Automated Scientific Literature Review
using classical NLP techniques

Abstract

This project attempts to demonstrate the capabilities of an Automated Scientific Literature Review. For this version, a niche topic has been selected to follow through the pipeline of the project, and the topic is a healthcare-based aspect "Alzheimer's Disease Biomarkers". A healthcare topic is chosen because healthcare is a rapidly developing area with a huge variety of research published, and NLP can be deployed in several aspects to accelerate knowledge dissemination. This project is uploaded on: https://github.com/visalakshi2001/applied-nlp-project1[1]

1 Credits

This document has been adapted from the instructions for earlier ACL proceedings, including those for ACL-2012 by Maggie Li and Michael White, those from ACL-2010 by Jing-Shing Chang and Philipp Koehn, those for ACL-2008 by Johanna D. Moore, Simone Teufel, James Allan, and Sadaoki Furui, those for ACL-2005 by Hwee Tou Ng and Kemal Oflazer, those for ACL-2002 by Eugene Charniak and Dekang Lin, and earlier ACL and EACL formats. Those versions were written by several people, including John Chen, Henry S. Thompson and Donald Walker. Additional elements were taken from the formatting instructions of the *International Joint Conference on Artificial Intelligence*.

2 Introduction

This system leverages state-of-the-art NLP techniques to efficiently process, summarize, and extract key information from a large corpus of research papers. By automating the literature review process, we aim to significantly reduce the time and effort required for researchers to gather relevant information and identify important trends in the field. I collected 1000-2000 research papers, using PubMed PMC's API and scraped the contents of the paper from the sections of the response. Utilized Cohere's embeddings stored in Pinecone for efficient storage and retrieval. Then, using classical NLP models for summarization, named entity recognition to extract focused content and main genetic biomarkers explored in the paper. The resulting Streamlit application offers researchers a tool for quick, comprehensive literature reviews as a compilation of the entire pipeline.

3 Data Collection and Storage

The data collection process focused on acquiring a comprehensive dataset of research papers related to Alzheimer's Disease Biomarkers. I utilized the PubMed Central (PMC) API to access open-access papers in XML format. The structured notation of XML allowed for efficient scraping and extraction of content using specialized parsing techniques. The extracted text underwent a thorough cleaning process using regular expressions (regex) to remove irrelevant information and standardize the format. This resulted in a clean, full-text corpus of research papers. To enable efficient retrieval and similarity search, we employed Cohere's embedding model to convert the cleaned text into high-dimensional vector representations. These embeddings capture the semantic meaning of the text, allowing for more nuanced and context-aware searches. The embedded tokens were then stored in a Pinecone vector database. This database structure allows for quick retrieval of relevant full-text articles based on semantic similarity.

[1] GitHub: https://github.com/visalakshi2001/applied-nlp-project1

4 Model Training and Evaluation

I evaluated multiple models for each task to determine the most effective approach for our specific use case in Alzheimer's Disease Biomarkers research. For text summarization, we compared three state-of-the-art models:

1. T5 (Text-to-Text Transfer Transformer)
2. BART (Bidirectional and Auto-Regressive Transformers)
3. PEGASUS (Pre-training with Extracted Gap-sentences for Abstractive Summarization)

BART was selected as our primary summarization model due to its highest average performance across precision, recall, and F1 scores (see analysis below).

Named Entity Recognition (NER) is crucial for extracting protein and gene biomarkers from the papers. Utilizing the Clinical model BioBERT (Bidirectional Encoder Representations from Transformers for Biomedical Text Mining) for its superior performance in identifying biomedical entities within our specific domain. BioBERT is trained on Clinical dataset JNLPBA for extracting Protein, DNA, Cell Type, Cell Line, RNA.

While I initially planned to fine-tune these models on our specific dataset, unexpected hardware issues caused prevention from completing this process. However, the attempt and associated code are included in the repository for future reference and potential implementation.

The evaluation of these models was conducted using standard metrics: ROUGE scores (ROUGE-1, ROUGE-2, ROUGE-L) and Bert scores. The analysis using BERT scores are included in the section Figures.

5 Compilation (User-Interface)

We utilized Streamlit to create an intuitive user interface for our Automated Scientific Literature Review System. The app efficiently processes user keywords, presenting information to K most relevant papers in a standardized format. This app is optimized for 12 papers, and runs fairly fast for more papres. While executing this app and running it on local device is flawless and smooth with no exceptions, but deployment attempts [2] were hindered by model size limitations on the chosen platform. Future work may explore alternative hosting solutions to make the system widely accessible.

[2]URL (currently down): autoreview.streamlit.app

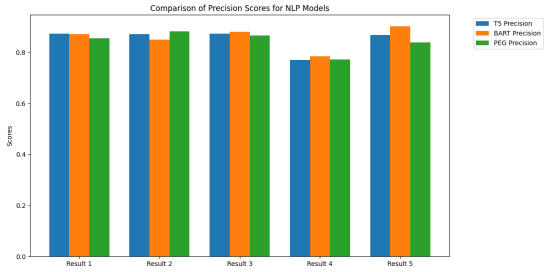
6 Conclusion

This project demonstrates the potential of NLP techniques in automating scientific literature reviews, particularly in the domain of healthcare. By leveraging classical models for summarization, named entity recognition, and relevance scoring, I've attempted to streamline the process as fine as possible. The integration of PubMed data, efficient storage solutions, and a user-friendly interface showcases the practical application of AI in accelerating medical research. While challenges persist, the system's local performance highlights its potential to revolutionize how researchers interact with vast amounts of scientific literature.

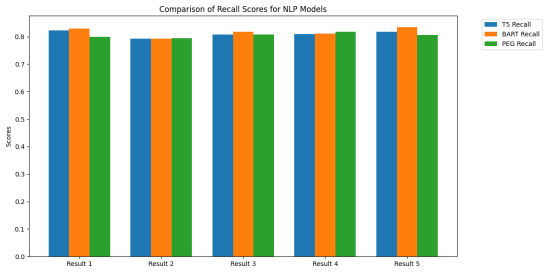
7 Limitations and Future Directions

Our system, while effective, has areas for improvement. The inability to fine-tune models due to hardware constraints potentially limits performance. The system currently lacks integration with real-time database updates, restricting it to a static dataset. Future work could explore incorporating more advanced NLP techniques, such as zero-shot learning or few-shot learning, to enhance adaptability. Expanding the system to cover multiple medical domains and implementing a more robust deployment solution would increase its utility. Additionally, incorporating user feedback mechanisms could help refine the relevance ranking and summary quality over time.

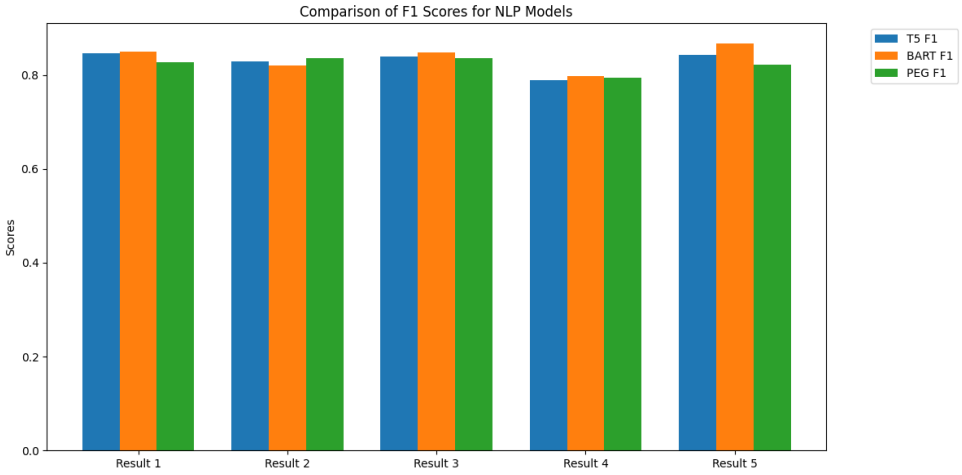
8 Figures



(a) Precision



(b) Recall



(c) F1 Score

Figure 1: Performance comparison of NLP models for Precision, Recall, and F1 Score