News Headline & Summary Generation

through Abstractive Text Summarization

AML Project

Kuber Shahi

Problem Statement

- **Given a news article, generate a summary of two-to-three sentences and a headline for the article.**
- The summary of the article should be abstractive rather than extractive.

Motivation

- News summarizer has widespread application
- Comes in handy to quickly get updated with the news without going into detail.
- Most people don’t want to read the entire article to get the news. Hence, the popularity of news summarizers like Inshorts.
- Not just limited to texts. Can be used to summarize anything that’s text such as reviews, journals, textbooks, articles, etc.

Abstractive vs Extractive Summarization

- In extractive summarization, important sentences and phrases are extracted from the original article as part of the summary.
- In abstractive summarization, new sentences are generated as part of the summary and the sentences in the summary might not be present in the article.
- Summary may use paraphrasing techniques
- Abstractive challenging than extractive as it is difficult to generate new grammatically and semantically correct summary.

PEGASUS for Abstractive Summarization

- PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization
- Transformer Encoder-Decoder Model
- Pre-training objective: Mask Language Modelling (MLM) and GSG Gap-Sentence Generation
- 16-Layers of Encoder and 16 Layers of Decoder

MLM and GSG

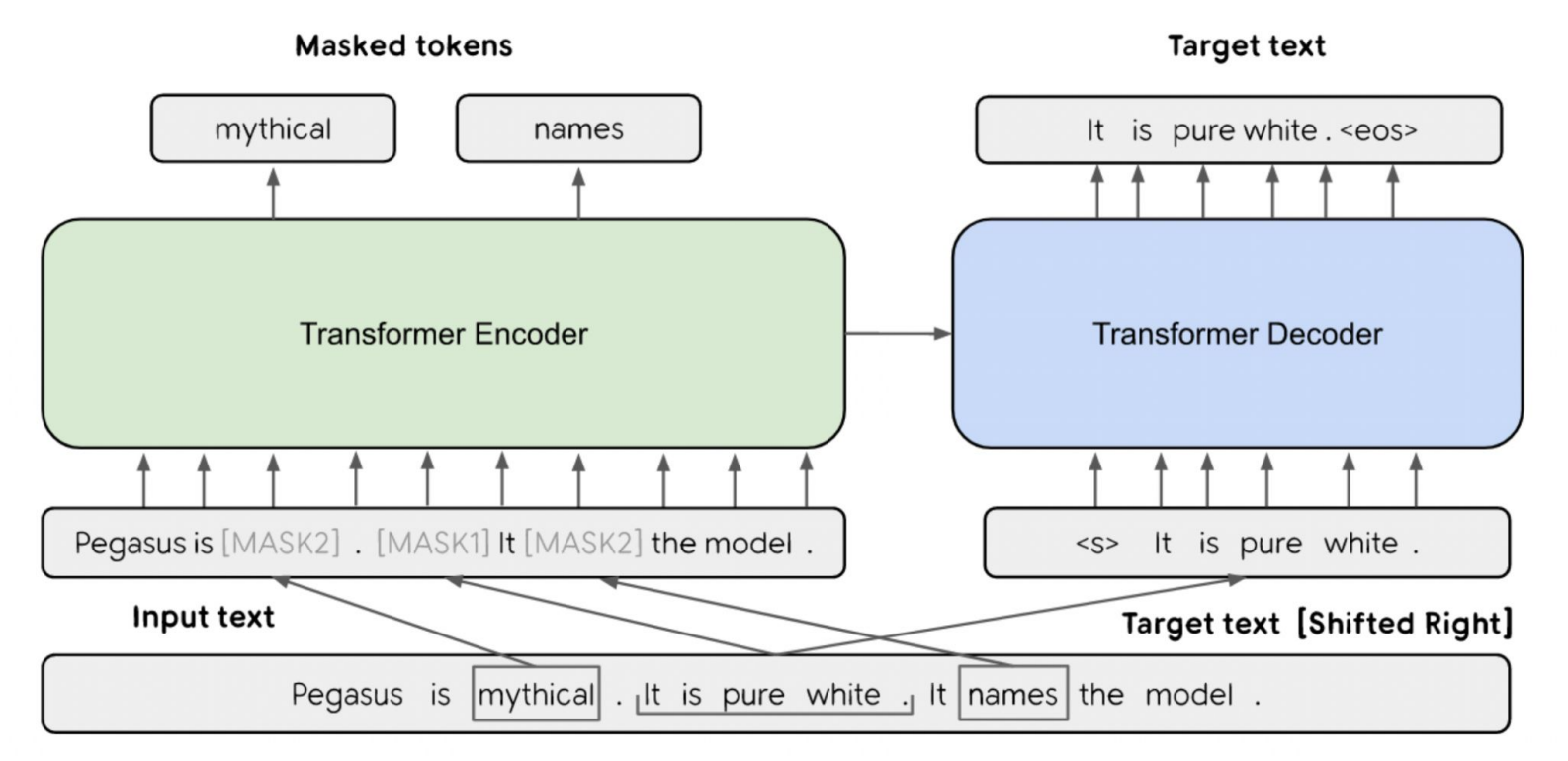
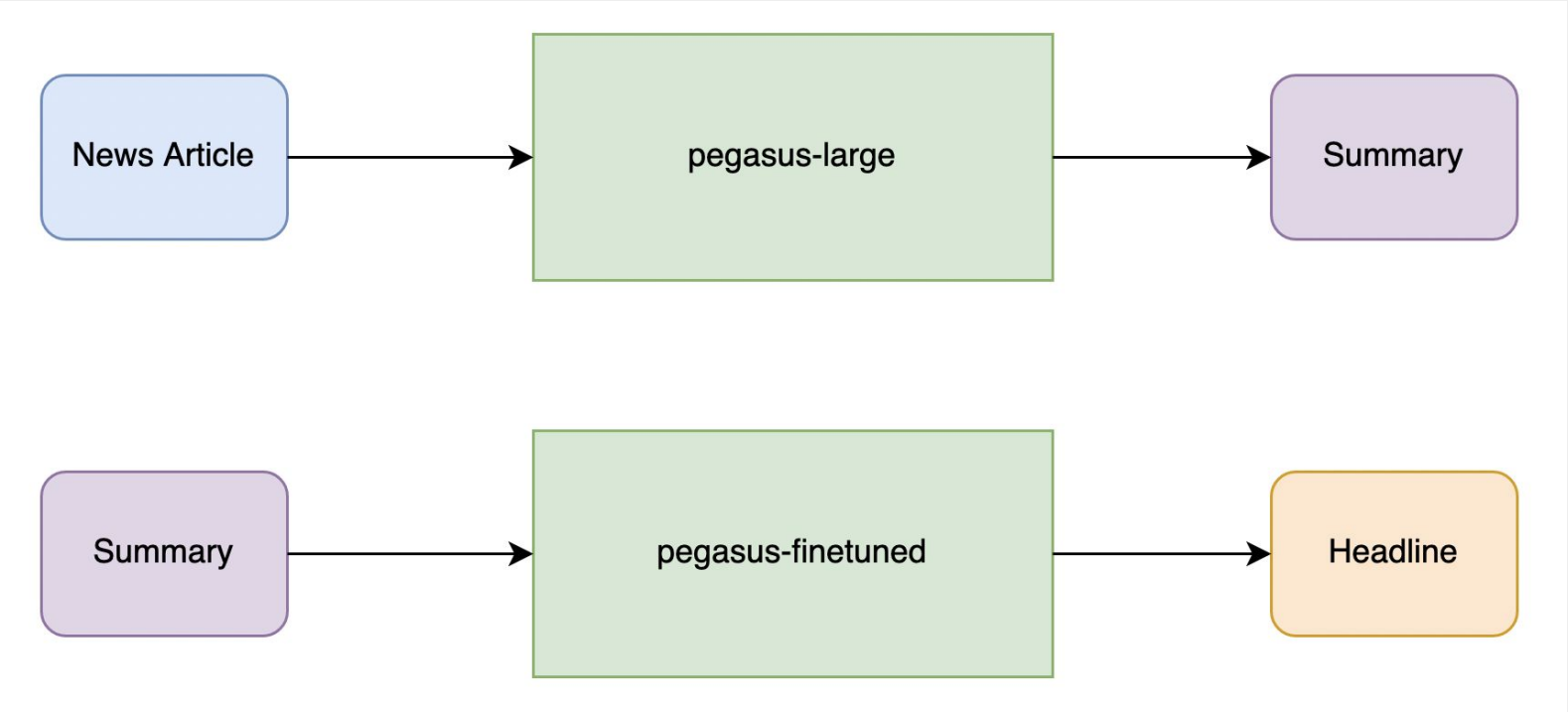


Fig 1.1: PEGASUS architecture overview (source: pegasus paper)

Pretraining

- Pretraining on two large general text corpora: C4 (350M webpages) and Hugenews (1.5B articles)
- Pretrained PyTorch Model available on Hugging Face Transformer as “pegasus-large”
- Has 568 Million parameter
- Vocab-size: 96103
- Maximum Token Length: 1024

Implementation



Fine-tuning dataset

- Using the ‘Indian News Summary’ Dataset from Kaggle scrapped from different news sources, primarily Inshorts
- Has around 98,000 thousand example. Did 80-20 train-test split

headlines	text
upGrad learner switches to career in ML & AI with 90% salary hike	Saurav Kant, an alumnus of upGrad and IIIT-B's PG Program in Machine learning and Artificial Intelligence, was a Sr Systems Engineer at Infosys with almost 5 years of work experience. The program and upGrad's 360-degree career support helped him transition to a Data Scientist at Tech Mahindra with 90% salary hike. upGrad's Online Power Learning has powered 3 lakh+ careers.
Delhi techie wins free food from Swiggy for one year on CRED	Kunal Shah's credit card bill payment platform, CRED, gave users a chance to win free food from Swiggy for one year. Pranav Kaushik, a Delhi techie, bagged this reward after spending 2000 CRED coins. Users get one CRED coin per rupee of bill paid, which can be used to avail rewards from brands like Ixigo, BookMyShow, UberEats, Cult.Fit and more.
New Zealand end Rohit Sharma-led India's 12-match winning streak	New Zealand defeated India by 8 wickets in the fourth ODI at Hamilton on Thursday to win their first match of the five-match ODI series. India lost an international match under Rohit Sharma's captaincy after 12 consecutive victories dating back to March 2018. The match witnessed India getting all out for 92, their seventh lowest total in ODI cricket history.

Fine-tuning

- Inference on a single example took around 1 minute in average
- Single Example,

```
 "The trailer of actress Manisha Koirala's comeback film 'Dear Maya' has been released. The film will revolve aroun
d the story of a lonely middle-aged spinster, portrayed by Manisha. Written and directed by Sunaina Bhatnagar, the
film also stars Pakistani VJ Madiha Imam and is scheduled to release on June 2.",
```

```
"Trailer of Manisha Koirala's comeback film 'Dear Maya' out",
```

- Pegasus-large generation:

```
'The film will revolve around the story of a lonely middle-aged spinster, portrayed by Manisha.',
```

- Pegasus-finetuned

```
I allows stock exchanges to extend trading hours in equity derivatives', "Manisha Koirala's comeback film 'Dear May
a' trailer released", 'My business is my business, none of your business: Daisy Shah', 'Khaleda sentenced to 7 year
```

Fine-tuning

- Fine-tuned on only 10000 examples for 10 epochs as training which took around 1 hr on gpus
- Evaluation metric: Rouge-1-F

  Evaluated on 2500 examples

  Before fine-tuning: 0.13

  After fine-tuning: 0.08
- Rouge-1-F decreases? Needs to be finetuned on more data.

Demo with Test example

Article:Elon Musk Would Reverse Twitter's Ban on Trump

Summary:

```
Summary Decoded.
['Elon Musk said on Tuesday that he would “reverse the permanent ban” of former President Donald J. Trump on Twitter and let
him back on the social network, in one of the first specific comments by Mr. Musk, who struck a deal last month to buy Twitt
er for $44 billion, said at a Financial Times conference that the company’s decision to bar Mr. The billionaire, who also le
ads the electric carmaker Tesla and the rocket company SpaceX, has called himself a “free speech absolutist” and has said he
is unhappy with how Twitter decides what can and cannot be posted online.']
CPU times: user 6min 8s, sys: 6min 46s, total: 12min 55s
Wall time: 2min 21s
```

Headline:

```
['Elon Musk to let Trump back on Twitter']
CPU times: user 49.1 s, sys: 53.5 s, total: 1min 42s
Wall time: 21.7 s
```

GitHub Link:

All implementation files are uploaded on GitHub at
https://github.com/kubershahi/ashoka-aml

Thank You