Project 3

Background

Thanks to technological advances, animal geneticists have an ever-expanding tool chest with which to study the inheritance of traits in livestock in order to improve production. With the rise of high-throughput technology, vast amounts of genotype/phenotype data are being rapidly generated. Animal QTLdb and other genotype/phenotype databases would greatly benefit from automated and expedited curation tools.

In this project, we want to write the title for each abstract.

Data

You will use the same data from project 2: QTL_text.json and QTL_test_unlabeled.tsv

QTL_text.json is in the following format (all fields are strings)

{"PMID":
"Journal":
"Title":
"Abstract":
"Category":}

Each entry in this json file corresponds to a paper

1. "Title": this is the title of the paper
2. "Abstract": this is the abstract of the paper. Usually it contains many sentences.
3. "Category": this is either '0' or '1'. '0' means the paper is not related to animal QTLdb curation, and '1' means the paper is related to animal QTLdb curation. In this project, this is the label.

QTL_test_unlabeled.tsv is a tab-separated values (TSV) file where each row represents a research paper. It consists of four columns with the following data types:

1. PMID (int): A unique numerical identifier assigned to each publication in the PubMed database.
2. Title (string): The title of the research paper
3. Abstract (string): this is the abstract of the paper. Usually it contains many sentences.
4. Label (int): An integer value, which is currently unlabeled (e.g., a placeholder 0 for future classification). 0 means the paper is not related to animal QTLdb curation, and 1 means the paper is related to animal QTLdb curation.

Each field is separated by a tab (\t).

Pre-processing and other requirements

You will need to pre-process the QTL_text.json to the proper format and split it into a training and development set if needed. You can use any existing libraries/implementations for this project. If you use a github repo, please add a readme file in your submission so that we can run your code.

Task

Train/prompt a generative model to predict the "title" of each paper in QTL_test_unlabeled.tsv using abstract. You can use QTL_text.json as the training dataset. Note that you cannot use PMID field in training or testing.

Submission

You will also need to submit the following documents on Canvas

1. your code (optional: readme)
2. a report (max 2 pages) in pdf using this template: https://www.overleaf.com/latex/templates/association-for-computational-linguistics-acl-conference/jvxskxpnznfj

Your report needs to include:

Title: project 3

Author: your name, email

Method: describe the method you used.

Result:

You will need to report the BLEU score, ROUGE-2, and ROUGE-L of your generated titles comparing to the ground truth titles as reference text.

BLEU: https://huggingface.co/spaces/evaluate-metric/bleu

ROUGE: https://huggingface.co/spaces/evaluate-metric/rouge

Discussion (optional): If you trained several models, you can discuss your findings.

Appendix (optional, does not count into the 2-page limit): Include links to all tutorials/websites/github that you referred to for this project. If you used AI assistant for coding, please include screenshots.