CS7.401. Intro to NLP

Project Outline

Group Name: RantProMax

Topic: Text summarization specialized for news headline generation.

Member list:

Roll no.	Name
2022201006	Jaffrey Joy
2022201013	Amit Marathe
2022201046	Atharva Vaidya

Project Description

Text summarization involves condensing a longer piece of text while retaining key information and the overall meaning of the original text. The resulting summary is a shorter and more concise version of the original. This project plans to overcome some of the concerns of existing text summarization architectures (recurrent neural nets and their variants with attention, transformer), specifically used for converting circumlocutory news articles into crisp headlines. A major challenge in headline generation is the preservation of named entities. Named entities are names of people, places, products, etc., lost in this process due to the scarcity of the word and the model's inability to learn its latent meaning. These named entities are important for better headline summarization. In this project, we try out approaches that preserve these named entities in the output headline resulting in a better summarization.

“Why use many word, when few word do trick.” - Malone et al.

Available Datasets

1. **NewSHead**[1]: This dataset consists of news article headlines and their corresponding summaries. The evaluation metrics used for this dataset are ROUGE (Recall-Oriented Understudy for Gisting Evaluation) and BLEU (Bilingual Evaluation Understudy).
2. **TAC2011**[2]: This dataset contains multi-document summarization data from the TAC 2011 conference. The evaluation metrics used for this dataset are ROUGE and Pyramid.
3. **DUC03+04**[3]: This dataset consists of news articles and their corresponding summaries from the DUC (Document Understanding Conference) 2003 and 2004. The evaluation metrics used for this dataset are ROUGE, Pyramid, and F-measure.
4. **Multi-News**[4]: This dataset contains news articles and their summaries from multiple sources. The evaluation metrics used for this dataset are ROUGE and BLEU.

We are also planning to build our own dataset by scraping from various online news websites and a popular news aggregator called *Inshorts*, where a variety of news articles and their corresponding headlines are readily available.

Literature Review on previously conducted research

Apart from being dependent on an optimal function, text summarization also relies on a sentence similarity measure, up to a certain extent. It can significantly improve the efficiency of abstractive summarization techniques. Masum el al. modified existing models and algorithms to generate headlines [5]. In addition to this, some further processing, like forming classifications for named entities, have been carried out by them to improve system accuracy and minimize possible problems [5].

Hanunggul and Suyanto presented a comparison between the two types of attention: global and local. They observed that a larger number of words pertaining to the original summary were produced in the global attention-based model. While in the local attention-based model, more sets of words from the original summary were produced. The reason behind such an outcome is the subsets of input words get considered instead of entire input words in the local attention implementation [6].

In 2015, Sutskever el al. made the first attempt to summarize the text using recurrent networks based on encoder decoder architecture [7]. Later in 2016, the model designed by Loung el al. on the famous CNN/daily mail dataset was the stepping stone for abstractive summarization [8]. The next model to give promising results was the pointer generator [9], a hybrid model that not only points the words directly to the summary but also generated new ones from the vocabulary.

Masum et al. developed an efficient way of summarizing text using sequence-to-sequence RNNs [10]. They proposed a method of successfully reducing the training loss that occurs while training the model. The steps involved in their methodology include data preprocessing, counting vocabulary size and then going on to adding word embeddings and passing it to the encoder decoder layer with LSTM. One of the limitations of their work is that it does not provide good results for large text inputs.

A paradigm shift in natural language processing occurred with the introduction of the transformer [11]. It uses self-attention as a way to find dependencies instead of recurrence. This design gave state-of-the-art results on a plethora of NLP tasks and was efficient due to the parallelization in the architecture.

Project Milestones and Schedule

- We plan on working with the input data, cleaning and preprocessing the articles from various sources, using libraries like Spacy for NER while simultaneously working on the basic vanilla RNN seq-seq model.
- Following the results of the previous stage we plan on implementing more sophisticated architectures leading up to the Transformer

Week	Milestone
1 - 2	Scraping data from various sources to create a dataset and Data Preprocessing
2 - 3	Basic Seq-Seq Encoder Decoder Model
3 - 4	Seq-Seq models that implement Attention
4 - 6	Transformer
7	Documentation

References:

[1] Yuning Mao et al., WWW '20: Proceedings of The Web Conference 2020. Pages 1773–1784. https://doi.org/10.1145/3366423.3380247

[2] Karolina Owczarzak and Hoa Trang Dang. 2011. Overview of the TAC 2011 summarization track: Guided task and AESOP task. In Proceedings of the Text Analysis Conference (TAC 2011).

[3] Over Paul and Yen James. 2004. An introduction to duc-2004. In Proceedings of the 4th Document Understanding Conference (DUC 2004).

[4] Alexander Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. 2019. Multi-News: A Large-Scale Multi-Document Summarization Dataset and Abstractive Hierarchical Model. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Pages 1074–1084. https://doi.org/10.18653/v1/P19-1102

[5] Masum KM, Abujar S, Tusher RTH, Faisal F, Hossai SA (2019) Sentence similarity measurement for Bengali abstractive text summarization. In: 2019 10th international conference on computing, communication and networking technologies (ICCCNT), Kanpur, India, 2019, pp 1–5. https://doi.org/10.1109/ICCCNT45670.2019.8944571

[6] Hanunggul PM, Suyanto S (2019) The impact of local attention in LSTM for abstractive text summarization. In: 2019 international seminar on research of information technology and intelligent systems (ISRITI), Yogyakarta, Indonesia, 2019, pp 54–57. https://doi.org/10.1109/ISRITI48646.2019.9034616

[7] Sutskever I, Vinyals O, Le QV (2014) Sequence to sequence learning with neural networks

[8] Luong M-T, Pham H, Manning CD (2015) Effective approaches to attention-based neural machine translation

[9] See A, Liu PJ, Manning CD (2017) Get to the point: summarization with pointer-generator networks

[10] Mohammad Masum K, Abujar S, Islam Talukder MA, Azad Rabby AKMS, Hossain SA (2019) Abstractive method of text summarization with sequence to sequence RNNs. In: 2019 10th international conference on computing, communication and networking technologies (ICCCNT), Kanpur, India, 2019, pp 1–5. https://doi.org/10.1109/ICCCNT45670.2019.8944620

[11] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017) Attention is all you need