International Journal of Engineering Inventions
e-ISSN: 2278-7461, p-ISSN: 2319-6491
Volume 13, Issue 3 [March. 2024] PP: 178-185

Comparative Study of Sentence Paraphrasing In Indian Languages

Kartiki Kotawar[1], Maithili Tawde[2], Niraj Waghchoure[3], Taniksha Datar[4] , Prof. Dr. Arati Deshpande[5]
Department of Computer Engineering, SCTR's Pune Institute of Computer Technology, Pune, India

Abstract *- Paraphrasing, the process of expressing the same meaning in alternative linguistic forms, is a fundamental task in natural language processing. Its importance is particularly pronounced in the context of Indian languages, characterized by rich linguistic diversity and cultural nuances. However, the development of effective paraphrasing models for these languages presents unique challenges, including limited paraphrase datasets and complex linguistic variations. The aim is to the advancement of paraphrasing techniques in Indian languages by using transformer models, specifically Muril. By bridging the gap between cutting-edge NLP technology and the linguistic complexities of India, we aim to provide practical insights and solutions for researchers, developers, and language enthusiasts interested in enhancing paraphrasing capabilities in this diverse linguistic landscape.*

Keywords: *alternate linguistic forms, cutting-edge, MURIL*

Date of Submission: 16-03-2024 Date of acceptance: 31-03-2024

I. INTRODUCTION

Language is the most common and powerful tool used for communication and connecting people across diverse cultures and regions. The ability to adapt and manipulate text in multiple languages is essential for effective communication. To address this need, we introduce a Large Language Model (LLM) based approach designed to rephrase, simplify, shorten, and stylize text in several Indian languages, including Hindi, Marathi, Telugu, Tamil, Punjabi, and Odia. By harnessing the power of LLMs, we aim to bridge the linguistic and cultural gaps that can often hinder effective communication. Our approach can be invaluable for various purposes, such as content creation, translation, education, and accessibility.

Key features of our LLM-based approach include:

Rephrasing: Our system can rephrase text, preserving the original message's meaning while adapting it to different styles, tones, or structures. This can be invaluable for content localization and enhancement.

Simplification: We provide the ability to simplify complex language, making information more accessible to a broader audience. This is particularly useful for educational materials and ensuring inclusivity.

Shortening: Our system can compress lengthy text while retaining its core message.

Provide administrators with effective tools to manage user accounts, roles, and permissions efficiently.

By incorporating a wide range of Indian languages, our approach enables users to engage with their audience, regardless of linguistic and cultural differences. Whether you need to adapt a message, make it more accessible, or simply communicate more effectively, our LLM-based approach offers a versatile and dynamic solution to meet your text transformation needs. We are excited to present this tool as a step towards a more interconnected and linguistically inclusive world.

By offering these capabilities across multiple Indian languages, our approach bridges the gap between languages and cultures, making it easier to engage with diverse audiences. Whether it's adapting content for different regions, enhancing accessibility, condensing information, or personalizing text, our LLM-based approach empowers users to communicate more effectively in the linguistic diversity of the Indian subcontinent.

Moreover, this technology has the potential to facilitate cross-cultural understanding and cooperation by ensuring that language barriers do not impede the exchange of ideas and information. It aligns with the ever-evolving digital landscape, where effective communication is crucial in connecting people and cultures, fostering inclusivity, and supporting diverse communication needs.

II. LITERATURE REVIEW

[6] focuses on five diverse tasks, namely, biography generation using Wikipedia infoboxes, news headline generation, sentence summarization, paraphrase generation and, questions generation. It describes the created datasets and uses them to benchmark the performance of several monolingual and multilingual baselines that leverage pre-trained sequence-to-sequence models. The results exhibit the strong performance of multilingual language-specific pre-trained models and the utility of models trained on our dataset for other related NLG tasks.

[7] presents IndicBART, a multilingual, sequence-to-sequence pre-trained model focusing on 11 Indic languages and English. IndicBART utilizes the orthographic similarity between Indic scripts to improve transfer learning between similar Indic languages. It evaluates IndicBART on two NLG tasks: Neural Machine Translation (NMT) and extreme summarization. The experiments on NMT and extreme summarization show that a model specific to related languages like IndicBART is competitive with large pre-trained models like mBART50 despite being significantly smaller.

The main objective of [5] is to convert the existing sentence into a different form by keeping the semantics or meaning the same. This will help in converting the complex sentence into a simpler one. The system mainly deals with Hindi Sentences and their different forms. It takes a sentence as input and produces another sentence without changing its semantics after applying synonyms and antonyms replacement methods. Reframing of Hindi sentences can be used to change a complex sentence in simplified form data.

[1] introduces a novel sentence-to-vector encoding framework tailored for advanced natural language processing tasks. By leveraging an encoder-decoder model trained on sentence paraphrase pairs, they ensure that the latent representation effectively encodes sentences with semantically similar information in close vector spaces. The utility of these sentence representations is further illustrated through their successful application in two distinct tasks: sentence paraphrasing and paragraph summarization. Such embeddings are particularly conducive for recurrent frameworks commonly employed in text processing. The experimental results presented underscore the efficacy of these vector representations in advanced language embedding tasks.

In [2] a novel approach was introduced to paraphrase generation utilizing generative pre-training. The crux of their methodology revolves around representing and predicting specific spans within exemplar sentences. This unique approach contrasts with traditional paraphrase generation methods that often rely on end-to-end sequence transformations. The authors argue that by focusing on distinct spans within exemplars, they can achieve more coherent and contextually accurate paraphrases. Their generative pre-training model is meticulously evaluated against standard benchmarks, and the results underscore the effectiveness of span-based representations in improving paraphrase quality. The study not only presents a compelling methodological advancement in the domain of paraphrase generation but also paves the way for future research to leverage span.

[3] introduced a paraphrase generator tailored for the Dravidian language, Kannada, utilizing a dictionary lookup approach. Central to their method is the integration of a morphological analyzer, dictionary lookup, and morphological generator to produce paraphrases for provided sentences. The morphological analyzer deciphers the internal word structures, providing both syntactic and morphological properties. The dictionary lookup, on the other hand, enables synonym substitution, ensuring the synonym occupies the same position in the paraphrased sentence as the original word did in the source sentence. Following this, the morphological generator employs suffix tables to generate inflections of the root word. The proposed tool's effectiveness is assessed across various news domains, offering a unique contribution to NLP applications for the Kannada language.

[4] applied sequential recurrent neural networks to a fairly high-level cognitive task, i.e. paraphrasing script-based stories. Using hierarchically organized modular sub-networks, which are trained separately and in parallel, the complexity of the task is reduced by effectively dividing it into subgoals. The system uses sequential natural language input and output and develops its own 1/0 representations for the words. The representations are stored in an external global lexicon, and they are adjusted in the course of training by all four subnetworks simultaneously, according to the FGREP method. By concatenating a unique identification with the resulting representation, an arbitrary number of instances of the same word type can be created and used in the stories. The system is able to produce a fully expanded paraphrase of the story from only a few sentences, i.e. the unmentioned events are inferred. The word instances are correctly bound to their roles, and simple plausible inferences of the variable content of the story are made in the process.

[12] introduces MUSS, a Multilingual Unsupervised Sentence Simplification system that does not require labeled simplification data. MUSS uses a novel approach to sentence simplification that trains strong models using sentence-level paraphrase data instead of proper simplification data. These models leverage unsupervised pretraining and controllable generation mechanisms to flexibly adjust attributes such as length and lexical complexity at inference time. It further presents a method to mine such paraphrased data in any language from Common Crawl using semantic sentence embeddings, thus removing the need for labeled data. We evaluate our approach on English, French, and Spanish simplification benchmarks and closely match or outperform the previous best-supervised results, despite not using any labeled simplification data. It pushes the state of the art further by incorporating labeled simplification data. [12]

[13] uses the adversarial paradigm and introduces a new adversarial method of dataset creation for paraphrase identification: the Adversarial Paraphrasing Task (APT), which asks participants to generate semantically equivalent (in the sense of mutually implicative) but lexically and syntactically disparate paraphrases. These sentence pairs can then be used both to test paraphrase identification models (which get barely random accuracy) and then improve their performance. To accelerate dataset generation, we explore the automation of APT using T5 and show that the resulting dataset also improves accuracy.

[14] presented an approach for generating sentence-level paraphrases, a task not addressed previously. This method learns structurally similar patterns of expression from data and identifies paraphrasing pairs among them flexible pattern-matching procedure allows us to paraphrase an unseen sentence by matching it to one of the induced patterns. This approach generates both lexical and structural paraphrases.

[8] has proposed an approach for abstractive text summarization using a generative adversarial network to perform multilingual text summarization. The results are at par with existing deep learning frameworks for summarization. Improvements can be made in terms of hyperparameter tuning and the size of the dataset. The input used to train the network needs to be larger to achieve better results. The current work uses significantly less data to train the network than the other text generation models and hence, acquiring a large dataset to improve the model is one of the future work.

[9] presented a progressive approach to train a DRL-based unsupervised paraphrasing model. The method provides a warm start to the DRL-based model with a pre-trained VAE (i.e., trained on non-parallel corpus). Then, the model progressively transitions from VAE's output to acting according to its policy. It also proposes a reward function that incorporates all the attributes of a good paraphrase and does not require parallel sentences.

Recent strides in Natural Language Processing (NLP) have led to the creation of the IndicNLPSuite. [10] includes vast corpora, tailored FastText word embeddings, compact ALBERT language models, and an IndicGLUE benchmark for NLU tasks. Derived from news crawls, the corpora cover 11 major Indian languages. FastText word embeddings enhance NLP capabilities, and ALBERT language models ensure efficiency. The IndicGLUE benchmark comprises diverse NLU evaluation datasets. Available under a Creative Commons license, these resources are poised to accelerate Indic NLP research and diversify language understanding, benefiting a large population.

The surge in digital communication has led to increased paraphrasing. Arabic, with its linguistic complexity, presents challenges. [11] proposes an Arabic paraphrase identification method using Siamese recurrent neural networks, global word vectors, and cosine similarity. It addresses the lack of Arabic paraphrase datasets and shows effectiveness through SemEval validation, contributing to NLP challenges. This research provides essential insights into paraphrase detection, highlighting its significance in Arabic linguistic contexts and NLP applications.

III. METHODOLOGY

In this research, we present a comprehensive methodology for the development of a Large Language Model (LLM) based approach aimed at rephrasing, simplifying, shortening, and stylizing text in multiple Indian languages, including Hindi, Marathi, Telugu, Tamil, Punjabi, and Odia. The proposed methodology covers data collection, model selection, training, evaluation, validation, iterative improvement, deployment, and scalability considerations.

Data Collection and Preprocessing: Data collection is a critical first step in our methodology. We gather a diverse corpus of text from a variety of sources in the target languages. This corpus is meticulously preprocessed to remove noise, ensuring that the data is consistent and free from irregularities that might affect the performance of the language model.

Model Selection: Choosing the right Large Language Model (LLM) is pivotal to the success of this project. We evaluate various LLMs based on their performance, language support, and scalability. The selected model will serve as the foundation for the subsequent steps in our approach.
Training and Fine-Tuning: The training process involves exposing the chosen LLM to the preprocessed dataset. It learns to rephrase, simplify, shorten, and stylize text through iterative exposure to this data. Fine-tuning further customizes the model, ensuring its

IV. LITERATURE REVIEW TABLE

Paper	Methodologies	Dataset used	Advantages	Limitations	Key findings/results
[1]	RNN encoder with LSTM	Visual Caption Datasets,The SICK dataset ,TACoS Multi-Level Corpus	Applicable in a variety of tasks	Styles and topics of the sentences in this dataset are limited	Reversing the encoder sentences helped the model learn long dependencies over long sentences.
[2]	GPT-2 ,Bernoulli distribution	QQP-Pos,ParaNMT-small	Allows the model to provide various paraphrased sentences in testing, corresponding to different secondorder-masking levels. Outperforms competitive baselines in semantic score	All of the results are based on a greedy technique.	The model developed a template masking technique, named first-order masking, to masked out irrelevant words in exemplars utilizing POS taggers. So that,the paraphrasing task is changed to predicting spans in masked templates.
[3]	Dictionary Lookup and Morphological Analyzer.	Kannada ratnakoosha	Developed tool can handle up to 4300 words .	It can handle mostly nouns.It can be improved by using concept noun phrase and verb phrase.	The methods used for paraphrasing includes using Dictionary lookup method that is synonym representation and Morphological analyzer gives the internal structure of the words.Then, the inflection of root word is generated through suffix tables.
[4]	Sequential RNN	Short stories.	The system is able to infer events which are left unmentioned in the story, and make simple plausible inferences of the variable content of the story.	A mechanism should be developed for representing multiple scripts and their interactions (e.g. a phone script or robbery script occurring within a restaurant script).	Paraphrasing script-based stories. Using four hierarchically organized modular subnetworks, which are trained separately and in parallel, the complexity of the task is reduced by effectively dividing it into subgoals.
[5]	Natural language Processing	Input text by user	This work will generate new sentences based on specified rules and after synonyms and antonyms replacements.	This work can be extended to make a decision support system that will work is a similar manner like humans and can respond just like humans by understanding the different forms of sentences given to it as an input.	Uses an algorithm which contains synonym and antonym replacement and applying reframing rules.
[6]	Pre- trained IndicBART	Annotate 250 examples in	In general, multilingual models outperform	Data creation relies on resources like parallel	Fine-tuning IndicBART gives substantially better

	and mt5 models	WikiBio and Paraphrasing, and 100 examples in Headline Generation	their monolingual counterparts.	corpora, monolingual corpora and Wikipedia of reasonable sizes. The approach may not apply to languages where such resources are scarce.	results than fine-tuning mT5.
[7]	Pre-trained IndicBART and mBART50 models	WAT 2021 MultiIndicMT test-set and the FLORES101 devtest ,the multilingual XL-Sum dataset	IndicBART supports 11 Indian languages and English, and utilizes the orthographic similarity of Indic scripts to enable better cross-lingual transfer.	Should plan to focus on training models on longer text chunks (documents) and larger text corpora, incorporating advances in multilingual pre-training, cross-lingual transfer and cross-lingual tasks for Indian languages.	IndicBART (IB) and mBART50 are competitive with each other where the former performs slightly better for Marathi, Punjabi, Tamil and Telugu.
[12]	Sentence Simplification and keeping the meaning of the sentence same	Newsela. Wikilarge	MUSS outperforms our strongest baseline by +8.25 SARI for French, while matching the pivot baseline performance for Spanish.	Further improvements could be achieved by using monolingual BART models trained for French or Spanish, possibly outperforming the pivot baseline.	Languages like English, French and Spanish that use labeled simplification data,
[13]	Adversarial Paraphrasing Task using Semantic Textual Similarity	Human-generated APT dataset , T5 base generated dataset.	The paper showed that RoBERTabase trained on TwitterPPDB performed poorly on APT benchmarks, but this performance was increased significantly when further trained on either our human- or machine-generated datasets.	future research into improving the models performance can be very valuable.	Using adversarial paradigm to recognize different sentences that have same meaning
[14]	Multiple Sequence Alignment	Articles produced between September 2000 and August 2002 by the Agence France-Press(AFP) and Reuters news agencies.	Produced lexical and structural paraphrases	Semantic and Reversal Paraphrases not implemented	Utilize lattices to identify structural similarities, confirming paraphrases, and employ mate lattices for generating output sentences from input.
[8]	Generative Adversarial Networks (GAN)	Multiling 2015 Dataset	Results are at par with existing deep learning frameworks for summarization.	Relatively smaller dataset used	Proposed an approach for abstractive text summarization using a generative adversarial network to perform multilingual text summarization.
[9]	Natural language generation,Unsupervised	Quora, WikiAnswers , MSCOCO, and Twit-	Reward function that incorporates all the attributes of a good paraphrase	Uses highly labelled data and cannot be trained by scarce dataset.	Method achieves up to 90% and 34% performance gains for the BLEU and

	learning,Deep Reinforcement Learning,Discrete Space Search	ter datasets			the i-BLEU metrics compared to state-of-the-art unsupervised methods, respectively.
[10]	IndicCorp: Indian Language Corpora, IndicGLUE: Multilingual NLU Benchmark, IndicBERT	WikiAnn NER dataset	Resources derived from this dataset outperform other pre-trained embeddings on many NLP tasks.	Multiple-choice QA and Cross-Lingual Sentence Retrieval prove to be the more challenging tasks.	Presented the IndicNLPSuite, a collection of large-scale, general-domain,sentence-level corpora of 8.9 billion words across 11 Indian languages, along with pre-trained models (IndicFT,IndicBERT) and NLU benchmarks (IndicGLUE).Accuracy: en to Indic:40.29%, Indic to en: 48%
[11]	CNN,LSTM,Bi-LSTM recurrent neural network	King Saud University Corpus of Classical Arabic KSUCCA	Combinations to improve the task of Arabic paraphrase detection in terms of statistical regularities in the context of sentences	Accuracy can be improved	F1 scores: 88.76% with Bi-LSTM

effectiveness in handling text transformation tasks in our target languages.

Evaluation Metrics: To objectively measure the model's performance, we employ specific evaluation metrics designed to assess its effectiveness in rephrasing, simplifying, shortening, and stylizing text while preserving the context and accuracy of the original content.

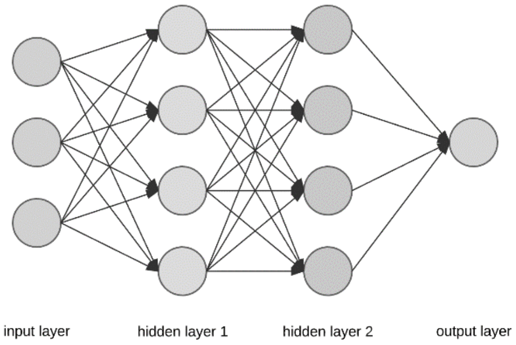
Validation Process: In addition to quantitative metrics, we implement a qualitative validation process. Model outputs undergo human evaluation to ensure the quality of the transformed text. This ensures that the text is not only modified but also remains contextually accurate and maintains its intended meaning.

Iterative Improvement: Our methodology emphasizes an iterative approach. Feedback and results from evaluations guide the continuous refinement of the model. Parameters are adjusted, and the dataset is expanded to enhance performance, ensuring the model's ability to handle diverse text transformation tasks.

Deployment Strategy: Upon achieving a satisfactory level of performance, the LLM-based solution is deployed. Particular emphasis is placed on user-friendliness and efficiency, making it suitable for various applications and users with different levels of technical expertise.

V. ARCHITECTURE

Neural networks are a fundamental component of artificial intelligence (AI) and machine learning. They are a computational model inspired by the structure and function of the human brain. Neural networks consist of interconnected nodes, often referred to as "neurons" or "artificial neurons." These artificial neurons work together to process and analyze complex data, recognize patterns, and make predictions or decisions.



The neural networks form the base for the transformer model which we are going to use for achieving paraphrasing in Indian languages. The Transformer model is a revolutionary architecture in the field of natural language processing (NLP) and deep learning. It was introduced in a 2017 paper titled "Attention Is All You

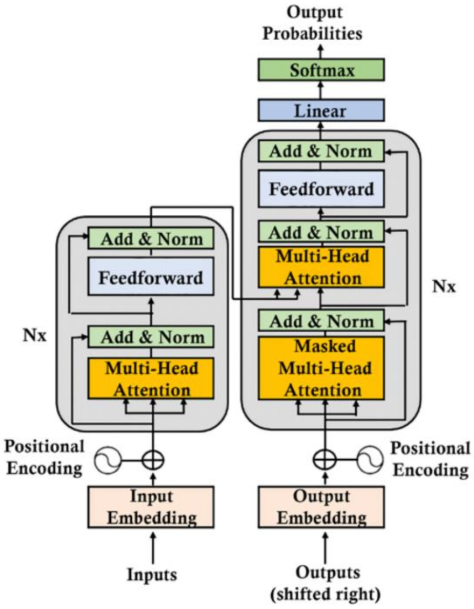
Need" by Vaswani et al., and it has since become the foundation for many state-of-the-art NLP models. The Transformer architecture offers several key innovations:

Self-Attention Mechanism: The core of the Transformer is the self-attention mechanism. Instead of processing words in a sequence, as in traditional RNNs or LSTMs, the Transformer model considers all words in a sentence simultaneously. It assigns different attention scores to different words, allowing the model to weigh the importance of each word in the context of the entire sentence.

Positional Encoding: Since the Transformer doesn't have built-in notions of word order (like an RNN or LSTM), it relies on positional encodings to consider the position of words in the input sequence. These positional encodings are added to the word embeddings, ensuring the model can differentiate between words based on their positions

Layer Normalization and Feed-Forward Networks: After the self-attention layers, Transformers include feed-forward neural networks and layer normalization. These components help in further capturing patterns and relationships within the data.

Encoder-Decoder Architecture: While Transformers were initially designed for sequence-to-sequence tasks (e.g., machine translation), they have also been adapted for other tasks, including text classification and language modeling. In these cases, the architecture is often simplified, using only the encoder part of the Transformer.



The model that we are planning to use which is based on transformer architecture is BERT, which stands for Bidirectional Encoder Representations from Transformers, is a groundbreaking natural language processing (NLP) model introduced by Google AI in 2018. BERT revolutionized the field of NLP and achieved state-of-the-art results on a wide range of NLP tasks.

VI. CONCLUSION

The realm of text generation, while advanced in the English language, has yet to fully explore the richness and diversity of Indian languages. With languages like Hindi, Marathi, Telugu, and others, each carrying its unique syntax, grammar, and cultural nuances, the challenge becomes multi-dimensional. Leveraging Large Language Models (LLMs) to paraphrase and adapt content in these languages not only offers a technological solution but also opens a window to the heart of India's linguistic heritage.

As we navigate the forefront of this technological advancement, the need for collaboration is more crucial than ever. Engaging with native linguists, cultural experts, and the broader community will be pivotal in refining and ensuring the authenticity of these models. Such an endeavor promises not only to cater to the immediate needs of writers and translators but also to foster a broader sense of linguistic inclusivity and representation. The journey, while challenging, heralds a new era of embracing and celebrating the linguistic tapestry of India.

ACKNOWLEDGEMENTS

We are deeply grateful to Prof. Dr. Arati Despande ma’am, our respected Head of the Department (HOD), and our department for the support, guidance, and help provided during the research conducted for the survey paper. Their expertise, encouragement, and mentorship have been essential in shaping our journey. Their unwavering belief in our capabilities and commitment to our academic growth has been a source of inspiration. We truly appreciate their dedication to our success. Thank you for being pillars of knowledge and support, and for your pivotal roles in our project's progress.

REFERENCES

[1]. C. Zhang et al., "Semantic sentence embeddings for paraphrasing and text summarization," 2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP), Montreal, QC, Canada, 2017, pp. 705-709, doi: 10.1109/GlobalSIP.2017.8309051..

[2]. T. -C. Bui, V. -D. Le, H. -T. To and S. K. Cha, "Generative Pre-training for Paraphrase Generation by Representing and Predicting Spans in Exemplars," 2021 IEEE International Conference on Big Data and Smart Computing (BigComp), Jeju Island, Korea (South), 2021, pp. 83-90, doi: 10.1109/BigComp51126.2021.00025.

[3]. A. Gadag and B. M. Sagar, "Paraphrase generator using dictionary lookup for Kannada language," 2016 2nd International Conference on Next Generation Computing Technologies (NGCT), Dehradun, India, 2016, pp. 164-168, doi: 10.1109/NGCT.2016.7877408.

[4]. Miikkulainen and Dyer, "A modular neural network architecture for sequential paraphrasing of script-based stories," International 1989 Joint Conference on Neural Networks, Washington, DC, USA, 1989, pp. 49-56 vol.2, doi: 10.1109/IJCNN.1989.118677

[5]. Sethi, Nandini & Agrawal, Prateek & Madaan, Vishu & Singh, Sanjay. (2016). A Novel Approach to Paraphrase Hindi Sentences using Natural Language Processing. Indian Journal of Science and Technology. 9. 10.17485/ijst/2016/v9i28/98374

[6]. Aman Kumar, Himani Shrotriya, Prachi Sahu, Amogh Mishra, Raj Dabre, Ratish Puduppully, Anoop Kunchukuttan, Mitesh M. Khapra, and Pratyush Kumar. 2022. IndicNLG Benchmark: Multilingual Datasets for Diverse NLG Tasks in Indic Languages. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 5363–5394, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics

[7]. Raj Dabre, Himani Shrotriya, Anoop Kunchukuttan, Ratish Puduppully, Mitesh Khapra, and Pratyush Kumar. 2022. IndicBART: A Pre-trained Model for Indic Natural Language Generation. In Findings of the Association for Computational Linguistics: ACL 2022, pages 1849–1863, Dublin, Ireland. Association for Computational Linguistics

[8]. Rupal Bhargava, Gargi Sharma, Yashvardhan Sharma,Deep Text Summarization using Generative Adversarial Networks in Indian Languages,Procedia Computer Science,Volume 167,2020

[9]. A. B. Siddique, Samet Oymak, and Vagelis Hristidis. 2020. Unsupervised Paraphrasing via Deep Reinforcement Learning. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '20). Association for Computing Machinery, New York, NY, USA, 1800–1809. https://doi.org/10.1145/3394486.3403231.

[10]. Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., 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, Online. Association for Computational Linguistics

[11]. Mahmoud, A., Zrigui, M. BLSTM-API: Bi-LSTM Recurrent Neural Network-Based Approach for Arabic Paraphrase Identification. Arab J Sci Eng 46, 4163–4174 (2021). https://doi.org/10.1007/s13369-020-05320-w

[12]. Louis Martin, Angela Fan, Éric de la Clergerie, Antoine Bordes, and Benoît Sagot. 2022. MUSS: Multilingual Unsupervised Sentence Simplification by Mining Paraphrases. In Proceedings of the Thirteenth Language Resources and Evaluation Conference, pages 1651–1664, Marseille, France. European Language Resources Association..

[13]. Nighojkar, Animesh & Licato, John. (2021). Improving Paraphrase Detection with the Adversarial Paraphrasing Task. 7106-7116. 10.18653/v1/2021.acl-long.552.

[14]. Regina Barzilay and Lillian Lee. 2003. Learning to Paraphrase: An Unsupervised Approach Using Multiple-Sequence Alignment. In Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, pages 16–23.