—an F1 score of 88.7%.

By providing this low-resource paraphrase dataset, we aim to equip researchers and practitioners with a valuable resource to advance further research in Marathi NLP.

8 Limitations

Compared to paraphrase dataset for high-resource languages, this dataset is relatively small (8K pairs). Moreover, the presence of code-mixed sentences introduces minor noise especially when using BERT models trained specifically using Marathi. Additionally, the dataset evaluation was limited to BERT-based models; incorporating LLMs could offer a more comprehensive assessment.

Acknowledgement

This work was carried out under the mentorship of L3Cube, Pune. We would like to express our gratitude towards our mentor, for his continuous support and encouragement. This work is a part of the L3Cube-MahaNLP project (Joshi, 2022b).

References

Alham Fikri Aji, Tirana Noor Fatyanosa, Radityo Eko Prasojo, Philip Arthur, Suci Fitriany, Salma Qonitah, Nadhifa Zulfa, Tomi Santoso, and Mahendra Data. 2022. Paracotta: Synthetic multilingual paraphrase corpora from the most diverse translation sample pair. *Preprint*, arXiv:2205.04651.

Ajwad Akil, Najrin Sultana, Abhik Bhattacharjee, and Rifat Shahriyar. 2022. Banglaparaphrase: a high-quality bangla paraphrase dataset. *arXiv preprint arXiv:2210.05109*.

M Anand Kumar, Shivkaran Singh, B Kavirajan, and KP Soman. 2016. Shared task on detecting paraphrases in indian languages (dpil): An overview. In *Forum for Information Retrieval Evaluation*, pages 128–140. Springer.

Regina Barzilay and Kathleen McKeown. 2001. Extracting paraphrases from a parallel corpus. In *Proceedings of the 39th annual meeting of the Association for Computational Linguistics*, pages 50–57.

Rahul Bhagat and Eduard Hovy. 2013. What is a paraphrase? *Computational linguistics*, 39(3):463–472.

Ruisheng Cao, Su Zhu, Chenyu Yang, Chen Liu, Rao Ma, Yanbin Zhao, Lu Chen, and Kai Yu. 2020. Unsupervised dual paraphrasing for two-stage semantic parsing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6806–6817, Online. Association for Computational Linguistics.

Asang Dani and Shailesh R Sathe. 2024. A review of the marathi natural language processing. *arXiv preprint arXiv:2412.15471*.

Samruddhi Deode, Janhavi Gadre, Aditi Kajale, Ananya Joshi, and Raviraj Joshi. 2023. L3cube-indicsbert: A simple approach for learning cross-lingual sentence representations using multilingual bert. In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, pages 154–163.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Ashwini Gadag and BM Sagar. 2016. A review on different methods of paraphrasing. In *2016 International conference on electrical, electronics, communication, computer and optimization techniques (ICEECCOT)*, pages 188–191. IEEE.

Veena Gangadharan, Deepa Gupta, L Amritha, and TA Athira. 2020. Paraphrase detection using deep neural network based word embedding techniques. In *2020 4th International Conference on Trends in Electronics and Informatics (ICOEI)(48184)*, pages 517–521. IEEE.

Silin Gao, Yichi Zhang, Zhijian Ou, and Zhou Yu. 2020. Paraphrase augmented task-oriented dialog generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 639–649, Online. Association for Computational Linguistics.

Ali Hejazizo. 2021. Combining variational sampling and metropolis–hastings sampling for paraphrase generation.

Ethan Hunt, Ritvik Janamsetty, Chanana Kinares, Chanel Koh, Alexis Sanchez, Felix Zhan, Murat Ozdemir, Shabnam Waseem, Osman Yolcu, Binay

Dahal, and 1 others. 2019. Machine learning models for paraphrase identification and its applications on plagiarism detection. In *2019 IEEE International Conference on Big Knowledge (ICBK)*, pages 97–104. IEEE.

Suramya Jadhav, Abhay Shanbhag, Amogh Thakurdesai, Ridhima Sinare, and Raviraj Joshi. 2024. On limitations of llm as annotator for low resource languages. *arXiv preprint arXiv:2411.17637*.

Ananya Joshi, Aditi Kajale, Janhavi Gadre, Samruddhi Deode, and Raviraj Joshi. 2023. L3cube-mahasbert and hindsbert: Sentence bert models and benchmarking bert sentence representations for hindi and marathi. In *Science and Information Conference*, pages 1184–1199. Springer.

Raviraj Joshi. 2022a. L3cube-mahacorpus and mahabert: Marathi monolingual corpus, marathi bert language models, and resources. *arXiv preprint arXiv:2202.01159*.

Raviraj Joshi. 2022b. L3cube-mahanlp: Marathi natural language processing datasets, models, and library. *arXiv preprint arXiv:2205.14728*.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul NC, Avik Bhattacharyya, Mitesh M Khapra, and Pratyush Kumar. 2020. Indicnlpsuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for indian languages. In *Findings of the association for computational linguistics: EMNLP 2020*, pages 4948–4961.

Jenna Kanerva, Filip Ginter, Li-Hsin Chang, Iiro Rastas, Valtteri Skantsi, Jemina Kilpeläinen, Hanna-Mari Kupari, Aurora Piirto, Jenna Saarni, Maija Sevón, and 1 others. 2024. Towards diverse and contextually anchored paraphrase modeling: A dataset and baselines for finnish. *Natural Language Engineering*, 30(2):319–353.

Nina Khairova, Anastasiia Shapovalova, Orken Mamyrbayev, Nataliia Sharonova, and Kuralay Mukhsina. 2022. Using bert model to identify sentences paraphrase in the news corpus. In *COLINS*, pages 38–48.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, and 1 others. 2021. Muril: Multilingual representations for indian languages. *arXiv preprint arXiv:2103.10730*.

Kalpesh Krishna, John Wieting, and Mohit Iyyer. 2020. Reformulating unsupervised style transfer as paraphrase generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 737–762, Online. Association for Computational Linguistics.

D Kubal and H Palivela. 2021. Unified model for paraphrase generation and paraphrase identification.

Aman Kumar, Himani Shrotriya, Prachi Sahu, Raj Dabre, Ratish Puduppully, Anoop Kunchukuttan, Amogh Mishra, Mitesh M. Khapra, and Pratyush Kumar. 2022. Indicnlg suite: Multilingual datasets for diverse nlg tasks in indic languages.

Anoop Kunchukuttan, Divyanshu Kakwani, Satish Golla, Gokul N. C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. Ai4bharat-indicnlp corpus: Monolingual corpora and word embeddings for indic languages. *Preprint*, arXiv:2005.00085.

Pawan Lahoti, Namita Mittal, and Girdhari Singh. 2022. A survey on nlp resources, tools, and techniques for marathi language processing. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(2):1–34.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. *Preprint*, arXiv:1909.11942.

Nitin Madnani and Bonnie J Dorr. 2010. Generating phrasal and sentential paraphrases: A survey of data-driven methods. *Computational Linguistics*, 36(3):341–387.

Affan Hilmy Natsir, Indriana Hidayah, and Teguh Bharata Adji. 2023. Deep learning in paraphrase generation: A systematic literature review. In *2023 IEEE 7th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, pages 118–123. IEEE.

Hemant Palivela. 2021. Optimization of paraphrase generation and identification using language models in natural language processing. *International Journal of Information Management Data Insights*, 1(2):100025.

Quoc Long Phan, Tran Huu Phuoc Doan, Ngoc Hieu Le, Ngoc Bao Duy Tran, and Tuong Nguyen Huynh. 2022. Vietnamese sentence paraphrase identification using sentence-bert and phobert. In *International Conference on Intelligence of Things*, pages 416–423. Springer.

Hassan Shahmohammadi, MirHossein Dezfoulian, and Muharram Mansoorizadeh. 2021. Paraphrase detection using lstm networks and handcrafted features. *Multimedia Tools and Applications*, 80(4):6479–6492.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

Jianing Zhou and Suma Bhat. 2021. Paraphrase generation: A survey of the state of the art. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5075–5086, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.