BnPC: A Gold Standard Corpus for Paraphrase Detection in Bangla, and its Evaluation

Sourav Saha[1*], **Zeshan Ahmed Nobin**[1*], **Mufassir Ahmad Chowdhury**[1*],
Md. Shakirul Hasan Khan Mobin[1*], **Mohammad Ruhul Amin**[2], **Sudipta Kar**[3]
Shahjalal University of Science and Technology, Bangladesh,[1]
{sourav95, zeshan07, mufassir73, shakirul34}@student.sust.edu,[1]
Fordham University, USA,[2] Amazon, USA[3]
mamin17@fordham.edu,[2] sudipkar@amazon.com[3]

Abstract

In this paper, we present a benchmark dataset for paraphrase detection in Bangla. Despite being the sixth most spoken language[1] in the world, paraphrase identification in Bangla is barely explored. Our dataset contains 8,787 ***human-annotated*** sentence pairs collected from 23 newspaper outlets' headlines in four categories. We explored several supervised modeling approaches to benchmark the dataset, including similarity metrics, linguistic features, and fine-tuned BERT models. We also conducted a zero-shot analysis to assess the performance of pre-trained BERT models, and we carried out both zero-shot and few-shot evaluations of the publicly accessible generative language model GPT 3.5 turbo. In the benchmark evaluations, when examining GPT-3.5 using a few-shot modeling approach, it becomes evident that the model can grasp paraphrases in a manner akin to fine-tuned mBERT language models with just a handful of example data points. Within the set of benchmarking trials, the fine-tuned BanglaBERT delivered the most remarkable performance, achieving a weighted-F1 score of 87.91. Noteworthy is that GPT-3.5 excelled in both zero-shot and few-shot experiments, attaining weighted-F1 scores of 51.51 and 80.53, in that order. We also performed a cross-dataset analysis and the outcomes suggest that the model trained in our dataset resembles both diversity and generalization when tested on the other dataset. Finally, we report a human evaluation experiment to obtain a better understanding of the paraphrasing task's limitations. We make our dataset and code publicly available.[2]

Keywords: Paraphrase Identification, Semantic Similarity, Benchmarking Dataset, Cross Dataset Analysis

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., 2013a), 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 value alignment, etc. 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; Xu et al., 2015a; Lan et al., 2017; He et al., 2020a) , Turkish (Demir et al., 2012), Russian (Pronoza et al., 2016), Arabic (Menai, 2019), Portuguese (Fonseca et al., 2016), Chinese (Zhang et al., 2019), among others.

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.

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 by 2050.[3] Bangla is the language of the

[*] Authors have equal contribution

[1] w.wiki/Pss

[2] https://github.com/Mufassir-Chowdhury/BnPC

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

BUCC 2024: The 17th Workshop on Building and Using Comparable Corpora, pages 69–84
20 May, 2024.

people of the Bengal region, now divided between Bangladesh and the Indian state of West Bengal, which are considered to be the region of fastest growing economies.[4] Because 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, the task of Bangla paraphrase identification has received limited attention. Akil et al. (2022) generated a synthetic Bangla paraphrase dataset consisting of 603,672 sentence pairs. Kumar et al. (2022) also experimented with six different NLG tasks across eleven Indic languages including the task of Bangla paraphrase generation. Meanwhile, Scherrer (2020) curated sentential paraphrases on 73 languages including Bangla, for which they considered only 1,440 Bangla sentences.

To address the scarcity of paraphrase detection dataset in Bangla language, we propose BnPC, a gold-standard Bangla paraphrase corpus. We outline the contributions of this study below:

- We propose BnPC, the largest gold standard paraphrase corpus in Bangla, consisting of 8,787 ***human-annotated*** pairs collected from 23 different newspaper outlets in Bangladesh. We present a few examples in Table 1.
- We report a benchmark evaluation on BnPC by exploiting several supervised learning approaches, such as the similarity metrics (BLEU, METEOR), bag-of-words approach (Word and Character n-grams), and fine-tuned language models.
- We carried out both the zero-shot and few-shot experiments over the publicly accessible GPT-3.5 turbo model using BnPC and present shortcomings we observed from GPT-3.5 responses.
- We performed a cross-dataset analysis by fine-tuning a monolingual and a multilingual BERT on BnPC and testing it on several other datasets. We show that models trained on BnPC resembles the capacity to provide better performance on diverse datasets.
- We also conducted a human evaluation experiment to get insights into the paraphrasing task's limitations.

2. Related Work

Over the recent years, a great deal of work has been accomplished in paraphrase detection. We discuss some of the notable works in this section.

Datasets for Paraphrase Identification: MSRP (Dolan and Brockett, 2005; Dolan et al., 2004) is the pioneering hand-labeled dataset extracted using heuristic techniques instead of the traditional machine translation method. Their approach obtained high lexical divergent paraphrase pairs, opening up new dimensions in the paraphrase identification field. Twitter Paraphrase Corpus (PIT-2015) (Xu et al., 2015a) is a realistic and balanced dataset collected from trending topics on Twitter containing a high degree of variation due to the use of informal language as well as more naturally occurring non-paraphrases. Twitter URL Corpus (TUC) (Lan et al., 2017) is a shared URL based growing paraphrase corpus with both formal and informal texts, where the authors mitigate the complications of extracting highly variant natural paraphrase sentence pairs on a large scale. Quora Question Pair (Chen et al., 2017) is a dataset containing interrogative sentence pairs that benefit the Q&A community by assisting in the detection of duplicate questions. PARADE (He et al., 2020b) is a domain-specific dataset where authors formed clusters of definitions focusing same aspect indicated by overlapping term and matched every two definitions from the same cluster together.

Approaches used in Paraphrase Detection: The noteworthy approaches for the task of paraphrase identification are MT metric based classifiers (Eyecioglu and Keller, 2015) combining lexical and compositional features. The modeling approaches include referential and machine translations (Finch et al., 2005; Biçici and Way, 2014), feature based approaches (Zarrella et al., 2015), supervised learning (Vo et al., 2015; Karan et al., 2015) using SVM and logistic regression (Satyapanich et al., 2015; Madnani et al., 2012a; van der Goot and van Noord, 2015), deep learning and BERT based approaches (Zhao and Lan, 2015; Bertero and Fung, 2015; Chandra and Stefanus, 2020).

Bangla Paraphrase Detection: TaPaCo (Scherrer, 2020) is a paraphrase corpus generated by populating a graph from the Tatoeba database and finding equivalent links between the sentence pairs with everyday sentences. They used a crowd-sourced method of paraphrase generation without assessing the capability of the translators. BanglaParaphrase (Akil et al., 2022) curated sentences from a Bangla blogging website using a machine translation (back-translation) and a novel filtering process based on PINC score (Chen and Dolan, 2011) (a metric based on lexical dissimilarity). IndicNLG used pivoting approach (Kumar et al., 2022) to extract paraphrases from a parallel corpus using English as the pivot.

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

In contrast, the BnPC dataset was created from human-generated text from newspaper headlines and labeled by three expert annotators validating all paraphrase pairs using a rigorous process to ensure the quality of the data.

3. Overview of BnPC Dataset

Data Collection: We constructed the BnPC corpus by gathering news headlines from 23 of the most popular[5] Bangla news portals. This is because headlines for similar news tend to be paraphrases. Thus we gathered news on four broad categories: *national, international, sports,* and *entertainment* over the four months starting from September to December of 2020. Alongside visiting individual news websites, we also utilized Google News[6] service to retrieve cluster of similar news, and a similar service from the Pipilika News[7].

Through manual inspection, we formed a total of 145 national, 158 international, 139 sports, and 175 entertainment related news clusters by selecting similar news of identical events. Each cluster contained different headlines focusing on different aspects of the same event reported by various news agencies. We followed different methods of paraphrasing to select paraphrasing pairs. These methods are presented in Table 2.

Annotation: Three of the native Bangla-speaking authors annotated the pairs. Each annotator was trained on different methods of paraphrasing according to Table 2. We decided to use five different paraphrase scores on a scale from 0 to 1 to reach a better labeling consensus among the annotators at the end of the process.

We discuss our score assignment for each of the 5 different paraphrasing decision: (1) “Not Paraphrase”: Score 0; (2) “Not-Paraphrase with Slight Similarity”: Score 0.25; (3) “Undecided”: Score 0.5; (4) “Paraphrase with Lexical Differences”: Score 0.75; and (5) “Paraphrase”: Score 1.0.

During the annotation, we followed the guidelines described in Bhagat and Hovy (2013). We averaged the scores of three annotators. Sample above the threshold score (0.5) were considered as paraphrase and below it as non-paraphrase in the final dataset. We discarded the ones with an average score of 0.5 as the annotators could not agree on whether the pairs were paraphrase or not. These samples were mostly partial paraphrases or had ambiguous meanings. A Fleiss’ Kappa score (Fleiss, 1971) of 0.61 indicates substantial inter-annotator agreement. We present some sample sentence pairs in Table 1.

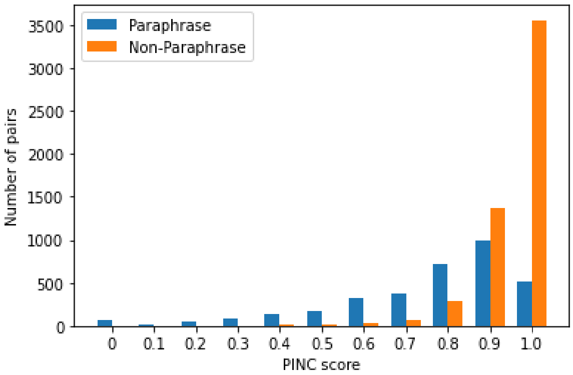

Figure 1: PINC score of paraphrase and non-paraphrase pairs of BnPC. PINC score denotes n-gram dissimilarity between two sentences. High PINC score denotes low lexical overlaps.

Statistics: As per Table 3, the class distribution of the dataset is slightly skewed towards the non-paraphrases, and 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. The high PINC score (Chen and Dolan, 2011) in Figure 1 for both paraphrase and non-paraphrase pairs indicates that the dataset contains more lexically diverse sentences. The diversity among the non-paraphrase pairs is more abundant.

Analysis: Paraphrase identification from real-world data is noisy and follows a wide range of methods compared to synthetically generated pairs. In our BnPC dataset, we analyzed various methods of paraphrases (Table 2). Often times, more than these methods are observed in paraphrase making in Bangla. This makes paraphrase detection in Bangla significantly more challenging for rule-based approaches.

4. Methodology

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

4.1. Evaluation Metric Based Approach

Following Madnani et al. (2012b) and Kravchenko (2017), we investigate paraphrase classifiers using machine translation (MT) evaluation metrics

[5] alexa.com/topsites/countries/BD
[6] news.google.com/?hl=bn
[7] news.pipilika.com

Methods of Paraphrase	Explanation	Sentence1	Sentence2
Change Of Order	Change of order involves changing the order of a word or phrase in a sentence	নতুন অ্যাটর্নি জেনারেল এ এম আমিন উদ্দিন (Newly appointed attorney general is A M Amin Uddin)	আমিন উদ্দিন নতুন এ্যাটর্নি জেনারেল (Amin Uddin is newly appointed attorney general)
Synonym Substitutions	It involves the replacement of a word or phrase of the sentence with one of its synonyms	পিস্তল কিনে ফেসবুকে ছবি দিলেন এমপি! (M.P. posted a photo on facebook after purchasing a pistol)	পিস্তল কিনে ফেসবুকে ছবি দিলেন সংসদ সদস্য (The Member of Parliament posted a picture on Facebook after buying a pistol)
Verbatim	It is a type of plagiarism where a sentence is copied without changing any aspect of the sentence	লাইফ সাপোর্টে ব্যারিস্টার রফিক-উল হক (Barrister Rafiq-ul-Haque on life support)	লাইফ সাপোর্টে ব্যারিস্টার রফিক-উল হক (Barrister Rafiq-ul-Haque on life support)
Ellipsis	Ellipsis involves the omission of clauses that are understood from the context of the remaining sentence	করোনায় আরো ১৮ জনের মৃত্যু (18 more people die from Corona)	করোনায় আরো ১৮ মৃত্যু (18 more die from Corona)
Punctuation Changes	Punctuation changes involve the change of punctuation used in the sentence	গুগল, ফেসবুক ও ইউটিউব থেকে রাজস্ব আদায়ের নির্দেশ	গুগল-ফেসবুক-ইউটিউব থেকে রাজস্ব আদায়ের নির্দেশ
Emphasization	Emphasization is a type of paraphrase where the exact same	চলতি মাসে দেশে আঘাত হানতে পারে ঘূর্ণিঝড় (A cyclone is expected to strike our country later this month)	চলতি মাসেই আঘাত হানতে পারে ঘূর্ণিঝড় (A cyclone is set to hit the country this very month)
Abbreviation	It involves shortened form of a word or phrase in one of the pairs	ইউরোপীয় ইউনিয়নের সঙ্গে সম্পর্কচ্ছেদের হুঁশিয়ারি রাশিয়ার (Russia issues a threat to sever ties with the European Union)	ইইউ ছাড়ার হুমকি দিল রাশিয়া (Russia threatens to leave EU)

Table 2: This table presents different methods of paraphrasing in our BnPC dataset. Most of the definitions are picked from Zhou et al. (2022).‡

	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 3: 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.

Root Word	
হার	
Type	**Example**
Root	৭৮ দিনে করোনা শনাক্তের হার সর্বোচ্চ (Corona detection rate is highest in 7 days)
Prefix	প্রথমবার ওয়েব সিরিজে জুটি বাঁধছেন সোহম-শ্রাবন্তী দর্শকদের উপহার (উপ + হার) দেবেন থ্রিলার লাভস্টোরি (Soham-Sravanti to tie the knot for the first time in web series, thriller Love Story to present to viewers)
Suffix	তছেন সু চি, আবার হারছে (হার + ছে) রোহিঙ্গরা? (Suu Kyi is winning, Rohingyas are losing again?)
Concatination	২৪ ঘণ্টায় করোনায় মৃত্যুহার (মৃত্যু + হার) কমেছে (Mortality rate in Corona has decreased in 24 hours)

Table 4: Examples of prefix, suffix, and concatenation usage in Bangla from our dataset.‡

like BLEU (Papineni et al., 2002a) and METEOR (Lavie and Denkowski, 2009) as these metrics provide a notion of lexical 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:

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

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

Here, α is a threshold, whose value was set by maximizing the performance on the training set (α=0.115 for BLEU and α=0.136 for METEOR).

4.2. 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.

4.3. Language Models

Pre-trained language models, particularly variants of BERT, have shown superior performance in a variety of natural language tasks. On the other hand, recent LLMs have shown superior quality in performing different NLP domain tasks. We use the Multilingual BERT (mBERT) (Devlin et al., 2018), RoBERTa (Liu et al., 2019c), XLM-RoBERTa (Conneau et al., 2019), and three different monolingual BERT models pre-trained on Bangla (Sarker, 2020; Bhattacharjee et al., 2021; Diskin et al., 2021)[8][9][10] from HuggingFace transformers (Wolf et al., 2020) and fine tune the

[8] huggingface.co/csebuetnlp/banglabert

[9] huggingface.co/sagorsarker/bangla-bert-base

[10] huggingface.co/neuropark/sahajBERT

binary prediction layer. We reported the zero-shot performance of mBERT, XLM-RoBERTa, BanglaBERT. Additionally, we perform zero-shot and few-shot approaches on publically available GPT 3.5 turbo. BanglaBERT (Bhattacharjee et al., 2021) was trained on 27.5 GB data crawled from 110 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).

5. Experiments and Results

5.1. Experimental Setup

We use 70% of the data for training, and equally divide the rest for development and testing. For the metric-based approaches, we remove the punctuations and for BOW-based methods, we preprocess 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 pretrained language models, we fine-tune ($\lambda{=}2{*}10^{-5}$, batch size 32) the models for 5 epochs with early stopping. For gpt-3.5-prompting we used the ChatGPT Platform API [11] with the following parameters: temperature=0 (0 for deterministic output), max_tokens=256, top_p=1, frequency_penalty=0, presence_penalty=0.

5.2. Results & Analysis

Table 5 presents the precision, recall, and weighted F1 scores of different models on the test set. 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 the sentence pairs. As a consequence, METEOR exhibits better performance for the task.

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 a 77.97 F1 score and longer ngrams' F1 score decreases gradually. However, character n-grams show better performance than the word n-grams in general. Usage of prefixes, suffixes, and word concatenation is heavy in Bangla, which we believe is the reason for the strength of character n-grams (Table 4). The combination of unigram and character bigram 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.

Model	P	R	F1
Baseline (Random)	50.56	50.67	49.62
Baseline (Majority)	34.86	59.04	43.83
BLEU	76.95	76.76	76.10
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 (E)	77.53	77.04	76.24
U+C2+E	78.83	78.19	77.41
bangla-bert-base (Zero-Shot)	51.54	58.68	45.02
mBERT (Zero-Shot)	26.39	48.87	23.82
XLM-RoBERTa (Zero-Shot)	34.86	59.04	43.83
sahajBERT (Zero-Shot)	55.29	48.85	46.85
BanglaBERT (Zero-Shot)	59.67	51.92	48.79
gpt-3.5-turbo (Zero-Shot)	71.69	62.27	51.51
gpt-3.5-turbo (Few-Shot)	80.53	80.63	80.53
bangla-bert-base (Sarker, 2020)	75.85	76.04	75.75
mBERT (Devlin et al., 2018)	82.54	82.42	82.47
XLM-RoBERTa (Conneau et al., 2019)	86.11	86.08	85.96
sahajBERT (Diskin et al., 2021)	86.55	86.37	86.19
BanglaBERT (Bhattacharjee et al., 2021)	**87.92**	**87.95**	**87.91**

Table 5: Results from different experiments of baseline, MT metrics, linguistic features, and pretrained LMs are reported in Precision (P), Recall (R) and weighted-F1 score.

Zero-shot performance of the models is significantly low (even compared to feature-based approaches). Among the zero-shot performance of the models, the GPT 3.5 turbo achieves the best results with an F1 score of 51.51. Interestingly, the GPT 3.5 turbo few-shot exhibits a significant performance boost. The few-shot (4-shot, two paraphrases, and two non-paraphrases) achieves an F1 score of 80.53 closer to the finetuning result of some LMs and surpassing all feature-based approaches indicating the paraphrase detection capabilities of large language models. We provide some interesting examples of LLM's failure in Table 7.

On our dataset, the best-performing model is BanglaBERT (Bhattacharjee et al., 2021), outperforming XLM-RoBERTa by a close margin. BanglaBERT is pre-trained on the highest volume of Bangla data (27.5 GB) to date. The competitive performance of XLM-RoBERTa results from its effective cross-lingual transfer learning.

To provide a performance comparison of the best-performing multilingual model with other

[11] platform.openai.com/

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
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 6: 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 BanglaBERT.[‡]

Sentence 1	Sentence 2	Reason
খুলনায় ২৪ ঘণ্টা বন্ধ থাকবে পরিবহন (Transportation will be closed in Khulna for 24 hours)	খুলনায় পরিবহন চলাচল বন্ধ ঘোষণা (Transport closure announced in Khulna)	Unless it's direct syntactic similarity the LLM model fails in case of bangla. The broader context is easier for humans to comprehend.
চাঁদপুরে আগুনে পুড়ে স্কুল শিক্ষিকার রহস্যজনক মৃত্যু (Mysterious death of school teacher in fire in Chandpur)	আগুনে অঙ্গার শিক্ষিকা (Teacher turned into cinder in a fire)	LLMs struggle with idiomatic expressions, often misinterpreting them.
ব্রিটেনে আর ফিরতে পারবেন না শামীমা (Shamima will not be able to return to Britain)	শামীমার যুক্তরাজ্যে ফেরার আবেদন নাকচ করলেন আদালত (The court rejected Shamima's request to return to the UK)	LLMs may not detect paraphrases when two sentences convey the same news but use different subjects.
ফের নানা হলেন ডিপজল (Deepzal became grandfather again)	মা হলেন ডিপজল কন্যা ওলিজা (Deepzal daughter Oliza became a mother)	LLMs may struggle to follow logical syllogisms accurately.
১০০১ দিন পর জেল থেকে মুক্তি পেলেন সৌদি অধিকারকর্মী (Saudi rights activist released from jail after 1001 days)	৩ বছর পর সৌদির নারী অধিকার কর্মী লুজাইনের মুক্তি (Saudi women's rights activist Luzain released from jail after 3 years)	LLMs can be confused by changes in units when interpreting or processing information.

Table 7: Examples of sentence pairs where LLMs fail to classify using few-shot approach.[‡]

datasets, we fine-tune XLM-RoBERTa on other substantial English datasets with the identical experimental setup. The F1 scores are 90.78 on MSRP (Dolan and Brockett, 2005)), 75.01 on PARADE (He et al., 2020)), and 88.31 on PIT (Xu et al., 2015a)). 85.96 F1 on BnPC falls in between these scores and provides a competitive benchmark result.

5.3. Comparison of Datasets

Cross Dataset Generation: As the other datasets don't have any non-paraphrase pairs, we added the non-paraphrase from our dataset. To compare the quality of the contemporary datasets with the BnPC, we also maintained the paraphrase and non-paraphrase ratio of BnPC on the other datasets. For BanglaParaphrase and IndicNLG we randomly sampled the equivalent number of paraphrases as BnPC and appended all our non-paraphrase pairs to them. Since these two datasets are substantially larger than BnPC we repeated this process three times for brevity and experimented with each of these datasets and averaged the results. Since TaPaCo has a smaller size than BnPC, we appended only a random portion of our non-paraphrase pairs to maintain the overall paraphrase and non-paraphrase ratio equivalent to BnPC. To ensure unbiased experiments, we include non-paraphrase pairs from our train, test, and validation sets into the corresponding sets of other datasets. (Fig: 2)

Results: To conduct cross-dataset testing, we implement both monolingual (sahajBERT) and multilingual (mBERT) models across various merged datasets. The models trained on BnPC consistently perform well across all datasets, achieving a minimum F1 score of 69.97 on IndicNLG. On the other hand, models trained on BanglaParaphrase excel across most datasets and face a downfall of performance on our gold standard BnPC dataset, scoring below 50%, while surpassing the 92% F1 score on other datasets. Models trained on TaPaCo demonstrate strong performance across most tests, with the notable exception of BnPC, where they yield the lowest F1 score of 44% among all the cross-dataset experiments. IndicNLG proves to be a strong performer across synthetic datasets, consistently achieving over 97%, and it delivers a respectable F1 score of 57.32 on our BnPC dataset. In summary, models trained on synthetic datasets display subpar performance when tested on our gold standard dataset.

We obtain the context of the paraphrase pairs by using BLEU (Papineni et al., 2002a) and ROUGE

[‡]denotes the sentences in these tables were translated using Google Translator for the clarity of the non Bangla speakers.

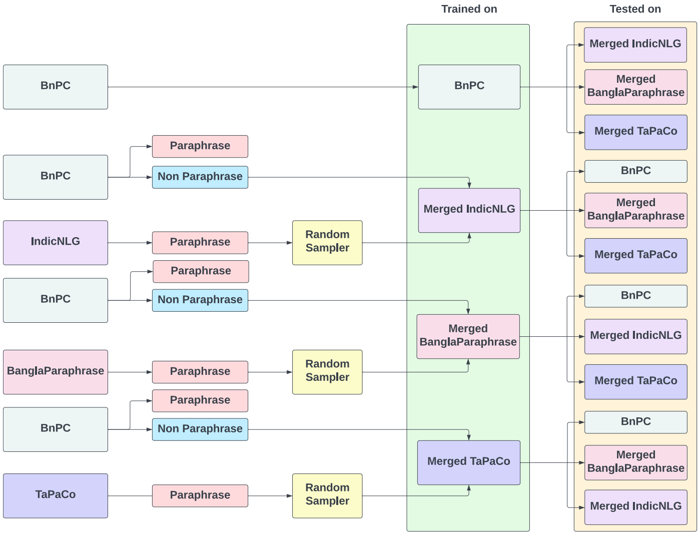

Figure 2: The workflow diagram of cross-dataset test. It shows the procedures for generating the merged datasets for cross-dataset experiments and the experimental procedures.

Model	Trained On	Tested On (BnPC)			Tested On (BanglaParaphrase)			Tested On (TaPaCo)			Tested On (IndicNLG)		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1
SahajBERT	BnPC	86.55	86.37	86.19	94.59	94.18	94.22	86.42	86.37	86.24	73.25	71.81	69.67
	BanglaParaphrase	73.21	61.31	49.04	99.52	99.54	99.54	98.84	98.83	98.82	93.56	92.87	92.73
	TaPaCo	75.84	59.11	**44.00**	84.16	78.35	76.14	100.00	100.00	**100.00**	90.08	88.07	87.60
	IndicNLG	71.02	64.94	57.32	97.61	97.55	97.56	98.29	98.22	98.22	97.18	97.15	97.15
mBERT	BnPC	80.61	80.63	80.62	89.85	88.61	88.70	84.88	84.54	84.62	65.14	65.71	63.75
	BanglaParaphrase	72.66	62.38	51.57	99.23	99.23	99.23	89.60	87.91	87.48	78.83	70.35	65.22
	TaPaCo	72.36	60.25	**46.77**	80.67	72.14	67.70	99.87	99.87	**99.87**	87.28	84.05	83.09
	IndicNLG	69.68	65.23	58.40	96.60	96.54	96.55	97.96	97.85	97.85	96.36	96.32	96.33

Table 8: The table shows the cross-dataset performance of monolingual (SahajBERT) and multilingual (mBERT) models. It contains precision (P), recall (R), and weighted-F1 scores of the models. The worst performances (row-wise) are shown in red and the best performances (row-wise) are shown in blue.

(Lin, 2004) metrics. We see that TaPaCo has the highest n-gram similarity since it mostly consists of simple and small sentences. IndicNLG shows the lowest n-gram similarity across all the metrics. BanglaParaphrase and BnPC have similar n-gram similarity across the metrics indicating a moderate n-gram overlap.

Analysis: From Table 8, we see that models trained on synthetic datasets show poor performance on human-generated data. On the other hand, models trained on BnPC show decent performance on synthetic datasets. Despite BnPC having moderate n-gram similarities, the failure of models trained on other datasets and tested on BnPC can originate from the wide distribution of paraphrases across the PINC Score spectrum. The BnPC paraphrases are spread across the spectrum from 0.0-1.0, which is absent in other datasets with the single highest being only 36.25% of the samples on 0.9. 82% of the data is within 0.6-1.0. and the other 18% data falls within 0.0-0.5 which is the highest among other datasets.

The monolingual Model trained on BanglaParaphrase did well except on BnPC and the Multilingual model trained on BanglaParaphrase did moderate performance on BnPC and IndicNLG. This can stem from the fact that BanglaParaphrase has paraphrases (∼98%) mostly spread within 0.6-0.9 PINC score with 44.48% data on 0.8. This makes it hard to perform well on a dataset with more distributed n-gram similarity and similar size. Models trained on other datasets and tested on BanglaParaphrase show a better performance except for TaPaCo which might originate from the smaller size of TaPaCo. TaPaCo has 42% smaller size than the other datasets. Models trained on other datasets show good performance on TaPaCo. This can be traced back to the smaller size of the dataset, sentences, and high n-gram similarity of TaPaCo paraphrase pairs shown in Figure 3 which is the highest among all the datasets. Making it easier to identify paraphrases in simple and small sentences. Models trained on TaPaCo show poor performances on all the datasets except on IndicNLG. IndicNLG has the lowest n-gram similarity among all the datasets. Because of this, models tested on IndicNLG show comparatively weaker performance.

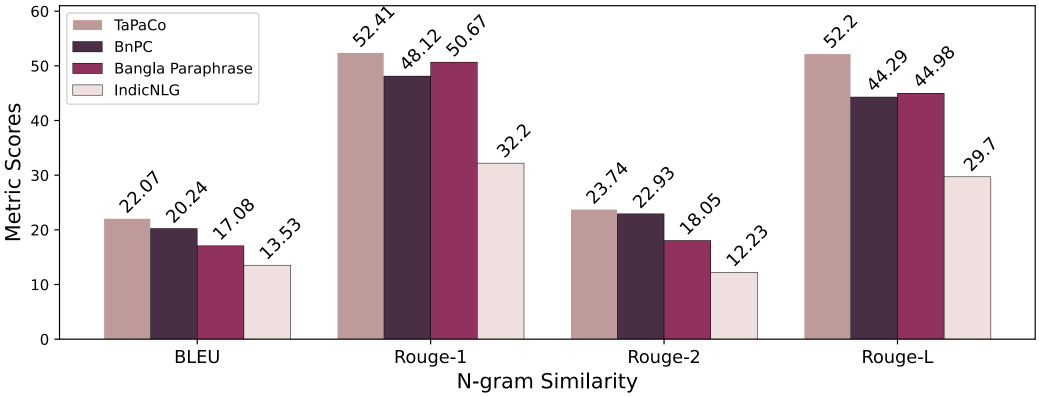

Figure 3: This Figure shows the N-gram similarity comparison of the datasets. For comparing the N-gram similarity we implement BLEU, [Rouge-N, Rouge-L](Lin, 2004) methods.

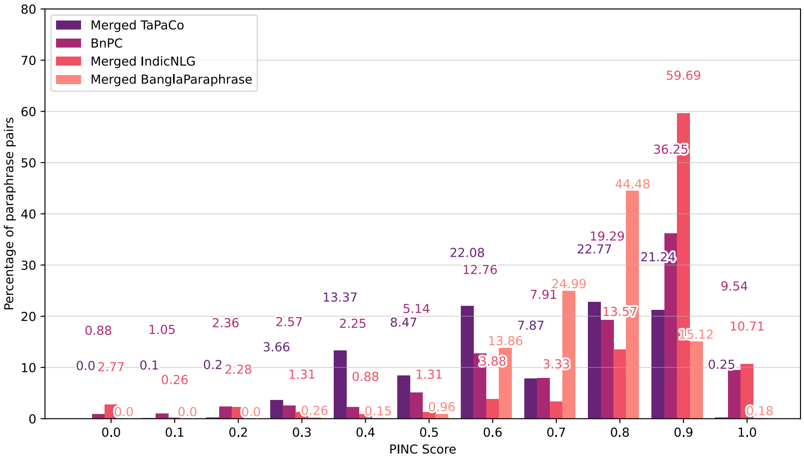

Figure 4: This Figure shows the PINC score comparison of the datasets. PINC score denotes n-gram dissimilarity between two sentences. High PINC score denotes low lexical overlaps between the sentence pairs.

Models trained on IndicNLG show a good performance except on BnPC. Figure 4 shows that almost 60% of their paraphrase pairs stem 0.9 PINC score. This is a probable reason for the IndicNLG's poor performance on the BnPC dataset. We exhibit that models trained with lower n-gram similarity tend to do well on datasets that have higher n-gram similarity on paraphrase identification task.

5.4. 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 native Bangla-speaking undergraduate students from different majors on a voluntary basis to ensure diversity in subjects. After instructing them about the task, we asked 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. Using Fleiss' Kappa metric, we calculate the inter-annotator agreement of those pupils and get a score of 0.47. The best-performing model's F1 score of 87.98 on this sample of data indicates that the job can be more difficult for humans to accomplish.

Analyzing the errors and interviewing the human subjects, we find that the main reasons are lack of domain knowledge, presence of numbers in the sentences, and pairs with long overlaps of spans. (Table 6).

6. Conclusion and Future Works

In this paper, we propose BnPC, the largest handcrafted Bangla dataset for paraphrase detection. Through our investigations to develop a benchmark classifier, we find that lexical features like character n-grams show competitive performance in identifying paraphrases. Similar performance can be achieved by simply using the machine translation metric-based classifiers. From our experiments, we see that the monolingual model BanglaBERT slightly outperforms the multilingual model XLM-RoBERTa on the BnPC dataset. Also, we find the GPT-3.5 turbo performs almost as well as fine-tuned language models. Our cross-dataset analysis shows that models trained on our dataset generalize more compared to contemporary datasets and we provide some quantitative analysis differentiating the datasets. Our dataset comprises formal data from newspaper headlines. So, a good direction for future work can be extending this dataset with different domains and topics' data, for example, conversational data. We release the corpus publicly to foster further work in this area.

Limitations

The study has some potential limitations. One potential limitation is that our dataset is comprised of formal data from news headlines which is different from the noisy data on social media. Social media data generally contains misspellings, and slang words creating challenges for paraphrase detection tasks, which is absent in our dataset. Other potential sources for curating a paraphrase dataset include blogs, books, and various academic writings. Moreover, our dataset comprises roughly 9K data leaving the scope for extending the dataset in the future.

Ethical Considerations

Dataset Release: The Copy Right Act. 2000[12] of People's Republic of Bangladesh allows reproduction and public release of copyright materials for non-commercial research proposals. We will release our BnPC dataset under a non-commercial license. Publicizing other supplementary materials like codes won't cause any copyright infringements.

Annotators' Compensation: All the annotators participated voluntarily in this research work.

Quality Assurance of the Dataset: All the annotations were done by native Bangla speakers. The Fleiss' Kappa score of our dataset showed substantial agreement, ensuring the quality of our dataset.

7. Bibliographical References

Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. Semeval-2012 task 6: A pilot on semantic textual similarity. In ** SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 385–393.

Alfred V. Aho and Jeffrey D. Ullman. 1972. *The Theory of Parsing, Translation and Compiling*, volume 1. Prentice-Hall, Englewood Cliffs, NJ.

Ajwad Akil, Najrin Sultana, Abhik Bhattacharjee, and Rifat Shahriyar. 2022. BanglaParaphrase: A high-quality Bangla paraphrase dataset. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 261–272, Online only. Association for Computational Linguistics.

Abdullah Al Hadi, Md. Yasin Ali Khan, and Md. Abu Sayed. 2016. Extracting semantic relatedness for bangla words. In *2016 5th International Conference on Informatics, Electronics and Vision (ICIEV)*, pages 10–14.

Firoj Alam, Shammur Absar Chowdhury, and Sheak Rashed Haider Noori. 2016. Bidirectional lstms—crfs networks for bangla pos tagging. In *19th International Conference on Computer and Information Technology (ICCIT), 2016*, pages 377–382. IEEE.

American Psychological Association. 1983. *Publications Manual*. American Psychological Association, Washington, DC.

Rie Kubota Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853.

Galen Andrew and Jianfeng Gao. 2007. Scalable training of L1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40.

[12] http://copyrightoffice.portal.gov.bd/sites/default/files/files/copyrightoffice.portal.gov.bd/law/121de2e9_9bc9_4944_bfef_0a12af0864a5/Copyright,2000(1)%20(2).pdf

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

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 and Lillian Lee. 2003. Learning to paraphrase: An unsupervised approach using multiple-sequence alignment. *arXiv preprint cs/0304006*.

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.

Dario Bertero and Pascale Fung. 2015. Hltc-hkust: A neural network paraphrase classifier using translation metrics, semantic roles and lexical similarity features. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 23–28.

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, Wasi Uddin Ahmad, Anindya Iqbal, M. Sohel Rahman, and Rifat Shahriyar. 2021. Banglabert: Language model pretraining and benchmarks for low-resource language understanding evaluation in bangla.

Ergun Biçici and Andy Way. 2014. Rtm-dcu: Referential translation machines for semantic similarity.

Victoria Bobicev and Marina Sokolova. 2017. Inter-annotator agreement in sentiment analysis: Machine learning perspective. In *RANLP*, volume 97.

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.

Andreas Chandra and Ruben Stefanus. 2020. Experiments on paraphrase identification using quora question pairs dataset.

Ashok K. Chandra, Dexter C. Kozen, and Larry J. Stockmeyer. 1981. Alternation. *Journal of the Association for Computing Machinery*, 28(1):114–133.

David Chen and William Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 190–200, Portland, Oregon, USA. Association for Computational Linguistics.

Zihan Chen, Hongbo Zhang, Xiaoji Zhang, and Leqi Zhao. 2017. Quora question pairs.

Jianpeng Cheng and Dimitri Kartsaklis. 2015. Syntax-aware multi-sense word embeddings for deep compositional models of meaning. *arXiv preprint arXiv:1508.02354*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

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'12)*, pages 4087–4091, Istanbul, Turkey. European Language Resources Association (ELRA).

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, pages 376–380.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Michael Diskin, Alexey Bukhtiyarov, Max Ryabinin, Lucile Saulnier, Quentin Lhoest, Anton Sinitsin, Dmitriy Popov, Dmitry V. Pyrkin, Maxim Kashirin, Alexander Borzunov, Albert Villanova del Moral, Denis Mazur, Ilia Kobelev, Yacine Jernite, Thomas Wolf, and Gennady Pekhimenko. 2021. Distributed deep learning in open collaborations. *CoRR*, abs/2106.10207.

William Dolan, Chris Quirk, Chris Brockett, and Bill Dolan. 2004. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources.

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

Qingxiu Dong, Xiaojun Wan, and Yue Cao. 2021. Parasci: A large scientific paraphrase dataset for longer paraphrase generation. *arXiv preprint arXiv:2101.08382*.

Asli Eyecioglu and Bill Keller. 2015. Asobek: Twitter paraphrase identification with simple overlap features and svms in proceedings of 9th international workshop on semantic evaluation (semeval).

Asli Eyecioglu and Bill Keller. 2016. Constructing a turkish corpus for paraphrase identification and semantic similarity. *Lecture Notes in Computer Science*, Computational Linguistics and Intelligent Text Processing. Proceedings of the 17th International Conference on Intelligent Text Processing and Computational Linguistics.:562–574.

Anthony Fader, Luke Zettlemoyer, and Oren Etzioni. 2013a. 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.

Anthony Fader, Luke Zettlemoyer, and Oren Etzioni. 2013b. 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.

Samuel Fernando and Mark Stevenson. 2008. A semantic similarity approach to paraphrase detection. In *Proceedings of the 11th annual research colloquium of the UK special interest group for computational linguistics*, pages 45–52.

Andrew Finch, Young-Sook Hwang, and Eiichiro Sumita. 2005. Using machine translation evaluation techniques to determine sentence-level semantic equivalence. In *Proceedings of the third international workshop on paraphrasing (IWP2005)*.

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.

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. Ppdb: The paraphrase database. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 758–764.

Dan Gusfield. 1997. *Algorithms on Strings, Trees and Sequences*. Cambridge University Press, Cambridge, UK.

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

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

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Kyuyeon Hwang and Wonyong Sung. 2017. Character-level language modeling with hierarchical recurrent neural networks. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5720–5724. IEEE.

Ali Ibrahim, Boris Katz, and Jimmy Lin. 2003. Extracting structural paraphrases from aligned monolingual corpora. Technical report, MASSACHUSETTS INST OF TECH CAMBRIDGE ARTIFICIAL INTELLIGENCE LAB.

Shankar Iyer, Nikhil Dandekar, and Kornél Csernai. 2017. First quora dataset release: Question pairs. data. quora. com.

Yangfeng Ji and Jacob Eisenstein. 2013. Discriminative improvements to distributional sentence similarity. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 891–896.

Rafael-Michael Karampatsis. 2015. Cdtds: Predicting paraphrases in twitter via support vector regression. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pages 75–79.

Mladen Karan, Goran Glavaš, Jan Šnajder, Bojana Dalbelo Bašić, Ivan Vulic, and Marie-Francine Moens. 2015. Tklbliir: Detecting twitter paraphrases with tweetingjay. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 70–74. ACL; East Stroudsburg, PA.

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

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

Wuwei Lan, Siyu Qiu, Hua He, and Wei Xu. 2017. A continuously growing dataset of sentential paraphrases. *arXiv preprint arXiv:1708.00391*.

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

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019a. Improving multi-task deep neural networks via knowledge distillation for natural language understanding. *arXiv preprint arXiv:1904.09482*.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019b. Multi-task deep neural networks for natural language understanding. *arXiv preprint arXiv:1901.11504*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019c. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Edward Loper and Steven Bird. 2002. Nltk: The natural language toolkit. *arXiv preprint cs/0205028*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization.

Nitin Madnani, Joel Tetreault, and Martin Chodorow. 2012a. Re-examining machine translation metrics for paraphrase identification. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 182–190.

Nitin Madnani, Joel R. Tetreault, and Martin Chodorow. 2012b. 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.

Rajat Pandit, Saptarshi Sengupta, Sudip Kumar Naskar, Niladri Sekhar Dash, and Mohini Mohan Sardar. 2019a. Improving semantic similarity with cross-lingual resources: A study in bangla—a low resourced language. In *Informatics*, volume 6, page 19. Multidisciplinary Digital Publishing Institute.

Rajat Pandit, Saptarshi Sengupta, Sudip Kumar Naskar, Niladri Sekhar Dash, and Mohini Mohan Sardar. 2019b. Improving semantic similarity with cross-lingual resources: A study in bangla—a low resourced language. *Informatics*, 6(2).

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002a. 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.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002b. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Diana Pérez and Enrique Alfonseca. 2005. Application of the bleu algorithm for recognising textual entailments. In *Proceedings of the First Challenge Workshop Recognising Textual Entailment*, pages 9–12. Citeseer.

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.

Jason Phang, Thibault Févry, and Samuel R Bowman. 2018. Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks. *arXiv preprint arXiv:1811.01088*.

Martin Potthast, Benno Stein, Alberto Barrón-Cedeño, and Paolo Rosso. 2010. An evaluation framework for plagiarism detection. In *Coling 2010: Posters*, pages 997–1005.

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

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Mohammad Sadegh Rasooli and Joel R. Tetreault. 2015. Yara parser: A fast and accurate dependency parser. *Computing Research Repository*, arXiv:1503.06733. Version 2.

Dwijen Rudrapal, Amitava Das, and Baby Bhattacharya. 2015. Measuring semantic similarity for bengali tweets using wordnet. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 537–544.

Vasile Rus, Rajendra Banjade, and Mihai C Lintean. 2014. On paraphrase identification corpora. In *LREC*, pages 2422–2429. Citeseer.

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

Taneeya Satyapanich, Hang Gao, and Tim Finin. 2015. Ebiquity: Paraphrase and semantic similarity in twitter using skipgrams. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 51–55.

Yves Scherrer. 2020. TaPaCo: A corpus of sentential paraphrases for 73 languages. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6868–6873, Marseille, France. European Language Resources Association.

Md Shajalal and Masaki Aono. 2018. Semantic textual similarity in bengali text. pages 1–5.

Manjira Sinha, Tirthankar Dasgupta, Abhik Jana, and Anupam Basu. 2014. Article: Design and development of a bangla semantic lexicon and semantic similarity measure. *International Journal of Computer Applications*, 95(5):8–16. Full text available.

Manjira Sinha, Abhik Jana, Tirthankar Dasgupta, and Anupam Basu. 2012. A new semantic lexicon and similarity measure in Bangla. In *Proceedings of the 3rd Workshop on Cognitive Aspects of the Lexicon*, pages 171–182, Mumbai, India. The COLING 2012 Organizing Committee.

Gaurav Singh Tomar, Thyago Duque, Oscar Täckström, Jakob Uszkoreit, and Dipanjan Das. 2017. Neural paraphrase identification of questions with noisy pretraining. *arXiv preprint arXiv:1704.04565*.

Rob van der Goot and Gertjan van Noord. 2015. Rob: Using semantic meaning to recognize paraphrases. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 40–44.

Ngoc Phuoc An Vo, Simone Magnolini, and Octavian Popescu. 2015. Fbk-hlt: An effective system for paraphrase identification and semantic similarity in twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 29–33.

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

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

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. 2015a. 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, Denver, Colorado. Association for Computational Linguistics.

Wei Xu, Chris Callison-Burch, and Bill Dolan. 2015b. 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.

Fabio Massimo Zanzotto, Marco Pennacchiotti, and Kostas Tsioutsiouliklis. 2011. Linguistic redundancy in twitter. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 659–669.

Guido Zarrella, John Henderson, Elizabeth Merkhofer, and Laura Strickhart. 2015. Mitre: Seven systems for semantic similarity in tweets. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pages 12–17.

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

Jiang Zhao and Man Lan. 2015. Ecnu: Leveraging word embeddings to boost performance for paraphrase in twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 34–39.

Shiqi Zhao, Haifeng Wang, Ting Liu, and Sheng Li. 2009. Extracting paraphrase patterns from bilingual parallel corpora. *Nat. Lang. Eng.*, 15(4):503–526.

Chao Zhou, Cheng Qiu, and Daniel E. Acuna. 2022. Paraphrase identification with deep learning: A review of datasets and methods.

Donglai Zhu, Hengshuai Yao, Bei Jiang, and Peng Yu. 2018. Negative log likelihood ratio loss for deep neural network classification.

8. Language Resource References

Akil, Ajwad and Sultana, Najrin and Bhattacharjee, Abhik and Shahriyar, Rifat. 2022. *BanglaParaphrase: A High-Quality Bangla Paraphrase Dataset*. Association for Computational Linguistics. PID https://doi.org/10.48550/arXiv.2210.05109.

Dolan, Bill and Brockett, Chris. 2005. *Automatically Constructing a Corpus of Sentential Paraphrases*. Asia Federation of Natural Language Processing, Third International Workshop on Paraphrasing (IWP2005). PID https://www.microsoft.com/en-us/research/publication/automatically-constructing-a-corpus-of-sentential-paraphrases/.

He, Yun and Wang, Zhuoer and Zhang, Yin and Huang, Ruihong and Caverlee, James. 2020. *PARADE: A New Dataset for Paraphrase Identification Requiring Computer Science Domain Knowledge*. Association for Computational Linguistics. PID https://doi.org/10.18653/v1/2020.emnlp-main.611.

Kumar, Aman and Shrotriya, Himani and Sahu, Prachi and Mishra, Amogh and Dabre, Raj and Puduppully, Ratish and Kunchukuttan, Anoop and Khapra, Mitesh M. and Kumar, Pratyush. 2022. *IndicNLG Benchmark: Multilingual Datasets for Diverse NLG Tasks in Indic Languages*. Association for Computational Linguistics. PID https://doi.org/10.18653/v1/2022.emnlp-main.360.

Scherrer, Yves. 2020. *TaPaCo: A Corpus of Sentential Paraphrases for 73 Languages v.1*. Zenodo. PID https://doi.org/10.5281/zenodo.3707949.

Xu, Wei and Callison-Burch, Chris and Dolan, Bill. 2015. *SemEval-2015 Task 1: Paraphrase and Semantic Similarity in Twitter (PIT)*. Association for Computational Linguistics. PID https://doi.org/10.18653/v1/S15-2001.

9. Appendix

9.1. 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 9: Alexa ranking of different news portals. (Collected on 08 October, 2021)

We used the Alexa ranking[13] to gather news from the most popular sites in the national and international domains. The global ranking and ranking in Bangladesh of the news portals are shown in Table 9.

9.2. 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.

All these issues create a problem to properly classify a pair as a paraphrase or not. Some debatable sentence pairs are added in Table 10.

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

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 10: Examples of debatable sentence pairs.