News Headline Generation in Telugu

Likhita Sunkari
IIT Bhubaneswar, India
likhitasai02@gmail.com

Abhik Jana
IIT Bhubaneswar, India
abhikjana1@gmail.com

Abstract

News headline generation has seen significant advancements in resource-rich languages like the English language, leveraging sophisticated natural language processing (NLP) techniques. However, similar progress has not been observed for low-resource Indian languages, particularly Telugu. We focus on implementing news headline generation given the news article using an abstractive summarization approach, which enables the generation of contextually rich headlines by interpreting and rephrasing content using deep learning techniques. We create a dataset of articles and headlines scraped[1] from the Telugu news website 'Sakshi'. We use pre-trained language models such as `mBART50`[2], `mT5`[3], and `IndicBART`[4] for fine-tuning using our dataset. Our findings show the effectiveness of fine-tuning pre-trained models for news headline-generation tasks in Telugu.

1 Introduction

While significant advancements have been made in news article generation for English, there is a notable lack of similar work for Indian languages, especially Telugu. Developing a model from the ground up is time-consuming and requires extensive effort, particularly due to the need for a large training dataset, which is currently unavailable for Telugu. To address this challenge, we prepare a Telugu dataset inspired by the work of B et al. (2021). Next, we fine-tune three pre-trained models, namely mT5 (Xue et al., 2021), mBART50 (Tang et al., 2020), and IndicBART (Dabre et al., 2022), using our dataset. The paper by Xue et al. (2021) introduces mT5, a multilingual adaptation of the T5 model, which is discussed in Raffel et al. (2023). mT5 is trained on the innovative multilingual dataset mC4. Tang et al. (2020) presented mBART50, a pre-trained model tailored for multilingual text generation across 50 diverse languages. Leveraging extensive pre-training on a wide array of languages, mBART50 demonstrates robust performance in various natural language processing tasks, including machine translation and text generation. Dabre et al. (2022) introduced IndicBART, a pre-trained model designed for text summarization in Telugu and 10 other Indian languages, as well as English. We achieve a BLEU score of 0.41 and a ROUGE score of 0.25 with the fine-tuned mBART50 model, which are promising for low-resource language like Telugu. In a nutshell, the dataset we create for Telugu news headline generation is the first of its kind, considering its size and diversity, and it would be immensely useful for the low-resource community.

2 Dataset Preparation

We create the dataset from the Telugu News website `Sakshi`[5] by taking the headline as the summary and the article as the source text. We compile the dataset with 13,023 Telugu news articles. We split the dataset into a training set of 10,418 items and a test set of 2,605 items for experimentation. While creating the dataset, we focus on articles from different domains, i.e., 'politics', 'entertainment', 'sports', 'business', etc. The steps for preparing the dataset are described below.

[1] Web scraping is adhered to Sakshi's terms and conditions, with robots.txt access permissions reviewed and content used as per copyright and fair use laws.

[2] https://huggingface.co/facebook/mbart-large-50

[3] https://huggingface.co/google/mt5-small

[4] https://huggingface.co/ai4bharat/IndicBART

[5] https://www.sakshi.com/getarchive

	mBART50	mT5	IndicBART
BLEU	0.41	0.33	0.3
ROUGE	0.25	0.17	0.08
Sentence Similarity	0.45	0.33	0.27

Table 1: Performances of fine-tuned language models

2.1 Web Scraping

HTTP Request: The *requests* library is used to send a POST request to the Sakshi news archive URL (`https://www.sakshi.com/getarchive`) to receive the JSON response containing news article URLs.
URL Extraction: The JSON response from the web scraping request is processed to extract and convert relative URLs to absolute URLs for further access using regular expressions.

2.2 Text Extraction and Formatting

HTML content extraction: Absolute URLs are iterated, and the `requests`[6] library is used to make an HTTP GET request to retrieve the HTML content for parsing.
Data Extraction from HTML: Relevant information, such as the page title and article text, is extracted and stored for each URL using `BeautifulSoup`[7].
Post-processing: The extracted data is saved in a .csv file, with operations to remove duplicates and unwanted characters performed during the process.

3 Experiments and Results

Using our proposed dataset, we fine-tune the mBART50, mT5, and IndicBART models with batch sizes of 2, 4, and 4, respectively, over 5, 10, and 10 epochs. We use the hugging face platform to fine-tune and test the model. In order to evaluate the performance of models, we use `BLEU-1`[8](Papineni et al., 2002), `ROUGE-1 F-Score`[9] (Lin, 2004). In addition, we also measure the semantic similarity between the gold standard headline and the generated headline. We name this `Sentence Similarity Evaluation`[10] (Lin, 2004) metric. Table 1 shows the evaluation scores for all three models. Based on these metrics, mBART50

	Telugu Text	Translated English Text
Reference Headline	త్వరలో మొబైల్ యూజర్లకు ప్రత్యేక కస్టమర్ ఐడీ	Exclusive customer ID for mobile users soon
mBART50 Generated Headline	మొబైల్ యూజర్లకు ప్రత్యేకమైన కస్టమర్ ఐడీ!	Unique customer ID for mobile users!
mT5 Generated Headline	వినియోగదారులకు ప్రత్యేక సేవలు	Special services for customers
IndicBART Generated Headline	ప్రత్యేకమైన కస్టమర్ డీని కేటాయించి	Assign a unique customer de

Figure 1: Gold standard headline (Reference) as well as model generated headlines

is the best-performing model, and IndicBART is the worst-performing model. Figure 1 represents sample headlines produced by fine-tuned models. After doing a manual inspection of generated headlines, we notice that mBART50 consistently produces grammatically correct and contextually relevant headlines. IndicBART, however, often generates grammatically incorrect and incoherent headlines despite using words from the article.

4 Conclusion

In this work, we have attempted the task of news headline generation in the Telugu language. We created a dataset and fine-tuned three state-of-the-art language models, namely mBART50, mT5 and IndicBART, using our dataset. We noticed the promising performance of the mBART50 model for the task, which would inspire the community to investigate more in this direction.

[6] https://pypi.org/project/requests/
[7] https://pypi.org/project/beautifulsoup4/
[8] https://huggingface.co/spaces/evaluate-metric/bleu
[9] https://pypi.org/project/rouge-score/
[10] https://pypi.org/project/sentence-similarity/

References

Mohan Bharath B, Aravindh Gowtham B, and Akhil M. 2021. Neural abstractive text summarizer for telugu language. *Preprint*, arXiv:2101.07120.

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*. Association for Computational Linguistics.

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.

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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the limits of transfer learning with a unified text-to-text transformer. *Preprint*, arXiv:1910.10683.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *Preprint*, arXiv:2008.00401.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.