BnPC: A Corpus for Paraphrase Detection in Bangla

Anonymous ACL submission

Abstract

In this paper, we present the first benchmark dataset for paraphrase detection in Bangla language. Despite being the sixth most spoken language[1] in the world, paraphrase identification in the Bangla language is barely explored. Our dataset contains 8,787 human-annotated sentence pairs collected from a total of 23 newspaper outlets' headlines on four categories. We explore different linguistic features and pre-trained language models to benchmark the dataset. We perform a human evaluation experiment to obtain a better understanding of the task's constraints, which reveals intriguing insights. We make our dataset and code publicly available.[2]

1 Introduction

Paraphrase identification is considered to be one of the pivotal and fundamental tasks of Natural Language Processing (NLP). When two different sentences express the same meaning, they are called paraphrases. Paraphrase identification has many implications on tasks like question answering (Fader et al., 2013), text summarization (Barzilay et al., 1999), plagiarism detection (Barrón-Cedeño et al., 2013), information retrieval (Wallis, 1993), first story detection (Petrović et al., 2012), and so on. As a result, extensive research has been conducted on paraphrase identification, and numerous paraphrase corpora have been developed in various languages like English (Dolan and Brockett, 2005), Turkish (Demir et al., 2012), Russian (Pronoza et al., 2016), Arabic (Menai, 2019), Portuguese (Zhang et al., 2019), Chinese (Fonseca et al., 2016), and so on.

A descendent of Sanskrit, Bangla is currently spoken by over 260 million people in the world and is set to become the third most spoken language in the fastest growing economies by 2050[3]. Bangla is the language of the people of the Bengal region, now divided between Bangladesh and the Indian state of West Bengal[4]. As a result of the technological advancements in Bangla speaking communities, the demand and usage of the Bangla language in the digital world continue to grow exponentially. Despite such a growing demand and need for digital Bangla resources, no public Bangla paraphrase corpus is available. In this paper,

- We propose Bangla Paraphrase Corpus (BnPC) consisting of 8,787 annotated pairs.
- We develop a benchmark paraphrase detection system by investigating bag-of-words approach and pre-trained language models.
- We also conduct a human evaluation experiment to get insights on the task.

2 Overview of BnPC

Data Collection As the headlines for an identical event tend to be paraphrases, we created BnPC by collecting news headlines from 23 most-popular[5,6] Bangla news portals. We gathered news on four broad categories: national, international, sports, and entertainment over four months from September to December of 2020. To collect news of identical events, we utilized Google News[7] and Pipilika News[8] (a Bangla search engine) generated news clusters alongside visiting individual news websites. Through manual inspection, we grouped

[1] https://en.wikipedia.org/wiki/List_of_languages_by_total_number_of_speakers

[2] The link is not revealed due to anonymity policy.

[3] https://www.washingtonpost.com/news/worldviews/wp/2015/09/24/the-future-of-language/

[4] https://www.britannica.com/place/Bengal-region-Asia

[5] https://www.alexa.com/topsites/countries/BD

[6] Source portal list in Appendix Table 5.

[7] https://news.google.com/?hl=bn&gl=BD&ceid=BD:bn

[8] https://news.pipilika.com/

Paraphrases with slight lexical differences

- কাল মিয়ানমারে জাতীয় নির্বাচন, রোহিঙ্গারা বঞ্চিত
  National elections in Myanmar tomorrow, Rohingyas deprived
- মিয়ানমারে কাল নির্বাচন : ভোট নেই রোহিঙ্গাদের
  Tomorrow's election in Myanmar: Rohingyas do not have votes

Paraphrases with significant lexical differences

- বিজিবি এখন জলে, স্থলে ও আকাশপথে বিচরণ করবে
  The BGB will now operate on water, land and air
- বিজিবির এয়ার উইংয়ের যাত্রা শুরু, ত্রিমাত্রিক বাহিনী ঘোষণা
  The BGB air wing begins its journey, announcing three-dimensional forces

Non-paraphrases with significant lexical similarity

- পদ্মা সেতুর ৩২তম স্প্যান বসতে পারে আজ
  The 32nd span of the Padma Bridge can sit today
- পদ্মা সেতুর ৩২তম স্প্যান বসতে পারে কাল
  The 32nd span of the Padma Bridge may sit tomorrow

Non-paraphrases with slight lexical similarity

- ফিটনেস টেস্টে সাকিবের বাজিমাত
  Shakib's shines in fitness test
- এক বছরেও 'ফিট' হতে পারেননি নাসির
  Nasir could not be 'fit' in a year

Table 1: Examples of paraphrase and non-paraphrase pairs with different amount of lexical overlap.

145 national, 158 international, 139 sports, and 175 entertainment related news events published by multiple news portals. Each group contained numerous headlines focusing different aspects of the same event. We removed headlines with issues like incomplete sentences, grammatical errors, code-mixing, duplicate sentence pairs, etc.[9] We generated 10,144 sentence pairs by taking sentences from the same groups.

Annotation We followed the guidelines described in Bhagat and Hovy (2013) to annotate candidate pairs. Three annotators were asked to quantify the possibility of being paraphrases with five levels using this scale; 0: Not paraphrase, 0.25: Not paraphrase having slight similarity, 0.5: Not sure or requires more context, 0.75: Paraphrase despite having some differences, 1: Paraphrase. We averaged the score of three annotators and discarded the ones with an average score of 0.5 as the annotators could not agree on whether the pairs are paraphrase or not. These samples were mostly partial-paraphrases or have ambiguous meanings.[10] A Fleiss' Kappa score (Fleiss, 1971) of 0.61 indicates substantial inter-annotator agreement. We present some sample sentence pairs in Table 1.

[9]Examples of discarded sentences are added in Appendix A.2

[10]Examples provided in the Appendix

	T	P	W/S	C/S
Paraphrase	3,426	38.99%	6.97	46.95
Non-Paraphrase	5,361	61.01%	7.32	48.86
Total	8,787	100.00%	7.18	48.11

Table 2: Distribution of T (total number), P (percentage), W/S (word per sentence), and C/S (character per sentence) between paraphrase and non-paraphrase sentence pairs in the dataset.

Statistics As per Table 2, the class distribution of the dataset is slightly skewed towards the non-paraphrases. Also, these non-paraphrase sentences tend to be a little longer than the paraphrase ones. There are 8,541 unique Bangla words (23.8%) in the dataset. We observe lexical diversity in the dataset as 35.19% sentence pairs have zero and 28.94% pairs have only one word in common.

3 Methodology

To develop a paraphrase classifier, we explore the metrics for machine translation evaluation, bag-of-words, and pre-trained language models.

3.1 Evaluation Metric Based Approach

Following Madnani et al. (2012) and Kravchenko (2017), we investigate paraphrase classifiers using machine translation (MT) evaluation metrics like BLEU (Papineni et al., 2002) and METEOR (Lavie and Denkowski, 2009) as these metrics provide a notion of similarity between a reference and a generated text. Given a candidate pair $X = (x_1, x_2)$ and a metric (e.g., BLEU), we classify the pair as a paraphrase or not paraphrase by the following equations:

$$f_{BLEU}(X) = \frac{BLEU(x_1, x_2) + BLEU(x_2, x_1)}{2}$$

$$\hat{y} = \begin{cases} \text{Paraphrase, } \textbf{if } f_{BLEU}(X) \geq \alpha \\ \text{Not Paraphrase, } \textbf{if } f_{BLEU}(X) < \alpha \end{cases}$$

Here, α is a threshold, whose value was set by maximizing the performance on the training set (α=0.249 for BLEU and α=0.136 for METEOR).

3.2 Bag of Words (BOW)

For each text in a candidate pair, we extract word n-grams (n=1, 2, 3) and character n-grams (n=2, 3, 4, 5) and use the cosine similarity scores for each n-gram set as features to train a Support Vector Machine (SVM) classifier. Additionally, we investigate training the model by dividing the mean word embedding vectors of the pair, by its norm and taking the quotient as input feature. We use the pre-trained FastText (Bojanowski et al., 2016) Bangla embedding (coverage=91.77%) for this purpose.

3.3 Pre-trained Language Model

Pre-trained language models, particularly variants of BERT, have shown superior performance in a variety of natural language tasks. We use the Multilingual BERT (MBERT) (Devlin et al., 2018) and two different BERT models pre-trained on only Bangla (Sarker, 2020; Bhattacharjee et al., 2021) from HuggingFace transformers (Wolf et al., 2020) and fine tune the binary prediction layer. BanglaBERT (Bhattacharjee et al., 2021) was trained on wikidump and 30 GB data crawled from 60 Bangla websites, whereas bangla-bert-base (Sarker, 2020) was trained on wikidump and 11 GB web crawled data from OSCAR (Ortiz Suárez et al., 2020).

4 Experiments and Results

4.1 Experimental Setup

We use 70% of the data for training and equally divide the rest of the data for development and test. For the metric based approaches, we remove the punctuations and for BOW based methods, we pre-process the data by removing punctuation and normalizing digits as it shows better results in the development set. As a set of simple baselines, we compare our results with a majority and a random baseline. We report our results using precision, recall, and weighted F1 score. We use Scikit-learn (Buitinck et al., 2013) implementations for SVM, cosine similarity, and n-gram extraction. For the pre-trained language models, we fine-tune (λ=10^{-5}, batch size 32) the models for 20 epochs with early stopping with a patience of 5 epochs.

4.2 Results

Table 3 presents the precision, recall, and weighted F1 scores of different models on the test set[11]. The MT metric-based approaches (BLEU, METEOR) perform relatively well compared to the baselines, with METEOR getting up to 77.08 F1 score. METEOR considers both unigram precision and recall, whereas BLEU solely measures precision when matching candidate sentences to reference sentences. As a consequence, METEOR exhibits a higher correlation with human judgments at the sentence level.

[11] Validation results are provided in Table 6

Model	P	R	F1
Baseline (Random)	50.56	50.67	49.62
Baseline (Majority)	34.86	59.04	43.83
BLEU	67.88	67.86	67.87
METEOR	77.28	77.40	77.08
Unigram (U)	76.67	75.97	74.93
Bigram (B)	74.59	73.67	72.21
Trigram (T)	73.88	66.36	59.46
U+B	76.30	75.82	74.90
U+B+T	76.42	75.90	74.95
Char-2-gram (C2)	79.07	78.62	77.97
Char-3-gram (C3)	78.61	78.41	77.87
Char-4-gram (C4)	78.06	77.76	77.12
Char-5-gram (C5)	77.52	76.97	76.12
C2+C3	78.72	78.41	77.80
C2+C3+C4	78.19	77.98	77.40
C2+C3+C4+C5	78.39	78.12	77.52
U+C2	79.22	78.77	78.11
U+C2+C3	78.73	78.34	77.68
U+C2+C3+C4	78.47	78.05	77.36
All n-grams	78.26	77.76	77.01
Word Embedding (Fastext) (E)	77.53	77.04	76.24
U+C2+E	78.83	78.19	77.41
BanglaBERT (Bhattacharjee et al., 2021)	67.32	67.58	67.45
bangla-bert-base (Sarker, 2020)	75.85	76.04	75.75
MBERT	**82.54**	**82.42**	**82.47**

Table 3: Results from different experiments of baseline, MT metrics, linguistic features, and pre-trained LMs are reported in Precision (P), Recall (R) and weighted-F1 score.

Unigram performs the best among the word n-grams with an F1 score of 74.93 and we notice a decline in F1 for the longer word n-grams. This pattern is consistent with the character n-grams as well. Character bigrams achieve 77.97 F1 score and longer ngrams' F1 score decrease gradually. However, character n-grams show better performance than the word n-grams in general. Usage of prefix, suffix, and word concatenation is heavy in Bangla, which we believe is the reason of the strength of character n-grams. The combination of unigram and character bigrams yields the highest F1 score of 78.11 among all the lexical feature combinations. We observe no improvement in this by integrating the embedding features.

We obtain the best result from MBERT (Devlin et al., 2018), surpassing the performance of the other two BERT models trained on only Bangla. This indicates that Bangla benefits from multilingual knowledge transferred from learning the other languages. This is not surprising as more than 10% of the training languages of MBERT are from the Indo-European languages like Bangla. Additionally, modern Bangla vocabulary is highly influenced by foreign words. Analysing the errors made by these models, we find that BanglaBERT

Sentence 1	Sentence 2	Label	*Subject	**Model
প্রধানমন্ত্রীর সংবাদ সম্মেলন শনিবার (The Prime Minister’s press conference is on Saturday)	প্রধানমন্ত্রীর সংবাদ সম্মেলন আজ (The Prime Minister’s press conference is today)	0	0	1
জাপানে শক্তিশালী ভূমিকম্পে আহত শতাধিক (Hundreds injured in strong earthquake in Japan)	জাপানের উপকূলে ৭ দশমিক ৩ মাত্রার ভূমিকম্প (7.3 magnitude earthquake off the coast of Japan)	0	1	0
করোনায় মৃত্যু প্রায় ২৪ লাখ (About 24 lakh died in Corona)	মৃত্যু ২৩ লাখ ৬৭ হাজার, আক্রান্ত ১০ কোটি সাড়ে ৭৭ লাখের বেশি (23 lakh 67 thousand deaths, more than 10 crore 77.5 lakh affected)	1	1	0
জাপানের উত্তরাঞ্চলে ৭.৩ মাত্রার ভূমিকম্প (7.3 magnitude earthquake shakes northern Japan)	জাপানে ৭.১ মাত্রার ভূমিকম্প (7.1 magnitude earthquake shakes Japan)	1	0	1
একসঙ্গে ১০০ ছবি নির্মাণের ঘোষণা! (Announcement to make 100 movies!)	ইতিহাসে প্রথম : ১০ সিনেমার মহরত, একশ'র ঘোষণা (First in history: 10 movie masterpieces, 100 announcements)	0	1	1
আমেরিকার এই কুখ্যাত জেল বন্ধ করতে পারেন বাইডেন (Biden might close this infamous prison in America)	গুয়ানতানামো বে কারাগার বন্ধ করতে চান বাইডেন (Biden wants to close Guantanamo Bay prison)	1	0	0

Table 4: Disagreement among subject, model, and actual label. Here 1 represents paraphrase and 0 represents non-paraphrase sentence pairs. *Subject’s prediction is taken using majority voting.**Prediction on Multilingual BERT.

(Bhattacharjee et al., 2021) typically mislabels pairs (as paraphrases) with high lexical overlap but low or no overlap in nouns. MBERT (Devlin et al., 2018) fails to detect paraphrases with no significant lexical overlap.

To assess the performance of pre-trained BERT on some paraphrase identification corpora in English, we fine-tune the BERT model on MSRP (Dolan and Brockett, 2005), PIT (Xu et al., 2015), PARADE (He et al., 2020) with the exact experimental setup. The F1 scores are 88.49 (MSRP), 68.11 (PARADE), and 48.55 (PIT). 82.47 F1 on BNPC falls between these scores and provides a competitive benchmark result.

4.3 Human Evaluation

We conduct a human evaluation study with 300 randomly selected examples from our test set to assess the human performance in the task. We take the help of five undergraduates from different majors to ensure diversity in subjects. After instructing them about the task, we ask them to classify each pair into either paraphrase or non-paraphrase. Then we compare their assigned labels against the ground truth. The individual F1 scores of the five annotators are 69.48, 72.25, 74.37, 74.58, and 84.13, yielding an average F1 score of 74.96. Our fine-tuned MBERT model earned an F1 score of 81.89 on this sample of data, indicating that the job is more difficult for humans to accomplish. Analysing the errors and interviewing the human subjects, we find that the main reasons for the errors are lack of domain knowledge, presence of number in the sentences, and pairs with long overlaps of spans. Some examples are presented in Table 4.

5 Conclusion and Future Works

In this paper, we propose BnPC, the first Bangla dataset for paraphrase detection. Through our investigations to develop a benchmark classifier, we find that lexcial features like character n-grams show competitive performance in identifying paraphrases. Similar performance can be achieved by simply using the METEOR score of the pairs. Our experiments show that multilingual knowledge is more helpful for this task than using monolingual pre-trained language models. We release the corpus publicly to foster further work in this area.

As this corpus is limited to only news headlines, models built with this data may not perform well in other domains. Therefore, a good direction for the future work can be extending this dataset with data from different domains and topics, for example conversational data. As our experiments show that, in an identical experimental setup, monolingual BERT models perform poorly than the multilingual BERT, further analysis can be done to objectify the specific multilingual knowledge that is outperforming the monolingual knowledge in this task. This phenomenon can be explored across multiple tasks, as (Bhattacharjee et al., 2021) showed that BanglaBERT outperformed MBERT in tasks like sentiment classification, emotion classification, document classification, named entity recognition, and natural language inference.

References

Alberto Barrón-Cedeño, Marta Vila, M. Antònia Martí, and Paolo Rosso. 2013. Plagiarism Meets Paraphrasing: Insights for the Next Generation in Automatic Plagiarism Detection. *Computational Linguistics*, 39(4):917–947.

Regina Barzilay, Kathleen R. McKeown, and Michael Elhadad. 1999. Information fusion in the context of multi-document summarization. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 550–557, College Park, Maryland, USA. Association for Computational Linguistics.

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

Abhik Bhattacharjee, Tahmid Hasan, Kazi Samin, Md Saiful Islam, M. Sohel Rahman, Anindya Iqbal, and Rifat Shahriyar. 2021. Banglabert: Combating embedding barrier in multilingual models for low-resource language understanding. *CoRR*, abs/2101.00204.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. 2013. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122.

Seniz Demir, İlknur Durgar El-Kahlout, Erdem Unal, and Hamza Kaya. 2012. Turkish paraphrase corpus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 4087–4091, Istanbul, Turkey. European Languages Resources Association (ELRA).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

William B Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Anthony Fader, Luke Zettlemoyer, and Oren Etzioni. 2013. Paraphrase-driven learning for open question answering. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1608–1618, Sofia, Bulgaria. Association for Computational Linguistics.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

E Fonseca, L Santos, Marcelo Criscuolo, and S Aluisio. 2016. Assin: Avaliacao de similaridade semantica e inferencia textual. In *Computational Processing of the Portuguese Language-12th International Conference, Tomar, Portugal*, pages 13–15.

Yun He, Zhuoer Wang, Yin Zhang, Ruihong Huang, and James Caverlee. 2020. Parade: A new dataset for paraphrase identification requiring computer science domain knowledge. *arXiv preprint arXiv:2010.03725*.

Dmitry Kravchenko. 2017. Paraphrase detection using machine translation and textual similarity algorithms. In *Conference on artificial intelligence and natural language*, pages 277–292. Springer.

Alon Lavie and Michael J Denkowski. 2009. The meteor metric for automatic evaluation of machine translation. *Machine translation*, 23(2-3):105–115.

Nitin Madnani, Joel R. Tetreault, and Martin Chodorow. 2012. Re-examining machine translation metrics for paraphrase identification. In *NAACL*.

Alaa Altheneyan; Mohamed Menai. 2019. Arpc a corpus for paraphrase identification in arabic text.

Pedro Javier Ortiz Suárez, Laurent Romary, and Benoît Sagot. 2020. A monolingual approach to contextualized word embeddings for mid-resource languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1703–1714, Online. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Saša Petrović, Miles Osborne, and Victor Lavrenko. 2012. Using paraphrases for improving first story detection in news and twitter. In *Proceedings of the 2012 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 338–346.

Ekaterina Pronoza, Elena Yagunova, and Anton Pronoza. 2016. *Construction of a Russian Paraphrase Corpus: Unsupervised Paraphrase Extraction*, volume 573, pages 146–157.

Sagor Sarker. 2020. Banglabert: Bengali mask language model for bengali language understading.

P. Wallis. 1993. Information retrieval based on paraphrase.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Wei Xu, Chris Callison-Burch, and Bill Dolan. 2015. Semeval-2015 task 1: Paraphrase and semantic similarity in twitter (pit). In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pages 1–11.

Bowei Zhang, Weiwei Sun, Xiaojun Wan, and Zongming Guo. 2019. Pku paraphrase bank: A sentence-level paraphrase corpus for chinese. In *NLPCC*.

A Appendix

A.1 Source Portals for Data Collection

Name	Global Ranking	Country Ranking
prothomalo.com	500	4
jugantor.com	1,193	5
kalerkantho.com	1,646	6
jagonews24.com	1,691	7
bdnews24.com	1,573	8
bd-pratidin.com	2,106	12
banglanews24.com	3,238	16
dhakapost.com	4,545	17
banglatribune.com	3,319	18
ittefaq.com.bd	3,652	21
samakal.com	7,497	27
24livenewspaper.com	7,811	35
rtvonline.com	8,901	36
somoynews.tv	5,275	37
newsbangla24.com	10,987	40
dainikshiksha.com	10,417	41
ntvbd.com	8,935	43
dailyinqilab.com	9,745	44
anandabazar.com	3,415	50
mzamin.com	12,376	63
priyo.com	33,966	169
abplive.com	2,353	227

Table 5: Alexa ranking of different news portals. (Collected on 08 October, 2021)

We used the Alexa ranking[12] to gather news from the most popular sites in the national and international domain. The global ranking and ranking in Bangladesh of the news portals are shown in Table 5.

A.2 Discarded Sentence Pair Examples

While annotating the dataset, we found some sentence pairs where the annotators could not agree if it was a paraphrase or not. We called these sentence pairs debatable. After careful analysis, we found that these sentence pairs are usually partial paraphrases, have partial information of the other sentence, or have uncertain sentence pairs.

- **Partial Paraphrases:** Partial paraphrase occurs when a section of a complex sentence incorporates the paraphrase of another sentence.
- **Partial Information:** One sentence lacks some information, making it impossible to determine if it is a paraphrase or not.
- **Generalization:** Certain phrases is generalized in one sentence, while it is specific in the other one.

[12] https://www.alexa.com/topsites/countries/BD

Model	P	R	F1
Baseline(Random)	38.81	50.00	43.70
Baseline(Majority)	35.00	59.16	43.98
BLEU	76.46	75.85	76.00
METEOR	83.38	83.45	83.34
Unigram (U)	82.71	80.19	78.97
Bigram (B)	78.16	76.32	74.82
Trigram (T)	75.66	65.94	58.13
U+B	80.83	79.04	77.89
U+B+T	80.46	78.36	77.04
Char-2-gram (C2)	81.51	80.80	80.18
Char-3-gram (C3)	83.12	82.09	81.45
Char-4-gram (C4)	82.60	81.41	80.67
Char-5-gram (C5)	81.61	80.19	79.28
C2+C3	82.45	81.75	81.19
C2+C3+C4	82.69	81.89	81.30
C2+C3+C4+C5	82.58	81.48	80.77
U+C2	83.79	82.16	81.36
U+C2+C3	83.79	82.09	81.27
U+C2+C3+C4	83.89	82.16	81.33
All n-grams	83.06	81.41	80.54
Word Embedding	84.98	83.11	82.32
E+U+C2	85.13	83.31	82.56
BanglaBERT	61.58	62.62	59.02
bangla-bert-base	79.23	78.83	78.20
MBERT	**83.73**	**83.79**	**83.74**

Table 6: Validation results from different experiments of baseline, MT metrics, linguistic features, and pre-trained LMs are reported in Precision (P), Recall (R) and weighted-F1 score.

All these issues create a problem to properly classify a pair as a paraphrase or not. Some debatable sentence pairs are added in Table 7.

A.3 Validation Set Results:

To accommodate further research, we provide the development set results in Table 6.

Sentence 1	**Sentence 2**	**Reason**
কোহলির বেঙ্গালুরুর এবারও খালি হাতে বিদায় (Kohli’s Bangalore left empty handed this time)	কোহলিদের বিদায়, টিকে থাকল হায়দরাবাদ (Farewell to Kohli, Hyderabad survived)	Partial Paraphrase
জরিপে এগিয়ে বাইডেন, এরপরও ট্রাম্প যেভাবে জিততে পারেন (Biden ahead in the polls, yet how can Trump win)	ট্রাম্প যেভাবে জয়ী হতে পারেন (The way Trump can win)	
সম্মাননা পেলেন অপূর্ব-মেহজাবীন (Apurba-Mehzabin got the honor)	মেহজাবীনের হাতে সম্মাননা (Honor in the hands of Mehzabin)	Partial Information
নতুন দায়িত্বে আফসানা মিমি (Afsana Mimi in new responsibilities)	শিল্পকলা একাডেমির পরিচালকের দায়িত্বে মিমি ও মিনি (Mimi and Mini are the directors of Shilpakala Academy)	
ঢাবির ঘ' ইউনিটের ভর্তি পরীক্ষা না নেয়ার সিদ্ধান্ত (Decision not to take admission test of DU D unit)	ঢাবির 'ঘ' এবং 'চ' ইউনিট থাকছে না (DU does not have ’D’ and ’F’ units)	
মুম্বাইয়ে হোটেলে অজি ক্রিকেটার ডিন জোন্সের মৃত্যু (Aussie cricketer Dean Jones dies at hotel in Mumbai)	ধারাভাষ্য দিতে এসে অকালেই হৃদরোগে আক্রান্ত হয়ে প্রয়াত প্রখ্যাত ক্রিকেটার (The late famous cricketer suffered a heart attack prematurely when he came to comment)	Generalization
১০০ ছুঁইছুঁই বেশিরভাগ সবজি (Most vegetables touches 100)	কমেনি পেঁয়াজের ঝাঁজ, সবজির বাজারও চড়া (The market for onions and vegetables is also booming)	
যুক্তরাষ্ট্র থেকে ২২৯০ কোটি রুপির অস্ত্র কিনছে ভারত (India is buying arms worth Rs 2,290 crore from the United States)	আমেরিকা থেকে অতিরিক্ত ৭২,০০০ অ্যাসল্ট রাইফেল কিনবে ভারত (India will buy an additional 62,000 assault rifles from the United States)	

Table 7: Examples of debatable sentence pairs.