README for PLABA

Kush Attal

Brian Ondov

Dina Demner-Fushman

Lister Hill National Center for Biomedical Communications, U.S. National Library of Medicine, National Institutes of Health, Bethesda, MD.

Contact information: ddemner@mail.nih.gov

Contents

Introduction

This is the README for the Plain Language Adaptation of Biomedical Abstracts (PLABA) dataset presented in the paper "A Dataset for Plain Language Adaptation of Biomedical Abstracts". The dataset can be used for evaluation of text adaptation systems. Users are able to use this dataset to evaluate sentence-level or document-level text adaptation of answers to consumer health questions, with or without taking a question-driven approach to the problem. Users are encouraged to read the paper for further details regarding the methodology to create the dataset.

Included in this README are technical details for the dataset provided in the Open Science Framework archive (https://osf.io/rnpmf/), the dataset's description and JSON structure, statistics about the data, and usage notes for when using PLABA for evaluating text adaptation.

Datasets

The dataset included in this repository is in .json format. The filename is 'data.json' and contains all the questions, abstracts corresponding to each question ordered by PubMed ID (PMID), and their corresponding, human-generated adaptations.

Here are the details regarding the key: value pairs in the dataset.

The full PLABA collection. A sample of the data format, for one record, is shown below:

Data format:

{

<Question ID>:

{

"question": "What is question-driven answer adaptation?"

"question_type": "C"

<Answer1 PMID>:

{

"Title": "Defining question-driven answer adaptation"

"abstract":

{

<Sentence_1>: "This is the first sentence of the abstract."

<Sentence_2>: "This is the second sentence of the abstract."

<Sentence_3>: "This is the third sentence of the abstract."

<Sentence_4>: "This is the fourth sentence of the abstract."

…

},

“adaptations”:

{

“adaptation1”:

{

<Sentence_1>: “This is the first sentence of the adaptation.”

<Sentence_2>: “This is the second sentence of the adaptation.”

<Sentence_3>: “This is the third sentence of the adaptation.”

<Sentence_4>: “This is the fourth sentence of the adaptation.”

…

},

…

}

…

}

…

}

The key: value pairs are described here:

Key	Value
<Question ID> (The unique string identifier of each question)	A nested JSON object with each answer to the question
"question"	The question string
"question_type"	String that indicates if question is a clinical question (C) or biological question (B)
<Answer# ID> (The unique PubMed string identifier of each answer’s abstract)	A nested JSON object containing the data (i.e., PubMed abstract and human adaptation(s)) for the specific answer
"Title"	The string of the PubMed abstract’s title
"abstract”	A nested JSON object containing the PubMed abstract separated by sentences
<Sentence_#>	The sentence string corresponding to the #th sentence of the abstract/adaptation
"adaptation1”	A nested JSON object containing the first human adaptation separated by sentences

PLABA Statistics

This section contains the statistics regarding the documents in the collection. The table here is also presented in the paper.

Data Type	Count	Words		Sentences	
		Average	S.d.	Average	S.d.
Questions	75	10	6	1	0
Abstracts	750	240	95	10	4
Adaptations	921	244	95	12	5

Usage

Code with examples of pre-processing the dataset and benchmarking text adaptation models using PLABA can be found at https://github.com/attal-kush/PLABA. We recommend that interested users first read the paper that presents this dataset, “A Dataset for Plain Language Adaptation of Biomedical Abstracts,” to gain a better idea of how to use PLABA for evaluation.