Creating Structured Viewing of Paper Abstracts

Group Name: Team Abstract Segmentation

Group Members: Awais Naeem, Christian Martin, Allison Pujol

Objective

Novel medical research findings – and being able to understand and communicate those findings – are crucial to a healthy and effective public health system. When embarking on the research process, many biomedical experts will look at a published paper's abstract to determine whether something is worth reading. However, many medical research abstracts use highly technical language that slows down overall readability and sentence comprehension [1]. Furthermore, a researcher may want to compare only one part of the abstract against another abstract (such as seeing different methods for the same biological process, etc). As an NLP intervention, our group thus proposes a structured viewing of paper abstracts through which will classify sentences according to the part of the paper it describes. Our goal is to improve the readability and structure of "tricky" abstracts; and while this project does focus on medical abstracts, we see the benefits of applications to other academic disciplines.

Dataset

The dataset we will be using is composed of 200,000 medical research paper abstracts [1] from PubMed [2]. In the dataset, each abstract is splitted into sentences with a distinct label assigned to each sentence i.e., "background," "methods," "results," or "conclusion," which is given as the first word in the line. There are also blank lines and comments (lines starting with a "#") that will need to be removed.

The raw training/testing/validation dataset files will firstly need to be converted into formatted csv files such that one column will list individual sentences of the abstracts and a second column will list the corresponding label without keeping track of the abstract identity. For the runtime evaluation, an abstract will be splitted using sentence tokenizer and each sentence will be assigned to a separate class label.

Methods

To find the best way of classifying the sentences, we will use & evaluate three different approaches to determine which is most effective.

The first method will be to use a traditional ML approach of creating GloVe/Word2Vec sentence embeddings and then using a model to predict classes.

The second method will be to use a fine-tuned LLM model, such as T5, in a Sequence Classification task.

The third method will be to use a deep learning approach that utilizes an RNN Model and LSTM/GRU for short-term memory.

To evaluate the models, we will split the dataset and compare the results based on the labels in the dataset. Beyond this, we'd also like to explore presenting this dissection of the abstract in a way that is useful to readers. While we don't have evaluation metrics, this includes presenting the experience through a UI and extracting entities using NER to be used as the keywords for any abstract.

Milestones

- Getting Medical Research Paper Abstracts dataset from Kaggle [2]
- Preprocessing the raw dataset files to extract the individual abstract sentences and corresponding labels into formatted csv files
- Implementation of ML multi-class classification algorithm (XGBoost) using GloVe/Word2Vec embeddings
- Training a deep recurrent neural network (RNN) using LSTM/GRU layers for multi-class classification
- Fine-Tuning an LLM Model for a sequence classification task
- Evaluation, analysis and comparison of the results (accuracy, precision, recall, F1-score) for different classification models using test data
- Extraction of entities in the abstracts using NER
- A simple GUI which will take an abstract and divide each sentence of the abstract under different labels using the best classification model as per the accuracy measurement
- Presentation
- Report Writing

Deliverables

1. A jupyter notebook containing all the code will be hosted on a public Github repository
2. A link of the dataset [2] to download the dataset from the Kaggle website
3. Project Report
4. Presentation with live demo including the simple GUI for abstract segmentation

References

[1] Tal August, Lucy Lu Wang, Jonathan Bragg, Marti A. Hearst, Andrew Head, and Kyle Lo. 2023. Paper Plain: Making Medical Research Papers Approachable to Healthcare Consumers with Natural Language Processing. ACM Trans. Comput.-Hum. Interact. 30, 5, Article 74 (September 2023), 38 pages. https://doi.org/10.1145/3589955

[2] 200000 Medical Research Paper Abstracts https://www.kaggle.com/datasets/anshulmehtakaggl/200000-abstracts-for-seq-sentence-classification

[3] PubMed https://pubmed.ncbi.nlm.nih.gov/