BioLaySum: More Readable Summarization for Biomedical Research Paper

Group #27: Xincheng Cai, Mengmeng Du

Introduction

Problem Identification

Background - Biomedical publications have lots of highly technical and specialist language, difficult to understand for the non-expert audiences

Goal - Create lay summaries with simplified language and added background information accessible to non-experts.

Dataset

eLife journal : 4587

- Train dataset: 4346
- validation dataset: 241

biochemistry and chemical biology	microbiology and infectious disease
cell biology	neuroscience
developmental biology	structural biology and molecular biophysics

Evaluation

Relevance - The similarity between the generated text and the reference text

Readability - text's reading difficulty based on word length and sentence length

Factuality - The degree of alignment between the generated text content and the source information

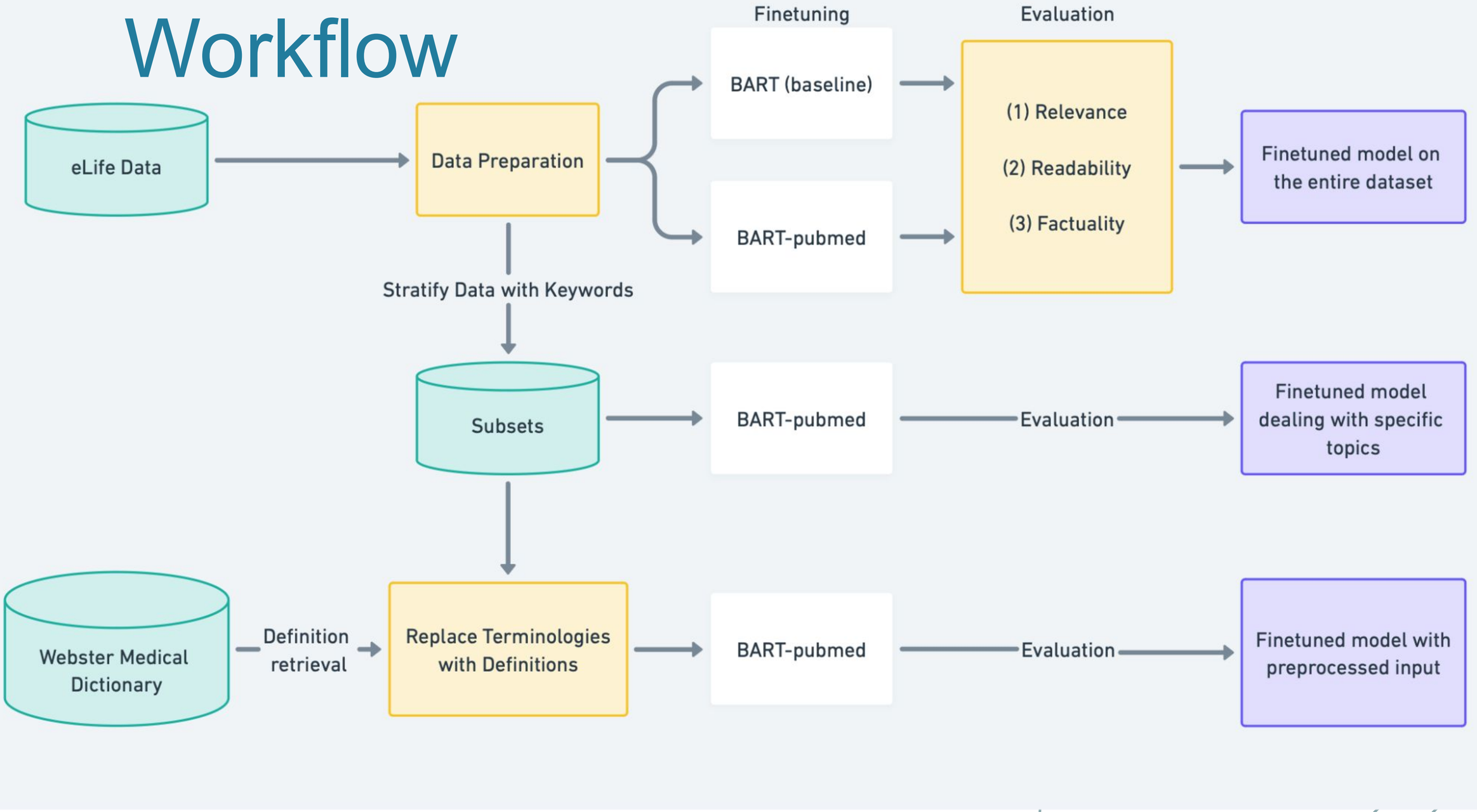
Relevance	ROUGE (1, 2, and L) BERTScore
Readability	FKGL and DCRS, CLI
Factuality	SummaC

Summary Model

Bart

Bart-pubmed

Workflow


Finetuning
Evaluation
eLife Data
Data Preparation
BART (baseline)
BART-pubmed
(1) Relevance
(2) Readability
(3) Factuality
Finetuned model on the entire dataset
Stratify Data with Keywords
Subsets
BART-pubmed
Evaluation
Finetuned model dealing with specific topics
Webster Medical Dictionary
Definition retrieval
Replace Terminologies with Definitions
BART-pubmed
Evaluation
Finetuned model with preprocessed input

Results

BART vs BART-pubmed

	BART	BART-pubmed
ROUGE-1	57.821059	54.922606
ROUGE-2	22.440689	19.132091
ROUGE-L	54.160794	51.062519
BERT-score	86.933277	86.070874
FKGL	9.169295	8.707054
CLI	10.619585	10.129627
DCRS	8.974730	8.711784

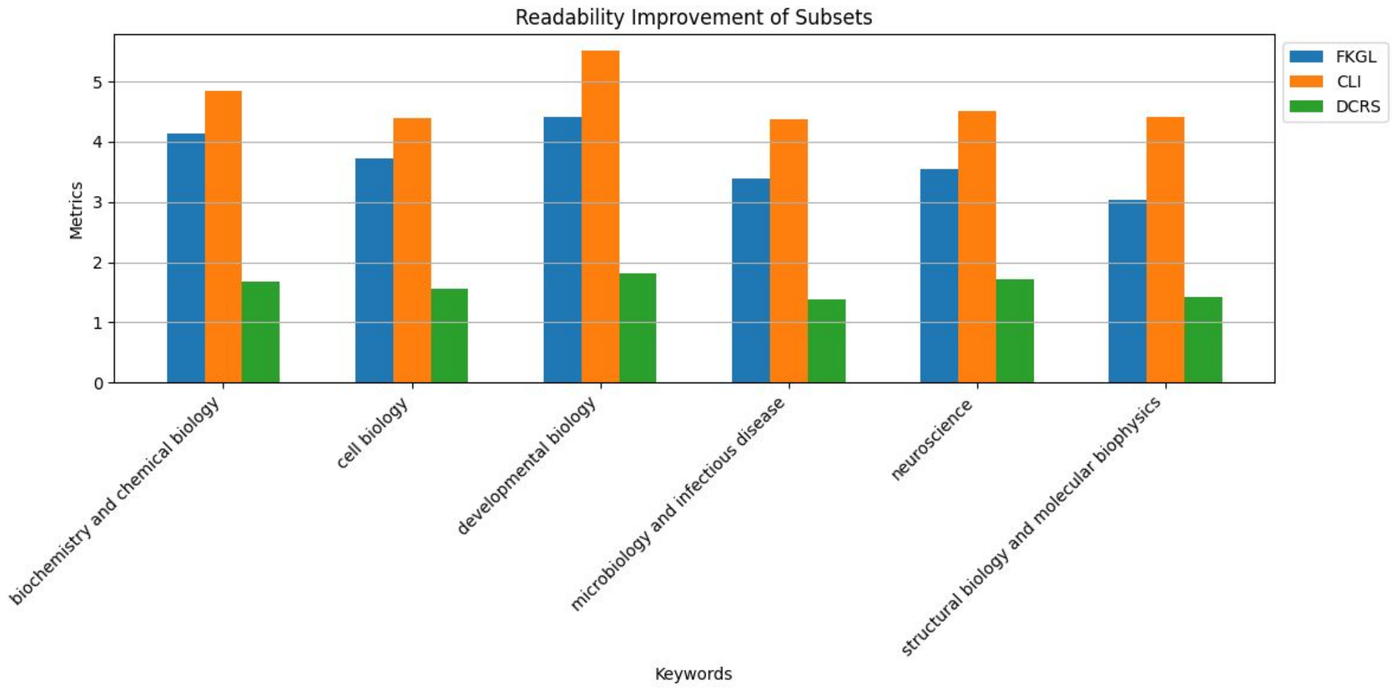
BART-pubmed on subsets

	rouge1	rouge2	rougeL	bert_score
biochemistry and chemical biology	0.526194	0.165373	0.491461	0.855682
cell biology	0.542733	0.179915	0.503583	0.861299
developmental biology	0.543000	0.188836	0.503814	0.860833
microbiology and infectious disease	0.544249	0.188292	0.506255	0.860753
neuroscience	0.553964	0.198521	0.515706	0.860500
structural biology and molecular biophysics	0.559834	0.193625	0.517510	0.860000

	avg_fkgl	avg_cli	avg_dcrs	summac_score
biochemistry and chemical biology	7.886207	9.607586	8.868966	0.417491
cell biology	8.863043	10.472826	8.824783	0.417462
developmental biology	8.233333	9.323810	8.530000	0.450472
microbiology and infectious disease	8.761538	10.846154	8.761538	0.395190
neuroscience	8.657971	10.360000	8.503043	0.431311
structural biology and molecular biophysics	9.492593	10.130000	9.184815	0.395497

Readability Improvement

Improvement = − (readability_output − readability_input)



Replace Terms with Definition vs Non-

Structural biology and molecular biophysics

	w/o_replacement	with_replacement
rouge1	0.538049	0.588711
rouge2	0.162231	0.200839
rougeL	0.495753	0.566689
bert_score	0.854332	0.841444
avg_fkgl	8.666667	8.137037
avg_cli	9.924074	9.618148
avg_dcrs	8.851111	8.070741
summac_score	0.414035	0.409548

Significance

- More readable summaries
- Facilitates knowledge translation, effective communication

Future Directions

- Apply the data preprocessing terminology replacement on all the subsets.
- Use the Longformer Encoder Decoder (LED) model to address the issue of limited input capacity.

THANK YOU