"Why Should English Have All the Fun?" : A Multilingual Approach for Low-Resource Transcript Summarization

Mufenas Muneer
mmch@usc.edu
Prof. Mohammed Rostami

Abstract

With the rapidly increasing internet content available globally, it is necessary to summarize it, mainly when the data contains redundant, repetitive information. Transcript summarization helps save time and effort by distilling essential information in a shorter, more understandable format. It can be beneficial for educational purposes, legal processing, or market research to identify valuable information from large amounts of data. Recent works on summarization mainly focus on the English language due to its high availability of resources. This paper proposes an abstractive multilingual transcript summarization model by building a pipeline to fine-tune state-of-the-art transformers (Vaswani et al., 2017) such as mBART and mT5 targeting English and Hindi languages. Since there are many summarization datasets for the English language, in our work, we address and assess the issue of low-resource languages such as Hindi by creating a dataset for the Hindi language by translating MediaSUM using State-of-the-Art results of Machine Translation. We combine this newly generated MediaSUM for Hindi with the existing English dataset, and fine-tune transformers on this bilingual corpus, and benchmark the best-performing model.

1 Introduction

The exponential growth of digital content has led to summarization becoming a vital technique in NLP for managing vast amounts of textual data. Summarization aims to create concise summaries that contain essential information from the text.

Although transformer-based models have advanced abstractive summarization (Neto et al., 2002), the limited resources available for low-resource languages are still a significant obstacle that hinders the ability of NLP models to perform well in these languages. Despite Hindi being the fourth most widely spoken language globally, with over 600 million speakers, there is a scarcity of annotated data available, which makes it challenging to train accurate summarization models, limiting the ability of Hindi speakers to access and process large amounts of data. This work focuses on developing a Multilingual abstractive transcript summarization model that includes both Hindi and English. To address the lack of resources for Hindi, a dataset is created by translating the existing MEDIASUM (Zhu et al., 2021) dataset into Hindi using state-of-the-art machine translation results. These two datasets are merged to create a benchmark for the best-performing model.

The project aims to enhance the productivity of Hindi and English speakers by providing them with essential and relevant information from long videos without spending significant time. Furthermore, we are tackling the problem of data scarcity for low-resource languages by creating a new in-house dataset. This will enable us to develop accurate summarization models for these languages and improve their accessibility.

2 Related Work

LSA and LexRank were found to be the top performers for news and documentaries, and films, respectively, in a summarization study Aparício et al. (2016). The study also showed similar human-generated abstractive summary scores across different domains, but combining scripts and subtitles for films was not effective. The study in Alrumiah and Al-Shargabi (2021) presents an LDA-based model for summarizing educational videos using subtitles. The results demonstrate that LDA outperforms other summarization models and that non-noun word extraction from generated keywords list improves the precision rate and human evaluation of the LDA summaries.

Most of the research on abstractive text summarization focuses on the English language due to the scarcity of well-annotated corpora for other

languages. Spotify's study in Tanaka and edgart (2022) shows that a multilingual model fine-tuned on podcast summarization data can perform as well as dedicated models fine-tuned on individual languages. In Singh et al. (2016), the authors present a bilingual summarization model in Hindi and English using an unsupervised deep learning model. To improve accuracy, they use a restricted Boltzmann machine to generate summaries.

Kakwani et al. (2021) presents Samanantar, large parallel corpora containing 49.7 million sentence pairs between English and 11 Indic Languages. The author proposes a multilingual NMT model, that outperforms existing models and baselines on publicly available benchmarks.

mT5, introduced by Xue et al. (2021), is a pretrained text-to-text transformer model trained on over 101 languages that outperforms existing models on various natural language processing tasks, including text generation, machine translation, and summarization. Hasan et al. (2021a) introduced CrossSum, a new dataset comprising 1.7 million article-summary pairs in over 1500 language pairs. They fine-tuned the MT5-CrossSum model on the dataset, achieving strong performance on both ROUGE and LaSE metrics, outperforming baseline models even on linguistically distant language pairs. Liu et al. (2020) proposed mBART, which outperforms previous approaches on various machine translation tasks and enables transfer learning to new language pairs, leading to significant performance gains for low-resource NLP tasks.

3 Problem Description

The growing demand for effective content summarization tools in multilingual settings poses significant challenges, particularly when dealing with languages such as Hindi. Existing summarization approaches primarily focus on English, with limited support for other languages like Hindi. Language-specific models may not be suitable for multilingual scenarios, where a single model capable of handling multiple languages is needed.

Developing an effective multilingual transcript summarization model poses several technical challenges that must be addressed. One key challenge is language variability (Ponti et al., 2019), as languages like English and Hindi have distinct grammar rules, vocabulary, sentence structures and. Another challenge is cultural nuances, where slang, colloquialisms, and cultural references vary significantly across multiple languages. Additionally, spoken language variations pose a challenge, as spoken language often deviates from formal written language and can include informal expressions, incomplete sentences, and pauses. The model must account for these language-specific characteristics, nuances and language variations to generate culturally appropriate summaries that convey the intended meaning accurately.It also faces the challenge of limited data availability, particularly for low-resource languages like Hindi, due to the scarcity of well-annotated corpora. To address this, it is important to generate datasets for natural language processing tasks.

The fine-tuned multilingual summarization model will be designed to effectively process spoken language, capturing key information while accounting for slang, colloquialisms, and cultural references that may vary across languages, and will generate concise and culturally appropriate summaries in both English and Hindi. As an additional benefit, this project also aims on creating a high-quality Hindi dataset that can be used as a reliable resource for researchers and developers working on Hindi summarization. The results of this project can have significant applications in various domains, facilitating content processing and information retrieval in multilingual settings.

4 Methods

Several important steps were taken for fine-tuning the models. The details for dataset preprocessing and training are listed as follows:

Dataset Generation: For building our dataset, we first extracted transcripts and summaries each at random from the extensive MediaSUM dataset (Zhu et al., 2021) resulting in an approximate total entries of 60,000. To preprocess the datasets, several steps were performed. Firstly, transcripts that were either too short or too long were removed, with the boundary conditions set between 40 and 1024 tokens. Secondly, punctuation and hyperlink removal were carried out. The dataset contained numerous backslashes used to differentiate between speakers, which added noise to the dataset. Finally, indexing issues were addressed, where inconsistent summaries were repeated within the transcripts, and these were also removed.

After preprocessing, we cut down the dataset further to 50,000 entries to work with the limited computational resources we were provided. The first

25,000 entries from this dataset were translated to Hindi using state-of-the-art translation models, IndicTrans which required an average of 8 seconds to translate one transcript-summary pair. These translations were verified using two methods. The first was BLEU score (Papineni et al. (2002)), where we achieved an impressive score of 33. This is a standard baseline BLEU score for Hindi translations from English (Kandimalla et al. (2022)). The second method was manually verifying if the translations are relevant to their English counterparts, and this was satisfied as well. From this, we concluded this translated model can not only be used further for fine-tuning but also helped us achieve our novel idea of fast generation of new translated datasets. To ensure robust training, these translated transcripts and summaries were combined with the English datasets and shuffled to avoid catastrophic forgetting (French, 1999). Due to the limited availability of computational resources, we focused solely on the MediaSUM dataset. Moving forward, we plan to utilize the Spotify Podcast dataset for further fine-tuning.

Model Selection: After an extensive survey, we decided to select top-performing multilingual summarisation models based on their ROUGE scores on English and Hindi datasets. Multiple versions of the following models were used for fine-tuning:

1. mBART (Liu et al. (2020)) is a multilingual extension of the BART model, designed for machine translation tasks. It was trained on 50 languages and can perform translation between any pair of these languages. Our hypothesis was that the model could generate more complete summaries by being exposed to more of the full input text. This extension is particularly crucial for transcripts as they are significantly longer than articles.

2. mT5 (Xue et al., 2021) was trained on a massive corpus of parallel data in 101 languages, making it one of the largest publicly available multilingual machine translation models. It can translate between any pair of these 101 languages; and has shown impressive results on various benchmarks.

Model Training: Our dataset was divided in the ratio of 8:1:1 where 80 percent of the dataset was used for training, 10 percent for validation and the remaining 10 percent for testing. Three versions of the mT5 model (mT5-multilingual-XLSum (Hasan et al., 2021b), mT5-base and mT5-small) were chosen and two versions of the mBART (mBART-25 and mBART-large-cnn) were selected for our model training task. All the models included are available at HuggingFace. These models were fine-tuned to the task of MediaSUM abstractive summarization. We employed these five different models with three different hyperparameter-tuning strategies and finetuned them to transcript summarization. We use the default settings that use 1024 tokens to train these models and to avoid truncation. We trained them for three epochs with a very small learning rate of 0.001 to reduce overfitting. We set the base batch size as 2 due to the low availability of computational resources.

5 Experimental Results

The experimental results demonstrate that the created dataset is a valuable resource for researchers and developers working on Hindi summarization. The dataset comprises diverse transcripts from various media sources and has achieved an impressive BLEU score of 33 for the translated dataset. This shows that the dataset was reliable for further tasks.

Model	1st Run	2nd Run	3rd Run
mT5-base	24	23.83	23.42
mBART-25	24.64	23.34	**24.69**
mT5-small	12.62	8.796	14.06
mBART-large-cnn	23.22	22.65	21.85
mT5-multilingual-XLSum	24.01	23.83	24.42

Table 1: ROUGE-Lsum scores for five different models

Dataset	ROUGE-1	ROUGE-2	ROUGE-Lsum
English (5000 entries - test set)	45.24	41.82	39.89
Hindi (5000 entries - test set)	17.32	13.97	13.20
Overall (10000 entries - test set)	28.41	25.10	24.69

Table 2: ROUGE scores of 3rd Run of mBART-25 on datasets divided by languages

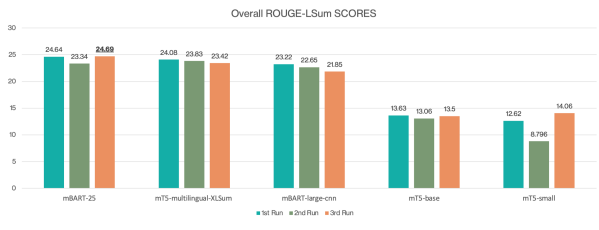

Figure 1: Rouge-LSum scores on all models

In order to establish a benchmark for comparison, we conducted experiments to determine the baseline performance of different models on the English Mediasum dataset for the summarization tasks. The baseline results indicate a ROUGE-1

score of 43 and a ROUGE-2 score of 23 using the mBART model (Goyal et al. (2021)).

For evaluation, we generated our summaries using Beam search (Ma and Zong (2020)) as it is generally preferred over greedy search (Akhmetov et al. (2021)) for abstractive summarization tasks because it allows the model to consider more diverse and accurate options for generating the output sequence. On the other hand, greedy search is a simpler and faster strategy, but it can only consider the most probable next word at each time step which may produce less diverse summaries. Therefore, we chose to use beam search with a size of 4 to generate high-quality abstractive summaries for our multilingual MediaSUM dataset.

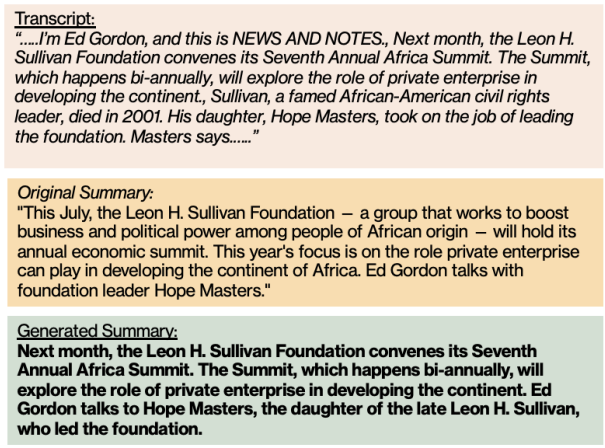

Figure 2: Sample of English abstractive summaries

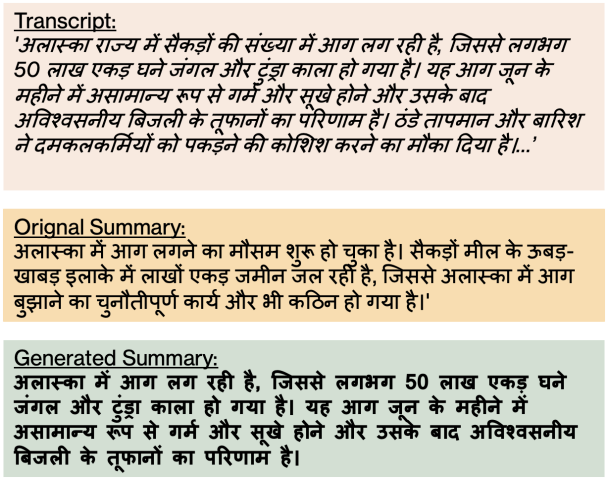

Figure 3: Sample of Hindi abstractive summaries

Figures 2 & 3 show that the summaries generated by our model are more detailed and comprehensive.

In our experiments, we evaluated the performance of our models using the ROUGE-Lsum (Lin, 2004) score, which is a metric that measures the similarity between generated summaries and human-written summaries. Our results showed that our system achieved a ROUGE-Lsum score of 24.69 on the combined English and Hindi test data, and hence we benchmark our results for the mBART-25 model. Specifically, the individual ROUGE-Lsum scores for Hindi and English were 13.20 and 39.89, respectively, when using the mBART-25 model. We decided to rely on ROUGE-Lsum scores instead of ROUGE1 because they are more suited for abstractive summaries.

However, we also found that the mT5-small model had the lowest performance among the models that we tested. This was likely due to the fact that the model was not able to capture the entire length of the input articles, which limited its ability to generate high-quality summaries. To further validate the quality of the generated summaries, we manually verified a subset of the summaries and computed similarity scores for the Hindi summaries. Our results showed that the similarity score for the Hindi summaries was 0.7, indicating that the generated summaries were highly relevant to the source content.

Overall, our results demonstrate that our system is capable of generating high-quality abstractive summaries for both English and Hindi text data and that our approach is effective.

6 Conclusion and Future Work

We successfully developed a robust multilingual transcript summarization model for English and Hindi, capable of effectively processing spoken language and generating concise and culturally appropriate summaries in both languages.

Our approach involved a specific methodology that included cleaning and extracting transcripts and summaries from MediaSum dataset, translating a portion of the transcripts and corresponding summaries into Hindi using state-of-the-art translation models, combining the Hindi and English datasets, and shuffling them to avoid catastrophic forgetting. Overall, the results demonstrate that the created dataset is a reliable resource for researchers and developers working on Hindi summarization, and our system has achieved promising results regarding summary quality. In the future, we also aim to expand our work to more low-resource languages, such as Marathi. We also intend to different techniques such as contextualized word embeddings and fine-tuning language models like GPT-3.

References

Iskander Akhmetov, Alexander Gelbukh, and Rustam Mussabayev. 2021. Greedy optimization method for extractive summarization of scientific articles. *IEEE Access*, 9:168141–168153.

Sarah Alrumiah and Amal Al-Shargabi. 2021. Educational videos subtitles' summarization using latent dirichlet allocation and length enhancement. *Computers, Materials and Continua*, 70:6205–6221.

Marta Aparício, Paulo Figueiredo, Francisco Raposo, David Martins de Matos, Ricardo Ribeiro, and Luís Marujo. 2016. Summarization of films and documentaries based on subtitles and scripts. *Pattern Recogn. Lett.*, 73(C):7–12.

Robert M French. 1999. Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, 3(4):128–135.

Tanya Goyal, Jiacheng Xu, Junyi Jessy Li, and Greg Durrett. 2021. Training dynamics for text summarization models. *arXiv preprint arXiv:2110.08370*.

Tahmid Hasan, Abhik Bhattacharjee, Wasi Uddin Ahmad, Yuan-Fang Li, Yong bin Kang, and Rifat Shahriyar. 2021a. Crosssum: Beyond english-centric cross-lingual abstractive text summarization for 1500+ language pairs. *CoRR*, abs/2112.08804.

Tahmid Hasan, Abhik Bhattacharjee, Md Saiful Islam, Kazi Samin, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021b. Xl-sum: Large-scale multilingual abstractive summarization for 44 languages.

D Kakwani, N Kumar, A Pradeep, K Deepak, V Raghavan, A Kunchukuttan, P Kumar, and MS Khapra. 2021. Samanantar: The largest publicly available parallel corpora collection for 11 indic languages. *arXiv preprint arXiv:2104.05596*.

Akshara Kandimalla, Pintu Lohar, Kumar Maji, and Andy Way. 2022. Improving english-to-indian language neural machine translation systems. *Information*, 13:245.

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.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation.

Ye Ma and Lu Zong. 2020. Attention-aware inference for neural abstractive summarization. *CoRR*, abs/2009.06891.

Joel Larocca Neto, Alex A. Freitas, and Celso A. A. Kaestner. 2002. Automatic text summarization using a machine learning approach. In *Advances in Artificial Intelligence*, pages 205–215, Berlin, Heidelberg. Springer Berlin Heidelberg.

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.

Edoardo Maria Ponti, Helen O'Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, Thierry Poibeau, Ekaterina Shutova, and Anna Korhonen. 2019. Modeling Language Variation and Universals: A Survey on Typological Linguistics for Natural Language Processing. *Computational Linguistics*, 45(3):559–601.

Shashi Pal Singh, Ajai Kumar, Abhilasha Mangal, and Shikha Singhal. 2016. Bilingual automatic text summarization using unsupervised deep learning. In *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*, pages 1195–1200.

Edgar Tanaka and edgart. 2022. Multilingual podcast summarization using longformers.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

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.

Chenguang Zhu, Yang Liu, Jie Mei, and Michael Zeng. 2021. Mediasum: A large-scale media interview dataset for dialogue summarization.