TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI
HANOI UNIVERSITY OF SCIENCE AND TECHNOLOGY

ONE LOVE. ONE FUTURE.

TRƯỜNG ĐẠI HỌC
BÁCH KHOA HÀ NỘI
HANOI UNIVERSITY
OF SCIENCE AND TECHNOLOGY

Subject: Project II

Abstractive Text Summarization

Student: Vu Lam Anh - 20214876
Supervisor: Associate Prof Nguyen Thi Kim Anh
Class Code: 738755

ONE LOVE. ONE FUTURE.

Outline

1. Introduction
2. Dataset
3. Approach
4. Evaluation
5. Experiment and Result
6. Future Work
7. References

Outline

1. Introduction
2. Dataset
3. Approach
4. Evaluation
5. Experiment and Result
6. Future Work
7. References

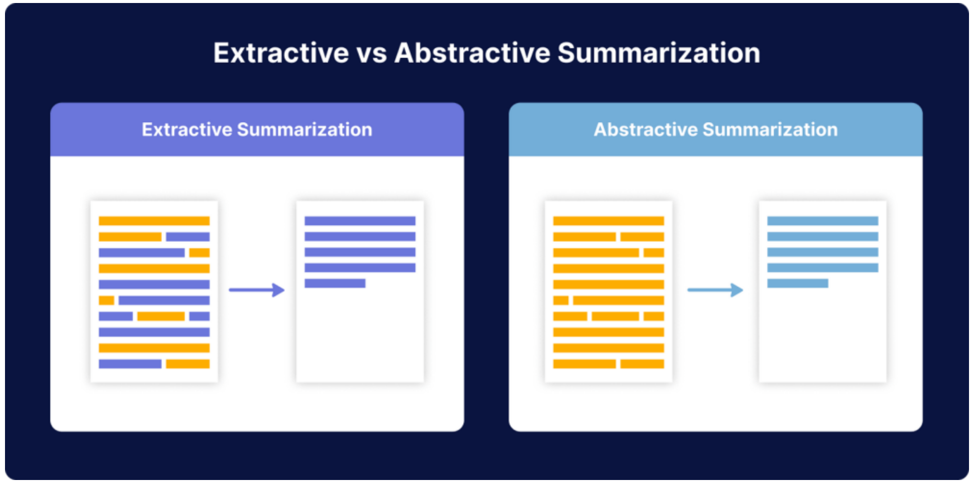
Introduction

- Text summarization is one of the main tasks in Natural Language Processing and has been applied in many areas.

- **Objectives:** Producing the shorter version of a long text or document while preserving the key information and meaning from the original document.

- **Types:**
  - Extractive summarization
  - Abstractive summarization

- In this report, we focus on the abstractive summarization



Extractive and Abstrace Summariation (Source: https://www.abstractivehealth.com/extractive-vs-abstractive-summarization-in-healthcare)

Outline

1. Introduction
2. Dataset
3. Approach
4. Evaluation
5. Experiment and Result
6. Future Work
7. References

Dataset

- In this report, we use the **NEWS SUMMARY** from the Kaggle platform

- It scraped news article and their summary from Hindu, Indian Times and Guardian from February to August, 2017

- This dataset consists of 4515 examples, each contains 2 features:
  - **Complete Article**: contains the whole text from original article
  - **Short Text**: contains text summaries the information from that article

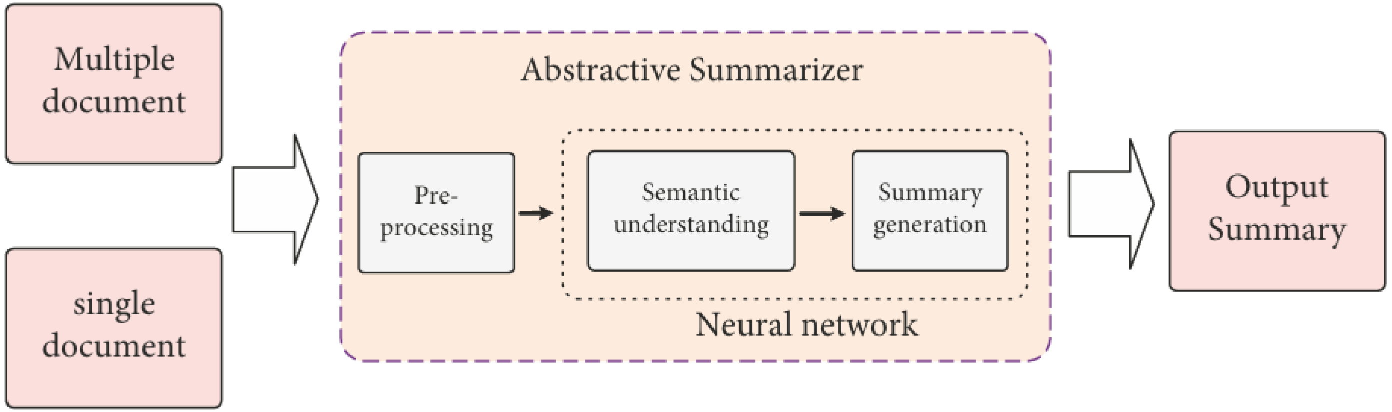
Outline

1. Introduction
2. Dataset
3. Approach
4. Evaluation
5. Experiment and Result
6. Future Work
7. References

Approach

❑ General approach

- In general, the abstractive summarization task can be divided into two intersection sub-tasks :
  - **NLU (Naltural Languge Understanding)**
  - **NLG (Naltural Language Generation)**



General workflow of abstractive summerization model. Source:[A Comprehensive Survey of Abstractive Text Summarization Based on Deep Learning (hindawi.com)]

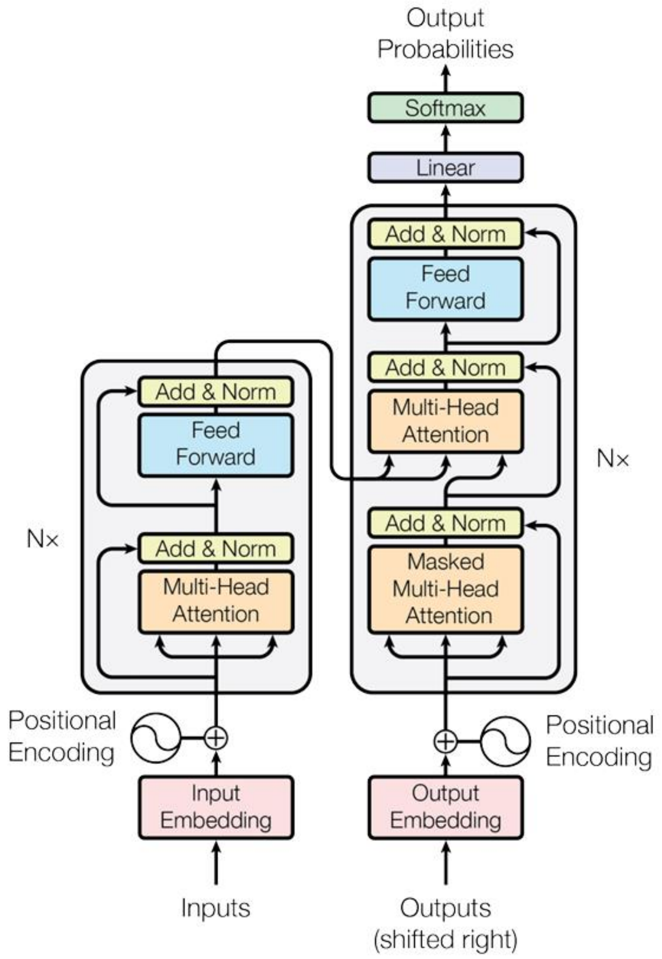
Approach

❑ BART Model

- Bart model is the model which pre-trains a model combining Bidirectional and Auto-Regressive Transformers

- It uses the standard sequence-to-sequence Transformer architecture (Vaswani et al., 2017) except, following GPT model, they modify ReLU activation functions to GeLUs.

- **Task** : It pre-trained the task of reconstructing the denoising text go back to the original text.

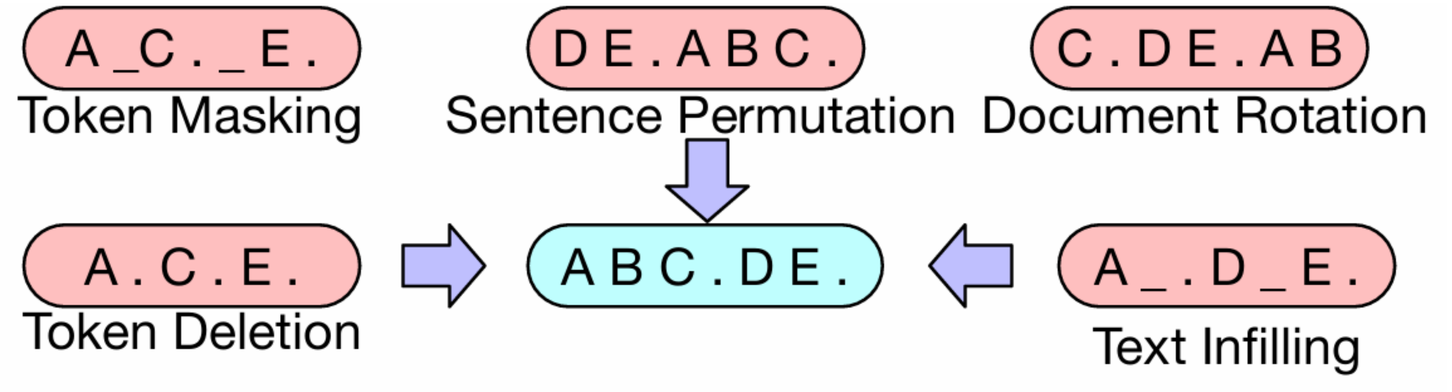


Transformer Architecture. Source:[1706.03762] Attention Is All You Need (arxiv.org)

Approach

❑ BART Model (cont)

- There are several transformations to corrupt text:
  - Token Masking
  - Sentence Permutation
  - Document Rotation
  - Token Deletion Text Infilling.



Transformations for noising the text. Source:[1910.13461] BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension (arxiv.org)

Approach

❑ Fine-tuning

- Fine-tuning is one of the most used method in machine learning nowaday.
- Instead of training from scratch as before, fine-tuning will adapt a pre-trained model for specific tasks.
- There are many ways to fine-tune such as:
  - Full fine-tuning
  - Parameter efficient fine-tuning(PEFT)

Outline

1. Introduction
2. Dataset
3. Approach
4. Evaluation
5. Experiment and Result
6. Future Work
7. References

Evaluation

- In general, the text generation task or specifically, in the text summarization, it is very hard to evaluate the accuracy of the text which is generated by the model since the natural language is diversity and no specific rule to tell which text is wrong or correct unless reading it.

- The following evaluation method to use:
  - **Family of Rouge Score**
  - **BertScore**

Evaluation

❑ Family of ROUGE Score

- ROUGE stands for Recall-Oriented Understudy for Gisting Evaluation.
- ROUGE computes the overlapping words between the generated summary and reference summary with different granularity corresponding to different ROUGE scores:
  - **ROUGE-N**
  - **ROUGE-L**
  - **ROUGE-S**
- In any kinds of ROUGE score, they all compute the precision, recall and f1-score using overlapping.

$$\frac{number\_of\_overlapping\_words}{total\_words\_in\_reference\_summary}$$

Recall formula

$$\frac{number\_of\_overlapping\_words}{total\_words\_in\_system\_summary}$$

Precision formula

Source: [What is ROUGE and how it works for evaluation of summaries? - Kavita Ganesan, PhD (kavita-ganesan.com)]

Evaluation

❑ Family of ROUGE Score (cont)

- **ROUGE-N** : measure compute the precision, recall and f1-score of overlapping of single word or pairs of words,... corresponding to **ROUGE-1** and **ROUGE-2**.

- **ROUGE-L** : measure the overlapping of longest matching sequence of words using Longest common sequence (LCS). An advantage of using LCS is that it does not require consecutive matches but in sequence matches that reflect sentence level word order. There are variants of ROUGE-L is **ROUGE-LSUM**, it measures the ROUGE-L of each sentences and takes the average.

- **ROUGE-S** : measures any pair of word in a sentence in order, allowing for arbitrary gaps. This can also be called skip-gram co-occurrence.

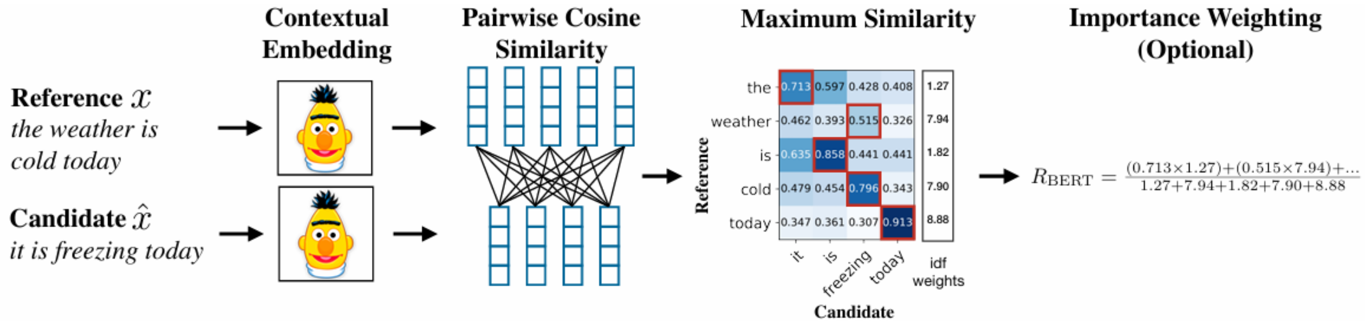
➢ In this report, we will only use ROUGE-1, ROUGE-2, ROUGE-LSUM, ROUGE-L to evaluate.

Evaluation

❑ BertScore

- The BERTSCORE, which is a language generation evaluation metric based on pre-trained BERT contextual embeddings.

- BERTSCORE can measure the semantics between the generated summary and reference summary.



The workflow to compute recall bert score. Source: [1904.09675] BERTScore: Evaluating Text Generation with BERT (arxiv.org)

- We choose model **microsoft/deberta-xlarge-mnli**, which is the model have the best correlation with human evaluation.

Outline

1. Introduction
2. Dataset
3. Approach
4. Evaluation
5. Experiment and Result
6. Future Work
7. References

❑ Experiment

- We will choose the pre-trained model: **sshleifer/distilbart-xsum-12-3**
- We use the Trainer API from transformers hugging face library to fine-tune model easier and more efficiently. We are fine-tuning with following configurations:
  - **Train and validation batch size**: 4
  - **Learning rate**: 3e-05
  - **Optimizer:** Adam with betas=(0.9,0.999) and epsilon=1e08
  - **Lr scheduler type**: linear
  - **Lr scheduler warmup steps**: 500
  - **Num epochs**: 5
  - **Label smoothing factor**: 0.1
- We compute **validation loss** on the validation data and compute the **ROUGE score** of the generated summary for test data at the end of each epoch

Experiment & Result

❑ Result

- Below show the results:

Training Loss	Epoch	Step	Validation Loss	Rouge1	Rouge2	Rougel	Rougelsum	Gen Len
3.4812	1.0	425	3.3209	47.7226	26.3282	35.5063	42.5426	66.523
3.2269	2.0	850	3.1838	50.4271	27.7047	37.2638	45.1897	77.115
2.9504	3.0	1275	3.1401	50.6362	28.2773	37.6	45.4901	74.992
2.8014	4.0	1700	3.1346	51.2942	28.4684	38.0877	46.0386	74.299
2.71	5.0	2125	3.1426	51.2701	28.3575	37.9263	45.8934	75.777

Experiment & Result

❑ Result (cont)

- We try the model with actual news from source: Ronaldo threatens to punch referee after getting sent off - VnExpress International . Below is the results:

> Portuguese striker Cristiano Ronaldo was sent off for the 12th time in his career after his elbow hit Al Nassr's defender Ali Al-Bulaihi's neck in the 86th minute of the Saudi Super Cup on Monday. Ronaldo raised his fist towards referee Mohammed Al-Hoaish and

- If you want to try the model more with different articles, you can go in this link: LA1512/fine-tuned-distilbart-xsum-12-3-news-summary · Hugging Face

Outline

1. Introduction
2. Dataset
3. Approach
4. Evaluation
5. Experiment and Result
6. Future Work
7. References

Future Work

- Find the best sets of hyper-parameter for fine-tuning.
- Find other metrics to evaluate model in other perspectives
- Find the evaluation method that doesn’t need the reference summary
- Try with other model such as T5 or PEGASUS model

Outline

1. Introduction
2. Dataset
3. Approach
4. Evaluation
5. Experiment and Result
6. Future Work
7. References

Outline

[1] Kondalarao vonteru : https://www.kaggle.com/datasets/sunnysai12345/news-summary

[2] Mengli Zhang,Gang Zhou,Wanting Yu,Ningbo Huang,and Wenfen Liu (2021): A Comprehensive Survey of Abstractive Text Summarization Based on Deep Learning (hindawi.com)

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin(2017):[1706.03762] Attention Is All You Need (arxiv.org)

[4] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, Luke Zettlemoyer(2019):[1910.13461] BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension (arxiv.org)

[5] Kavita Ganesan: What is ROUGE and how it works for evaluation of summaries? - Kavita Ganesan, PhD (kavita-ganesan.com)

[6] Chin-Yew Lin(2004): ROUGE: A Package for Automatic Evaluation of Summaries - ACL Anthology

[7] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, Yoav Artzi(2020):[1904.09675] BERTScore: Evaluating Text Generation with BERT (arxiv.org)

[8] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, Yoav Artzi(2020): Tiiiger/bert_score: BERT score for text generation (github.com)

[9] Sam Shleifer: sshleifer/distilbart-xsum-12-3 · Hugging Face

HUST

THANK YOU !

hust.edu.vn fb.com/dhbkhn