Abstractive Summarization of Nepali News Articles

1. Dataset/Corpus

For this project, I scraped different online Nepali news portals and extracted headlines and news articles across different categories. The dataset consists of around 372,000 entries. Because each entry consists of the news article, headline, and news category, the dataset could be further used for classification tasks and also for training large language models.

```
category
राजनीति            59975
देश/प्रदेश          51799
समाज              50207
अर्थ / वाणिज्य       43020
खेलकुद             42367
विश्व               37817
मनोरञ्जन            34178
विज्ञान र प्रविधि     23130
स्वास्थ्य            22392
शिक्षा               6755
Name: count, dtype: int64
```

Fig 1: News articles across different categories

	title	news	category
0	एमालेको प्रदेश प्रतिनिधिमा उदयपुरबाट सर्वसम्मत	उदयपुर : नेकपा एमाले कोशी प्रदेश कमिटीको प्रथम...	समाज
1	गौशालाबाट लुटियो ६ लाख ५० हजार रुपैयाँ	महोत्तरी : गौशाला नगरपालिका-११ भरतपुरका शम्भु ...	समाज
2	आगलागीमा ४ करोड बढीको क्षति, पीडितलाई तत्काल र...	मुगु : जिल्ला सदरमुकाम गमगढीमा शुक्रबार राति ...	समाज
3	तलेजुको दर्शन गर्न पूर्वराजा ज्ञानेन्द्र शाह भ...	भक्तपुर : पूर्वराजा ज्ञानेन्द्र शाह भक्तपुर आए...	समाज
4	वीरगन्जको खेतबाट ८ हजार ९१५ पिस लागूऔषध बरामद	वीरगन्ज : वीरगन्ज महानगरपालिका-२२ मनियारीस्थि...	समाज

Fig 2: Sample news from the dataset

2. Data Cleaning and Preprocessing:

- Removed special characters. (.,/”;\[{ …) except for (. and %) since they are used in numbers,
- Replace arabic numerals(0-9) to Nepali numerals(०-९)
- Removed English characters (Aa-Zz), emails, HTML elements, Emojis, and other unwanted non-Devanagari characters.
- Removed all the articles with news < 16 words.
- Removed all the articles with titles < 2 words or > 11 words

2.1. Preprocessing for mBART:

- Replace arabic numerals(0-9) to Nepali numerals(०-९)
- Removed English characters (Aa-Zz), emails, HTML elements, Emojis, and other unwanted non-Devanagari characters.
- Removed all the articles with news < 30 words
- Removed all the articles with titles < 4 words or > 15 words

3. Tokenizer:

For tokenization, Sentencepiece tokenizer based on the Byte Pair Encoding(BPE) algorithm was used, with a vocabulary size of 50 K.

4. Models

4.1. Attentive Seq2Seq Model

The model follows the Encoder-Decoder architecture with the following specification:

- **Encoder:**
    - Embedding Layer
    - Bi-LSTM

- **Decoder:**
    - Embedding Layer
    - LSTM
    - Bahdanau Attention
    - Concat → Context Vector and Decoder Hidden State
    - Dense
    - Softmax

4.1.1 Hyperparameters

Hyperparameters	Value
Vocab Size	50K
Encoder Sequence Length	256
Decoder Sequence Length	12
Embedding Dimension	100
Encoder Latent Dimension	128
Decoder Latent Dimension	256
Dropout	0.3
Batch Size	128
Epochs	18
L2 Regularization	0.01
Teacher Forcing Ratio	0.5
Coverage Weight	1.0

Table 1: Hyperparameters used for training the Seq2Seq Model

4.1.2 Attention Mechanism

The model uses the Bahdanau Attention, also known as Additive Attention, along with the Coverage Mechanism as discussed in the paper Get to the Point Summarization

4.1.3 Loss Function

The model uses a combination of [Cross Entropy] and Coverage Loss as discussed in the paper Get to the Point Summarization

4.2. Transformer

This model replicates the original Transformer architecture.

4.2.1 Hyperparameters

Hyperparameters	**Value**
Vocab Size	50K
Encoder Sequence Length	256
Decoder Sequence Length	12
Embedding Dimension (d_model)	256
Attention Heads (h)	8
Feed Forward Dim (d_ff)	2,048
Dropout	0.2
Encoder Layers	6
Decoder Layers	6
Batch Size	128
Learning Rate	1e-4
Epochs	20
L2 Regularization	0.01
Label Smoothing	0.1

Table 2: Hyperparameters used for training the Transformer Model

4.1.2 Loss Function

The model uses the Cross Entropy Loss function.

4.3. mBART (Multilingual Bi-Directional AutoRegressive Transformer)

Here, I finetuned the multilingual BART for summarization using QLoRA.

4.3.1 Hyperparameters

Hyperparameters	Value
Encoder Sequence Length	512
Decoder Max Length	26
Epochs	2
L2 Regularization	0.01
Batch Size	16
Learning Rate	2e-5
Model Quantization	4 bits
Lora Rank	16
Lora Scale Factor	32
Lora Dropout	0.1
Warmup Steps	1,000

Table 3: Hyperparameters used for finetuning mBART

5. Training Methodology

- **Attentive Seq2Seq**: 50% Teacher Forcing (50% of the time, ground truth label was provided)
- **Transformer**: Full Teacher Forcing

6. Inference

- Greedy Decoding
- Normalized Beam Search Decoding

7. Evaluation Metric

- BLEU
- Rouge

8. Results

Because the inference takes a very long time, Kaggle kernel could not handle it, so I have sampled a random 1000 news headline pairs and computed the metrics for that sample in the case of Attentive Seq2Seq and Transformer.
Greedy Search performed better on Attentive Seq2Seq, while Beam Search outperformed Greedy Search in Transformer

The mBART model has outperformed all the existing models for the abstractive summarization(headline generation) task.
Dhakal and Baral (2024)[8] achieved their scores with the same mBART QLoRA. However, the dataset I have used is 4 times larger than theirs, so this performance improvement is expected.

Model	BLEU	Rouge-1	Rouge-2	Rouge-L
Attentive Seq2Seq	3.17	19.15	4.9	18.68
Transformer	5.37	23.83	8.35	23.25
mBART+QLoRA	17.79	**40.5**	**22.6**	**39.05**
Dhakal and Baral (2024)[8]	-	35.9	19.99	34.88
Paudel (2022)[9]	-	15.74	3.29	15.21
Mishra et al. (2020)[10]	**22.1**	-	-	-
Thapa et al. (2024)[11]	-	20.42	15.89	17.76

Table 4: BLEU and Rouge comparision

References

[1] Kudo, T., & Richardson, J. (2018, August 19). *SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing*. arXiv.org. https://arxiv.org/abs/1808.06226

[2] Sennrich, R., Haddow, B., & Birch, A. (2015, August 31). *Neural Machine Translation of Rare Words with Subword Units*. arXiv.org. https://arxiv.org/abs/1508.07909

[3] Sutskever, I., Vinyals, O., & Le, Q., V. (2014, September 10). *Sequence to Sequence Learning with Neural Networks*. arXiv.org. https://arxiv.org/abs/1409.3215

[4] Bahdanau, D., Cho, K., & Bengio, Y. (2014, September 1). *Neural machine translation by jointly learning to align and translate*. arXiv.org. https://arxiv.org/abs/1409.0473

[5] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017, June 12). *Attention is all you need*. arXiv.org. https://arxiv.org/abs/1706.03762

[6] Papineni, K., Roukos, S., Ward, T., & Zhu, W. (2001). BLEU. *ACL Anthology*, 311. https://doi.org/10.3115/1073083.1073135

[7] Lin, C. (2004, July 1). *ROUGE: a package for automatic evaluation of summaries*. ACL Anthology. https://aclanthology.org/W04-1013/

[8] Dhakal, P., & Baral, D. S. (2024, September 29). *Abstractive Summarization of Low resourced Nepali language using Multilingual Transformers*. arXiv.org. https://arxiv.org/abs/2409.19566

[9] Paudel, N. *Attention based Recurrent Neural Network for Nepali Text Summarization*. Journal of Insitute of Science and Technology. https://www.academia.edu/97291141/Attention_based_Recurrent_Neural_Network_for_Nepali_Text_Summarization

[10] Mishra, K. R., Rathi, J., & Banjara, J. (n.d.). *Encoder Decoder based Nepali News Headline Generation*. https://www.ijcaonline.org/archives/volume175/number20/31565-2020920735/

[11] Thapa, P., Nyachhyon, J., Sharma, M., & Bal, B. K. (2024, November 24). *Development of Pre-Trained Transformer-based Models for the Nepali language*. arXiv.org. https://arxiv.org/abs/2411.15734

[12] Wu, Y., Schuster, M., Chen, Z., Le, Q., V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, Ł., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., . . . Dean, J. (2016, September 26). *Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation*. arXiv.org. https://arxiv.org/abs/1609.08144

[13] See, A., Liu, P. J., & Manning, C. D. (2017, April 14). *Get To The Point: Summarization*

with Pointer-Generator Networks. arXiv.org. https://arxiv.org/abs/1704.04368

[14] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2019, October 29). *BART: Denoising Sequence-to-Sequence Pre-training for natural language generation, Translation, and Comprehension*. arXiv.org. https://arxiv.org/abs/1910.13461

[15] Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., & Zettlemoyer, L. (2020, January 22). *Multilingual denoising pre-training for neural machine translation*. arXiv.org. https://arxiv.org/abs/2001.08210

[16] Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023, May 23). *QLORA: Efficient Finetuning of Quantized LLMS*. arXiv.org. https://arxiv.org/abs/2305.14314