Neural Machine Translation for Malayalam Paraphrase Generation

Christeena Varghese
Technical University of Applied Sciences
Würzburg-Schweinfurt
Würzburg, Germany

Sergey Koshelev
Institute of Linguistics, RAS
Moscow, Russia
s.koshelev@iling-ran.ru

Ivan P. Yamshchikov
CAIRO,
Technical University of Applied Sciences
Würzburg-Schweinfurt
Würzburg, Germany
ivan.yamshchikov@thws.de

Abstract

This study explores four methods of generating paraphrases in Malayalam, utilizing resources available for English paraphrasing and pre-trained Neural Machine Translation (NMT) models. We evaluate the resulting paraphrases using both automated metrics, such as BLEU, METEOR, and cosine similarity, as well as human annotation. Our findings suggest that automated evaluation measures may not be fully appropriate for Malayalam, as they do not consistently align with human judgment. This discrepancy underscores the need for more nuanced paraphrase evaluation approaches especially for highly agglutinative languages.

1 Introduction

Paraphrase generation is the task of rephrasing a given text while retaining its original meaning. Paraphrase generation has attracted considerable attention in natural language processing (NLP) and computational linguistics. Alternatively, paraphrasing can be defined as rewriting a sentence in a different form without losing its semantic information. Thus automated paraphrasing is an essential component of any successful NLP system. For example, paraphrasing is essential for an NLP system to pass the Turing test.

There are several notable ideas to generate paraphrases that are relevant in the context of this paper. First, the idea to use machine translation-inspired solutions for paraphrasing dates back to Quirk et al. (2004) who developed monolingual machine translation for paraphrase generation. Second, the idea of context-aware statistical paraphrase, that, to our knowledge, was first introduced in Zhao et al. (2009).

Recently, Li et al. (2017) presented an approach that integrated deep reinforcement learning to obtain automated paraphrases. Gupta et al. (2018) showed how deep generative networks could be used for paraphrase generation. Whereas Egonmwan and Chali (2019) used transformers for paraphrase generation. Zhou and Bhat (2021) provide a more detailed view of paraphrase generation.

In the context of linguistic diversity, the focus on paraphrasing extends beyond widely spoken languages to also include regional languages with rich linguistic nuances. Salloum and Habash (2011) addressed the challenge of dialectal to standard Arabic paraphrasing to enhance Arabic-English statistical machine translation. Their work signifies a critical effort to improve translation accuracy and fluency across different Arabic linguistic variants. Additionally, Mizukami et al. (2014) made a substantial contribution by creating a free, general-domain paraphrase database for the Japanese language. Furthermore,Gao et al. (2018) explores the enhancement of English-to-Chinese neural machine translation through paraphrase-based data augmentation.

This work addresses paraphrase generation in Malayalam. Malayalam is a Dravidian language spoken predominantly in Kerala. It is also spoken in Mahe and Lakshadweep of India, altogether resulting in a population of about 34 million. Malayalam is known for its complicated grammatical structures, complex verb conjugations, and extensive vocabulary.

There are several research projects addressing paraphrases identification Malayalam language and recognizing sentence similarities, see (Mathew and Idicula, 2013b) and (Gokul et al., 2017). Recently, (K. Nambiar et al., 2023) provided a Malayalam model for machine translation, text summarization, and question-answering.

As an extension of the above-mentioned stud-

Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 10–15

ies, this paper aims to address the complex area of Malayalam paraphrase generation. Our investigation focuses on developing a specialized dataset tailored to Malayalam paraphrases, leveraging insights from established paraphrase generation models. The main motivation behind this research is to address the lack of resources for non-English languages and to improve the capabilities of NLP systems in the context of languages with special linguistic features.

2 Related Works

A seminal work by Dolan and Brockett (2005) emphasizes the importance of developing effective models for paraphrase generation, considering the varying syntactic and semantic expressions across different languages. These challenges become more pronounced in highly agglutinative languages, where words can be formed by stringing together multiple morphemes, adding an additional layer of complexity to the generation process. Extending paraphrase generation to Malayalam, a language with a complex linguistic structure, demands special attention.

Scientific papers on multilingual NLP, such as the work by Huang et al. (2020), emphasize the need for language-specific adaptations in paraphrase generation models. The authors discuss the impact of linguistic diversity on the performance of NLP models, underscoring the importance of addressing language-specific challenges.

Mathew and Idicula (2013a) propose four similarity measures to predict the similarity between two sentences in Malayalam. Those are cosine similarity, Jaccard similarity, overlap coefficient, and containment measure.

A shared Task on Detecting Paraphrases in Indian Languages, namely, Hindi, Tamil, Malayalam, and Punjabi are proposed by Anand Kumar et al. (2018). It consisted of two subtasks: Subtask 1 is to determine whether a sentence pair is a paraphrase or not, and Subtask 2 is to determine whether a sentence pair is a semi-paraphrase or a paraphrase or not. Different members use different features such as stop words, lemmatization, POS tagging, synonyms, overlap, cosine similarity, Jaccard similarity, etc. Due to the complexity of sentences, the F1 score and accuracy of Task 1 are comparatively high compared to the accuracy of Task 2. They concluded that the agglutinative character of Malayalam and Tamil makes paraphrasing more challenging.

3 Data

This paper uses the GYAFC dataset Rao and Tetreault (2018) in English as a start for the paraphrasing pipeline. This dataset consists of informal and formal sentence pairs which are built using the Yahoo Answers L6 corpus. The sentences in this dataset are obtained from various domains including Entertainment, Music, Family, Relationships, etc. Around 1000 English sentence pairs are available in this dataset.

Though we explore the possibility of adopting English datasets for Malayalam paraphrasing, we also provide a sample of 800 Malayalam paraphrase pairs evaluated by crowd workers with overlap of five[1]. The details on datalabelling are provided in Section 5.

4 Methods

We try to explore four approaches that could potentially leverage the knowledge that we have for English and transfer it into Malayalam paraphrase. The first approach simply uses Google Translate on a random sample of 200 GYAFC paraphrases. We evaluate all four approaches on random 200 GYAFC sentence pairs.

The first model combines the output of Google Translate with MultiIndic Paraphrase Generation, a pre-trained model for paraphrase generation Kumar et al. (2022). A prior study by Zhou et al. (2018) served as the foundation for MultiIndic Paraphrase Generation, which extracts paraphrases from a parallel corpus. The model is developed using the Samanantar corpus Ramesh et al. (2022), which contains parallel corpora between English and all 11 Indic languages. 200 pairs of English phrases from the GYAFC dataset are translated into Malayalam using Google Translate. These translated Malayalam sentences are then fed into the MultiIndic Paraphrase Generation to obtain desired Malayalam paraphrase pairs. An illustrative example pertaining to this model can be found in Figure 1.

In the second approach, we use a set of English synonym word pairs[2] to generate paraphrases in English with a simple synonym replacement heuristic approach to paraphrase. The generated paraphrases

[1]Omitted to preserve anonymity in peer reivew.

[2]https://github.com/i-samenko/Triplet-net/blob/master/data/data.csv

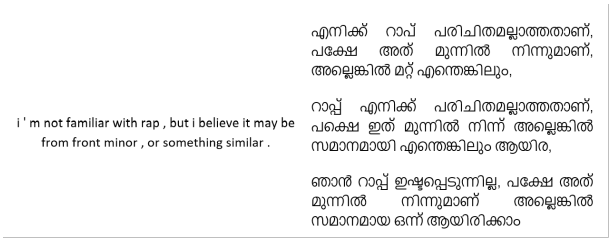

Figure 1: Result from Model 1

are then translated into Malayalam using Google Translate to obtain the Malayalam paraphrase set. Figure 2 exhibits an instance exemplifying this model, contributing to a deeper understanding.

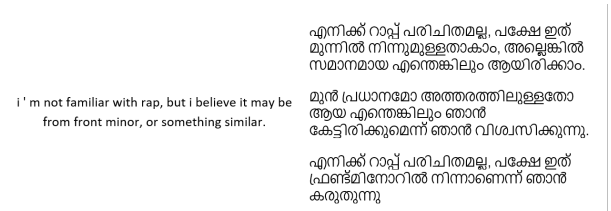

Figure 2: Result from Model 2

In the third approach, we use the bart-large-cnn model Lewis et al. (2019). Figure 3 contains an exemplar related to this model, offering additional clarity.

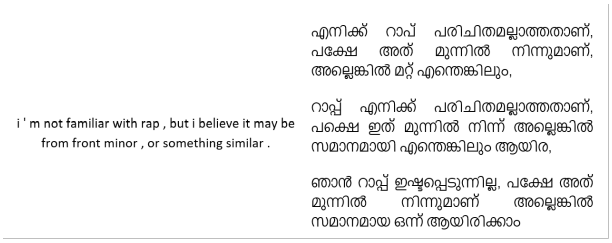

Figure 3: Result from Model 3

Finally, in the fourth model a pre-existing language translation model named, OPUS(Open Parallel Corpus) Tiedemann (2012). OPUS models are a collection of pre-trained multilingual machine translation models developed by the Helsinki NLP group. OPUS models are designed to handle translation tasks in several languages. They are trained to support translation between different language pairs, making them versatile for multilingual applications. Once again 200 pairs of sentences from the GYAFC dataset are passed to this model and Malayalam sentence pairs are generated. These translated sentences are then paraphrased by adjusting the beam-search parameters. Figure 4 includes an example associated with this model, providing supplementary clarity.

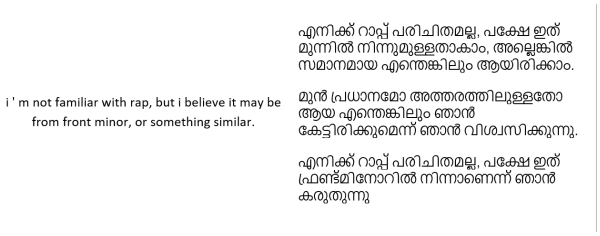

Figure 4: Result from Model 4

The num_beams parameter controls the number of beams to use in beam search. Beam search is a decoding algorithm that explores multiple possible sequences and selects the most likely ones. A larger num_beams value can increase diversity in generating phrases. Additionally, the num_return_sequences parameter determines how many different sequences to return. A higher value will result in more diverse paraphrases. Moreover, early_stopping is used to speed up the paraphrase generation process. These parameters collectively influence the diversity, quality, and speed of paraphrase generation.

Finally, we compare these paraphrase methods based on NMT with the paraphrase proposed for Malayalam in Anand Kumar et al. (2018).

5 Evaluation

Yamshchikov et al. (2021) have explored various metrics for the evaluation of paraphrases. They found BERTScore (Zhang et al., 2019) to be the most adequate metric for English paraphrases. However, there is no direct analogy of BERTScore for Malayalam and the most commonly used metrics do correlate with human judgment (Solomon et al., 2022) on par with BERTScore though not perfectly. Thus, in this work, we calculate the BLEU score (Papineni et al., 2002) and METEOR score (Lavie and Denkowski, 2009) for evaluating the phrases generated for a reference sentence. We also use cosine similarity used for paraphrase evaluation by Anand Kumar et al. (2018) to put our results in perspective, despite cosine similarity was found to have a lower correlation with the human evaluation of paraphrases. Finally, we have labelled 200 paraphrase pairs generated by each of the models with human labellers via crowd-sourcing platform. Each sentence was labelled with three or more native speakers of Malayalam. We measured a percentage of sentence pairs that were labelled as correct

[3]The self-reported evaluation metric.

Model	BLEU	METEOR	cosine similarity	human labels
MultiIndic Paraphrase (Kumar et al., 2022)	0.04	0.25	0.70	0.37
Synonym Replacement	0.05	0.28	0.60	**0.42**
BART (Lewis et al., 2019)	0.20	0.31	**0.96**	0.31
OPUS (Tiedemann, 2012)	**0.34**	**0.63**	0.83	0.23
Malayam Paraphrase (Anand Kumar et al., 2018)	-	-	0.79 [3]	-

Table 1: Average BLEU score, METEOR score, Cosine Similarity as well as the percent of paraphrases labelled as correct paraphrase by human labellers for various models.

paraphrases with high confidence. We publish the resulting human-labelled dataset of 800 sentence pairs to facilitate further research of paraphrasing in Malayalam.

Table 1 shows the results of the evaluation for 200 randomly sampled sentence pairs produced by four models that we test. It also puts these results into perspective comparing with the best result for Malayalam presented reported in Anand Kumar et al. (2018) denoted in the Table 1 as 'Malayalam Paraphrase'.

One can see that the OPUS model outperforms other models in terms of automated evaluation metrics. In the meantime, the paraphrases generated with MultiIndic Paraphrase Generation, specifically designed for Indian languages, show lower results on automated evaluation. Comparison of the proposed methods with the best Malayalam paraphrasing model described in Anand Kumar et al. (2018) also shows that on automated paraphrase evaluation metrics, direct application of machine translation methods, namely, BART or OPUS, leads to results that score higher in terms of BLEU, METEOR, and Cosine Similarity. However, this does not necessarily point at the weakness of the models but rather highlights the inadequacy of those popular evaluation metrics for Malayalam paraphrasing as well as the opportunity to leverage NMT to significantly expand the capabilities of Malayalam NLP.

Once we include human evaluation into the picture we see two crucial results. First, the most successful paraphrases, according to human judgement, as simply achieved by heuristic synonym replacement. This is not surprising. What is important is that humans also evaluate MultiIndic Paraphrase higher that BART or OPUS, despite those models higher scores on automated metrics.

6 Discussion

In this study, we check if one could use machine translation methods for paraphrasing in Malayalam. We test several methods of generating paraphrases in English, followed by their translation into Malayalam. This methodology was compared with the performance of Malayalam-specific paraphrase models.

Our findings reveal that using English for initial paraphrase generation and then translating to Malayalam can yield results that are on par with those from Malayalam-specific models. This has several important implications:

- Resource Optimization: This strategy showcases an efficient use of resources, leveraging the strengths of a high-resource language like English to benefit lower-resource languages;
- Model Versatility: The success of this approach suggests a potential shift in focus from developing language-specific models to enhancing translation-based methods;
- Expandability: such health check could be interesting for other Dravidian languages.

At the same time, one has to highlight certain limitations:

- Translation Dependence: The effectiveness of paraphrases is heavily reliant on the accuracy and nuances captured by the machine translation process;
- Evaluation Metrics Concern: A critical limitation is the potential inadequacy of automated evaluation metrics in accurately capturing the quality of paraphrases in Malayalam. This raises concerns about the reliability of any paraphrase results solely evaluated automatically without any human labels whatsoever;

- Model Reliance: The approach's success is contingent on the performance of the English paraphrase models employed.

7 Conclusion

This study evaluates how effective is the idea to apply the existing neural machine translation methods to paraphrase generation in Malayalam. The core finding of this paper is that the models specifically designed for agglutinative languages like Malayalam are showing performance on par with NMT machine translation pipelines that leverage available English resources. The study also highlights the demand for specific paraphrase evaluation metrics more suitable for Dravidian languages. Finally, we publish human-labelled dataset of paraphrases to facilitate further research on the topic.

References

M Anand Kumar, Shivkaran Singh, B Kavirajan, and KP Soman. 2018. Shared task on detecting paraphrases in indian languages (dpil): An overview. In *Text Processing: FIRE 2016 International Workshop, Kolkata, India, December 7–10, 2016, Revised Selected Papers*, pages 128–140. Springer.

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

Elozino Egonmwan and Yllias Chali. 2019. Transformer and seq2seq model for paraphrase generation. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 249–255.

Qinghong Gao, Pengjun Xie, Hua Wu, and Haifeng Wang. 2018. Improving english-to-chinese neural machine translation through paraphrase-based data augmentation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 3389–3398.

PP Gokul, BK Akhil, and Kumar KM Shiva. 2017. Sentence similarity detection in malayalam language using cosine similarity. In *2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)*, pages 221–225. IEEE.

Ankush Gupta, Arvind Agarwal, Prawaan Singh, and Piyush Rai. 2018. A deep generative framework for paraphrase generation. In *Proceedings of the aaai conference on artificial intelligence*, volume 32.

Zhen Huang, Shiyi Xu, Minghao Hu, Xinyi Wang, Jinyan Qiu, Yongquan Fu, Yuncai Zhao, Yuxing Peng, and Changjian Wang. 2020. Recent trends in deep learning based open-domain textual question answering systems. *IEEE Access*, 8:94341–94356.

Sindhya K. Nambiar, David Peter S, and Sumam Mary Idicula. 2023. Abstractive summarization of text document in malayalam language: Enhancing attention model using pos tagging feature. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(2):1–14.

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

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

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Zichao Li, Xin Jiang, Lifeng Shang, and Hang Li. 2017. Paraphrase generation with deep reinforcement learning. *arXiv preprint arXiv:1711.00279*.

Ditty Mathew and Sumam Mary Idicula. 2013a. Paraphrase identification of malayalam sentences - an experience. In *2013 Fifth International Conference on Advanced Computing (ICoAC)*, pages 376–382.

Ditty Mathew and Sumam Mary Idicula. 2013b. Paraphrase identification of malayalam sentences-an experience. In *2013 Fifth International Conference on Advanced Computing (ICoAC)*, pages 376–382. IEEE.

Masahiro Mizukami, Graham Neubig, Sakriani Sakti, Tomoki Toda, and Satoshi Nakamura. 2014. Building a free, general-domain paraphrase database for japanese. In *2014 17th Oriental Chapter of the International Committee for the Co-ordination and Standardization of Speech Databases and Assessment Techniques (COCOSDA)*, pages 1–4. IEEE.

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

Chris Quirk, Chris Brockett, and Bill Dolan. 2004. Monolingual machine translation for paraphrase generation.

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan Ak, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Divyanshu Kakwani, Navneet Kumar, et al. 2022. Samanantar: The largest publicly available parallel corpora collection for 11 indic languages. *Transactions of the Association for Computational Linguistics*, 10:145–162.

Sudha Rao and Joel Tetreault. 2018. Dear sir or madam, may I introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 129–140, New Orleans, Louisiana. Association for Computational Linguistics.

Wael Salloum and Nizar Habash. 2011. Dialectal to standard arabic paraphrasing to improve arabic-english statistical machine translation. In *Proceedings of the first workshop on algorithms and resources for modelling of dialects and language varieties*, pages 10–21.

Shaul Solomon, Adam Cohn, Hernan Rosenblum, Chezi Hershkovitz, and Ivan P Yamshchikov. 2022. Rethinking crowd sourcing for semantic similarity. In *Conference on Artificial Intelligence and Natural Language*, pages 70–81. Springer.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Lrec*, volume 2012, pages 2214–2218. Citeseer.

Ivan P Yamshchikov, Viacheslav Shibaev, Nikolay Khlebnikov, and Alexey Tikhonov. 2021. Style-transfer and paraphrase: Looking for a sensible semantic similarity metric. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14213–14220.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Shiqi Zhao, Xiang Lan, Ting Liu, and Sheng Li. 2009. Application-driven statistical paraphrase generation. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 834–842.

Jianing Zhou and Suma Bhat. 2021. Paraphrase generation: A survey of the state of the art. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 5075–5086.

Qingyu Zhou, Nan Yang, Furu Wei, Chuanqi Tan, Hangbo Bao, and Ming Zhou. 2018. Neural question generation from text: A preliminary study. In *Natural Language Processing and Chinese Computing: 6th CCF International Conference, NLPCC 2017, Dalian, China, November 8–12, 2017, Proceedings 6*, pages 662–671. Springer.