Social Computing Assignment 2
Intent Recognition and Entity Extraction from Healthcare Platform Data

Deadline: November 5, 2023 (end of day)

For any query about the assignment, you can contact TA Sourjyadip Ray (sourjyadipray@gmail.com).

Task: Build machine learning models for Intent Recognition and Entity Extraction from English and Indic language queries from Healthcare Platforms.

Reference paper: https://arxiv.org/pdf/2302.09685.pdf
Dataset: https://github.com/indichealth/indic-health-demo

Part 1: Intent Recognition

Task definition: Given a query from an online healthcare platform, identify the intent class the query belongs to. This is a multi-label classification task.

a) Intent Recognition in English Queries

Example:
Query: What are the instructions for storage and disposal of Janumet XR CP Tablet?
Tagset: <drug, disease, treatment, other>
Intent: drug

Models to be used:

- SVM https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html
- RoBERTa https://huggingface.co/roberta-base
- BioClinicalBERT https://huggingface.co/emilyalsentzer/Bio_ClinicalBERT

b) Intent Recognition in Indic Queries (Hindi)

Example:
Query: क्या कान के संक्रमण से वयस्कों के कान में दर्द हो सकता है?
Tagset: <drug, disease, treatment, other>
Intent: disease

Models to be used:

- XLMR https://huggingface.co/xlm-roberta-base
- Direct back translation to english and Bioclinical BERT
- Bridge language back translation and BioClinical BERT (Use bengali data for this, and use hindi as a bridge language)

Part 2: Entity Recognition

Task definition: Given a query from an online healthcare platform, predict the entity classes using the BIO tagging format.

More about BIO tagging:
https://medium.com/analytics-vidhya/bio-tagged-text-to-original-text-99b05da6664

a) Entity Recognition in English Queries

Example:
Query: What should I know about the storage and disposal of Neurobion RF Forte Injection?
Entity tagset: <drug, disease, treatment plan>
Entity:
Drug: Neurobion RF Forte Injection
BIO tags: O O O O O O O O O O B-drug I-drug I-drug I-drug

Models to be used:

- SVM
- RoBERTa
- BioClinicalBERT

b) Entity Recognition in Indic Queries (Hindi)

Example:
Query: डेंगू का पता लगाने के लिए कौन सा ब्लड टेस्ट करवाना चाहिए?
Entity tagset: <drug, disease, treatment plan>
Entity:
Disease: डेंगू
Treatment plan: ब्लड टेस्ट
BIO tags: B-disease O O O O O O B-treatment I-treatment O O

Models/methods to be used:

- XLMR https://huggingface.co/xlm-roberta-base
- Direct back translation to english and Bioclinical BERT
- Bridge language back translation and BioClinical BERT (Use bengali data for this, and use hindi as a bridge language)

Deliverables:

- Python notebooks for each method OR python files (3 methods per each language (english + hindi) for Intent Recognition and Entity Recognition = 12 models)
- README file containing instructions on how to run the code.
- Project Report containing Results and Analysis section.

General Guidelines:

- Report the results for both the IHQID-WebMD and IHQID-1mg datasets as in the reference paper.
- While submitting the code, only submit it for any 1 of the datasets for each method.
- The entity dataset has to be converted to the BIO tagging format. Marks will be awarded for this step.
- For the back translation methods for the Indic queries, any translation API such as Google Translate API can be used. Suggestion: https://pypi.org/project/deep-translator/
- Only the macro F1 metric needs to be reported in the results section in the project report. The report should also contain classification report (https://scikit-learn.org/stable/modules/generated/sklearn.metrics.classification_report.html) and confusion matrix (https://stackoverflow.com/questions/35572000/how-can-i-plot-a-confusion-matrix) of each method as part of analysis.
- You are allowed to use either pytorch or tensorflow frameworks.
- The model files need not be submitted.
- Please follow the following program format:
    1. Preprocessing (Data parsing and formatting)
    2. Model Creation
    3. Model Training (using train set)
    4. Model Evaluation on the Test set

Submission Guidelines:

- The code files should be in the format [task]_[method]_[lang].py (or ipynb).
  The short forms for file naming are as follows:
    - Task: Intent Recognition (ir), Entity Extraction (ee)
    - Method: SVM (svm), RoBERTa (rob), BioClinicalBERT (bcbert), XLMR (xlmr), Direct Translation (dtrans), Bridge Language (bridge)
    - Lang: English (en), Hindi (hi)

  Example filename: ir_rob_en.ipynb
- The running instructions for all codes need to be provided in a SINGLE README file.
- The Project Report only needs to have the following sections:
    - Results: All results in a task-wise tabular format (see reference paper).
    - Analysis: Classification report + Confusion matrix for each method.
- The naming format for the report should be the following [Roll number]_Report.pdf (Example: 22CS71P04_Report.pdf).

- The files need to be zipped and submitted with the following naming format [Roll number]_[Name]_Assignment.zip (Eg: 22CS71P04_SourjyadipRay.zip).

IMPORTANT: PLEASE FOLLOW THE NAMING CONVENTIONS STRICTLY. FAILURE TO DO SO WILL RESULT IN DEDUCTION OF MARKS.

Plagiarism: We will be employing strict plagiarism checking. If your code matches with another student's code, all those students will be awarded zero marks for the assignment. Therefore, please ensure there is no sharing of code.

Code Error: If the code doesn't run for a particular experiment, partial marks may be awarded based on structure and logic of code. If required, you may be contacted by the TAs to explain your code.

Additional Resources:

1. Support Vector Machines https://scikit-learn.org/stable/modules/svm.html
2. Using scikit learn for intent classification https://www.kaggle.com/code/hassanamin/atis-intent-classification-using-svm-and-spacy
3. Transformers https://arxiv.org/pdf/1706.03762.pdf
4. BERT https://arxiv.org/pdf/1810.04805.pdf
5. RoBERTa https://openreview.net/attachment?id=SyxS0T4tvS&name=original_pdf
6. XLM-R https://arxiv.org/pdf/1911.02116.pdf
7. Fine Tuning BERT for classification using transformers library https://luv-bansal.medium.com/fine-tuning-bert-for-text-classification-in-pytorch-503d97342db2
8. Jay Alammar Transformer blog http://jalammar.github.io/illustrated-transformer/
9. Jay Alammar BERT blog http://jalammar.github.io/illustrated-bert/
10. Jay Alammar Finetuning BERT http://jalammar.github.io/a-visual-guide-to-using-bert-for-the-first-time/

For any query about the assignment, you can contact TA Sourjyadip Ray (sourjyadipray@gmail.com).