Journal to Wiki Text Style transfer: Simplifying the medical literature for broader comprehension

Team Word Nerds

Alex Jonas, Annie Lam, Matthew L. Senjem, & Luis Silva
University of Minnesota
Minneapolis, MN 55455, USA
{jonas060,lam00058,senje001,silva364}@umn.edu

1 Introduction / Motivation

The motivation for translating medical literature to a Wikipedia structure is the simplification of the medical literature. Translating a medical research paper to a Wikipedia article is akin to summarizing the major findings of the paper in a digestible format. The simplification and increased digestibility of a Wikipedia article would allow broader comprehension for individuals not privy to reading medical research. For the select few who are – the result is less time needed to read and comprehend. Thus, the laity and scientific researcher alike are given another tool in their personal arsenal for interpreting and making sound decisions based off the latest medical research in a timely manner.

2 Literature survey

2.1 Text Style Transfer

In order to solidify our understanding of the text style transfer task and inform our methodology, we utilized a review of common definitions and methods used for the task. For a general definition of the task, text style transfer is a NLP task that aims to change the written style of a text, while preserving its original meaning (Toshevska and Gievska, 2022). There are multiple linguistic styles that can be targeted for this task, and their choice usually depends on the individual style, genre, as well as formality and politeness specific to the situation. For our project we will be attempting to transfer the typically dense language of academic writing to a more simplified, easy to understand style.

For the data used in this task, there are two main types of datasets: parallel and nonparallel. Parallel datasets have text that is directly transferred from the originally used text, usually by humans. Nonparallel datasets offer general categories of texts that can be compared but that are not directly associated between them. Our project will utilize a parallel dataset of wiki blog posts and academic journal articles. Evaluation metrics of text style transfer include by word overlap methods such as BLEU or ROUGE, and n-gram-based methods such as PINC. To evaluate our model, we will use the word overlap ROUGE metric.

There are two main stages in text style transfer, representation learning and sentence generation. And in these specific stages, different models have been attempted. For representation learning, Two types of RNNs, LSTM and GRU have been previously published, while GPT-2 has been favored in past published articles for sentence generation. For these stages, we will explore different models to find those best suited for our goal.

2.2 Summarization

For the summarization portion of our task, we will test out a transformer model introduced in the paper '*Investigating Efficiently Extending Transformers for Long Input Summarization*'(Phang et al., 2022). The authors of this article introduce a new transformer model called Pegasus-X that specifically targets long input summarization tasks. The most extensively researched transformer models are typically tested on short input lengths of around 500 to 2000 tokens (Phang et al., 2022). The input articles for our project will be much longer, with the academic articles commonly exceeding 4000 words. The PEGASUS-X model includes long input pretraining that allows it to handle inputs up to 16,000 tokens, making it well suited for our task (Phang et al., 2022). The model was found to outperform common models on the arXiv, Big Patent, and PubMed tasks (Phang et al., 2022).

2.3 Similar Tasks

To draw inspiration for our methodology, we researched papers working on similar tasks. In a previous work, (Liu et al., 2018) collected a large dataset named, WikiSum, (available on paperswithcode), and developed a method for generat-

ing Wikipedia articles from multi-document inputs. For the dataset curation, they crawled Wikipedia and followed links to the reference documents. They developed a two part approach where they first use extractive summarization to create a summary of the input text for each topic, then use a neural abstractive model, composed of a decoder-only transformer architecture to generate the final wiki article output. They explored several different extraction methods and neural abstraction methods, and showed that their decoder-only transformer model performed best at the abstraction task, among the models studied. Their decoder-only transformer model took as inputs the extracted text and the Wikipedia text concatenated together with a separating character, and the architecture consisted of an initial "memory compressed" transformer decoder block with an additional convolutional layer on the values and keys, followed by a "local attention" module where the inputs are split into smaller fragments before passing to the multi head attention block, then merged back together afterwards. We plan to explore using the dataset and methods from this paper in developing our model. For example, we will explore using the WikiSum dataset to pretrain our model, then use our curated dataset to fine tune the pre-trained model.

The next article provides guidance on how to organize and collect our data. Additionally, the paper is also useful for decoding within a Wikipedia summarization format. The paper introduces a dataset, WikiTableT, which contains Wikipedia article sections and their corresponding tabular data and various metadata (Chen et al., 2020). The dataset contains millions of instances while covering a broad range of topics and a variety of generation tasks. The resources considered for the creation of the WikiTableT are Wikidata table, info boxes in Wikipedia pages, hyperlinks in the passages (obtained from named entity recognition), and Wikipedia article structure (Chen et al., 2020). Given a Wikipedia article, the sectional data, infobox, and Wikidata table are used to generate text based off a Wikipedia section for a given Wikipedia article. The infobox and Wikidata table are termed the “background knowledge” for the article and used for all sections of a given Wikipedia article. The structure of this dataset provides insight into how we should approach our data collection methodology.

For models trained on this dataset the best decoding method was determined to be Beam-Search, and n-gram blocks proved particularly useful in mitigating grammatical errors. Human evaluation metrics were used to evaluate the models trained on WikiTableT. Relevance, support, grammar, coherence, and faithfulness were scored from 1 to 5. It was determined that people were unable to differentiate between written text and generated text from the neural models. Given the strength of the human evaluation results, we will look into using some of the methods mentioned in our model implementation.

3 Problem definition

The problem we are aiming to solve is to collect a dataset of scientific papers and corresponding Wikipedia articles, and use this dataset to build an automated system for generating Wikipedia style articles form scientific papers. We will collect a dataset of scientific papers, with corresponding wiki or blog posts summarizing the articles in wiki form. We will then use the resulting dataset to train a model for performing the wiki style summarization task for each article. We will explore various options for the model architecture, e.g. fine tune a pre-trained model from one of the references mentioned in the literature survey section above.

4 Proposed Idea and Hypothesis

4.1 Initial Idea

Our main idea to create a NLP tool that can summarize medical research papers and transform their style into a wiki format. The idea is that by building a model based on a parallel dataset of medical articles and wiki summaries scraped from the web, we will be able to simplify texts from the medical literature into a more palatable style of reading, in order to reach a wider audience.

4.2 Data Collection

We will explored a few different options for collecting the dataset. One option is using R Selenium, and another option, suggested by our mentor Zae, is using scrapy in combination with xpaths to get the summary and the link to the corresponding PDF file. Once we have the set of PDFs we will ingest the text from each PDF using PyPDF, and assemble the PDF along with the corresponding wiki text into training and test sets. For example, one simple method would be to access the Random page link from the Wiki Journal Club, extract the text on the

page as "Wiki text", extract the text from the linked PDF as "Original Text".

4.3 Model Development and Training

We will explore different model architectures for training a model to perform the wiki generation task, such as those mentioned in the literature survey above. We plan to utilize pre-trained models using the datasets listed above, and use our collected dataset to fine tune a pre-trained model. We will then use metrics such as ROGUE-L F1 score, and human evaluation, to assess the quality of the outputs.

4.4 Group Roles

For the initial development of the project, we plan to divide into two teams: The data collection team will be Luis Silva and Alex Jonas, and the model development and training team will be Annie Lam and Matthew Senjem. We plan to have regular meetings and interactions within the sub-teams and together as a group to iterate through our development process on both the data collection and model development portions of the project. The initial sub team assignments will help us subdivide the tasks and get an initial focus, but roles will likely be fluid over time, as we will work as a team on development on all aspects of the project.

4.5 Hypothesis

Our working hypothesis is that it is possible to build a NLP model that can effectively summarize and transfer styles from medical literature to wiki texts. It is our belief that an existing NLP model, perhaps with pre-training on an existing similar dataset and task, can be leveraged to perform this task, and fine tuned using the dataset we will collect, from parallel examples of previous papers that have been summarized as wiki texts.

5 Broader Impact

The impact of simplifying medical literature would be increased public scientific literacy and public engagement with science and technology, With the ultimate goal and hope of improving the overall health, education and well-being of a wider group of individuals in society.

6 Addressing In-Class Feedback

In the class presentation feedback, both classmates and the professor indicated that building the parallel dataset of medical blog summaries and their respective medical journal articles would be time intensive. So, we decided to split our team into two groups to map out tentative role assignments. Matt and Annie will work on building the model, and Alex and Luis will compile the dataset. This will allow us to work on both tasks simultaneously and optimize our workflow.

In terms of the implementation of the model, as the professor suggested, we will approach this problem as a combination task, with elements of both text style transfer, and text summarization tasks, and use methods from both of those problem domains. For the data collection portion, we will be using RSelenium in addition to exploring the BeautifulSoup python package suggested by our classmates and TAs for web scraping. We will start with a baseline of 20 articles, as recommended by our professor, and supplement with additional articles if needed. If time allows, as a final step for our project, we will consider human evaluation to supplement our evaluation metrics by having students evaluate the quality and helpfulness of the summaries, as our classmates suggested.

References

Mingda Chen, Sam Wiseman, and Kevin Gimpel. 2020. Generating wikipedia article sections from diverse data sources. *CoRR*, abs/2012.14919.

Peter J. Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. 2018. Generating wikipedia by summarizing long sequences.

Jason Phang, Yao Zhao, and Peter J. Liu. 2022. Investigating efficiently extending transformers for long input summarization.

Martina Toshevska and Sonja Gievska. 2022. A review of text style transfer using deep learning. *IEEE Transactions on Artificial Intelligence*, 3(5):669–684.