Generative model for news summarization

Miguel Cózar Tramblin A20522001
Carlos Munoz Losa A20521562
David Arias Cuadrado A20521638

1st March 2023

1 Description of the problem

News summarization is a crucial task for information retrieval and presentation. It involves condensing lengthy news articles into concise summaries, preserving the most important information and key points. In recent years, there has been a growing interest in using machine learning techniques, especially generative models, to improve the effectiveness of news summarization.

The objective of this project proposal is to develop a generative model for news summarization that can effectively summarize news articles in a coherent and informative manner. The model should be able to take in a news article as input and generate a concise summary that captures the essence of the article.

The proposed generative model for news summarization is expected to produce high-quality summaries that are informative and concise. The model should be able to capture the key points of a news article and present them in a coherent and readable format. The effectiveness of the model will be evaluated using standard metrics, and the results will be compared to existing state-of-the-art models for news summarization. Finally, we are aiming to train our model for making summaries with social networks length constraints so that the model can be used for publishing news summaries on Twitter.

2 State of the art

News summarization is the task of generating a shorter version of a news article while retaining the most important information. Generative models are a popular approach to this problem, where a model is trained to generate summaries based on input news articles. The state of the art in this area can be characterized by the following:

1. Transformer-based models: Transformer-based models, such as BERT[1], GPT-2, and T5, have shown great success in natural language processing tasks, including news

[1]Ma, K., Tian, M., Tan, Y. et al. What is this article about? Generative summarization with the BERT model in the geosciences domain. Earth Sci Inform 15, 21–36 (2022). https://doi.org/10.1007/s12145-021-00695-2

summarization. These models are pre-trained on large amounts of text data and fine-tuned on specific summarization tasks, resulting in high accuracy and fluency in generating summaries.

2. Encoder-Decoder architectures: Many state-of-the-art models for news summarization use an encoder-decoder architecture, where the input article is encoded into a fixed-length vector and then decoded into a summary. The encoder and decoder can be based on different types of neural networks, such as LSTMs, CNNs, and Transformers.[2]

3. Reinforcement Learning: Another approach to generative models for news summarization is using reinforcement learning, where the model learns to optimize a reward function that measures the quality of the generated summary. This approach has shown promising results in generating more concise and coherent summaries[3].

4. Evaluation Metrics: Finally, the state of the art in news summarization also includes the development of evaluation metrics that assess the quality of the generated summaries. Common metrics include ROUGE (Recall-Oriented Understudy for Gisting Evaluation), which measures the overlap between generated and reference summaries, and BLEU (Bilingual Evaluation Understudy), which measures the n-gram similarity between generated and reference summaries.

Overall, the state of the art in generative models for news summarization combines the latest advances in neural network architectures, pre-training techniques, reinforcement learning, and evaluation metrics to generate accurate, fluent, and concise summaries of news articles

3 Planning

- Week 1: Finalize project proposal and project plan.
  - Develop a detailed project plan with specific tasks, timelines, and milestones.
  - Finalize the project proposal and obtain necessary approvals.
  - Set up necessary tools and software for the project.
- Week 2: Collect and preprocess the data.
  - Collect a suitable dataset of news articles and corresponding summaries.
  - Preprocess the data, including cleaning, formatting, and tokenizing.
- Week 3: Develop and train the generative model.
  - Develop a generative model for news summarization
  - Train the model using the preprocessed dataset from Week 2.
- Week 4: Evaluate the performance of the model.

[2]P. Li, J. Yu,J. Chen and B. Guo, "HG-News: News Headline Generation Based on a Generative Pre-Training Model," in IEEE Access, vol. 9, pp. 110039-110046, 2021, doi: 10.1109/ACCESS.2021.3102741.

[3]Singh, R.K., Khetarpaul, S., Gorantla, R. et al. SHEG: summarization and headline generation of news articles using deep learning. Neural Comput & Applic 33, 3251–3265 (2021). https://doi.org/10.1007/s00521-020-05188-9

  - Evaluate the performance of the model using standard metrics such as ROUGE and BLEU.
  - Fine-tune the model based on evaluation results.
- Week 5: Implement length constraints and test the model.
  - Implement length constraints based on the specific requirements of social networks such as Twitter.
  - Test the model with the constraints to ensure that it produces summaries that comply with the length constraints.
- Week 6: Prepare final report and project presentation.
  - Prepare a final report summarizing the methodology, results, and conclusions of the project.
  - Prepare a presentation to share the findings and outcomes of the project.
  - Submit the final report and deliver the presentation.

Note: The above timeline is based on an aggressive schedule to complete the project in 6 weeks. The timeline may need to be adjusted based on the complexity of the project and available resources.