Introduction to Natural Language Processing

Final Project

TA : Guan-Yi Ho

joy861106.cs11@nycu.edu.tw

Task introduction

Fact checking with neural network.

Sometimes someone says something. We will consider that it ture or not, and find some evidence to support our opinion. If we sent some premise articles about claims and claimants into neural network, it may help us doing fact checking.

Here’re some examples of claims and it’s rating:

- OpIndia claimed Greta Thunberg's real name is Ghazala bhat. -> false
- Goldman Sachs, where Heidi Cruz is managing director, oversees Texas utilities. -> partial ture
- A woman found human ashes in a $2 blue heart ornament. -> true

Take the third claim as example. Here we found three links of newspaper or youtube about this claim as premise articles.


```
"premise_articles": {
  "https://boston.cbslocal.com/2018/01/11/plymouth-woman-finds-ashes-in-decoration-bought-at-savers/": "221_1.json",
  "https://www.youtube.com/watch?v=DuxWAQQDYk0": "221_2.json",
  "https://www.bostonglobe.com/metro/2018/01/11/plymouth-woman-claims-she-found-ashes-inside-trinket-savers/OXPB7nrVb7SWKxP5pf17OO/story.html": "221_3.json"
}
```


Use the sentence of these articles as evidence to prove that the third claim is true.

Implementation method

1. Evidence sentence extration:
    a. using technique of texting mining and information retrieval to vectorize the given premise articles into evidence vector for training.
    b. keywords you can search: Tokenization, Bag of Words, TF-IDF, DPR, Vectorizer, etc.
2. Claim Veracity Inference:
    a. Use any sequence model training with above vectors to verify claims.
    b. Target is passing the baseline that TA set.

Dataset

Link: Dataset (You can also download from Kaggle)

Dataset description is shown as right, and you can also find it on Kaggle.

Please use the method write in previous page to classify the claim into False/ Partial True/ True.

Files

- **train.json** - the training set
- **valid.json** - the valid set
- **test.json** - the test set
- **articles** - the premise articles provide

jsons

- `metadata` - definition of articles
  - `claimant` - the person who claims.
  - `claim` - the claim
  - `id` - claim id
  - `premise_articles` - {} of `url`, `json_file provide`
- `label` - definition of feature
  - `rating` - false/partial/ture of claim(0~2)
  - `original_rating` - original sentence to clearify rating
  - `id` - claim id

Kaggle competition (50%)

Kaggle link: INLP final Kaggle

Display team name: <student ID>

Each group one submission is enough.

Submission format:

- A 2361 X 2 .csv file, first row is for the column name and the last 2360 rows for your result.
- Column name should be id and rating.
- Result should be the id and rating(0~2),please make sure the order of your result is right!

You can submit at most 5 times each day.

No hint and bonus this time, just try to do your best before deadline.

The scoring metric will be macro F1, not accuracy!

Kaggle competition (50%)

Scoring criterion:

- Baseline(30%)
  - Pass the basic baseline(setting by guess all 0), you can get 10 points.
  - Pass the standard baseline(setting by TA’s simple model), you can get 20 points.
- Ranking(20%)
  - Determined by ranking with other teams on Kaggle.
  - The first one gets 20 points, the second one gets 18 points, and so on.(Sequentially -2pt.)

Report Submission(50%)

Submit a report contains 4 questions:

1. Briefly explain your choice to do evidence sentence extraction and compare the results of choosing different parameters(if has) in your method.(15pt.)
2. Describe how you implement your sequence model, including your choice of packages, model architectures, loss functions, hyperparameters (learning rate, epochs, etc.)etc.(15pt.)
3. All of method you have tried, and compare with your best method.(10pt.)
4. Do error analysis. You can use confusion matrix to illustrate the whole model performance, or try to analyze the dataset's problem like unbalanced number of labels could lead what difficulty on training and how to overcome it. Share anything you observe. (10pt.)

Please answer the questions in detail.

Report Submission(50%)

Please provide your commands to run your code in the last part of report.

TA will follow this to run your code, so please briefly explain it as possible.

If you don’t write this, you will get -5 points.

You don’t need to contain the dataset in your E3 submission, because the size is too large and TA will use same dataset to test your code.

So, please also provide your location of dataset in your code for TA to put dataset in the last part of report.

If you don’t write this, you will get -5 points.

E3 Submission

- Deadline:
    - Submit Zip to E3 before 1/13 11:59PM
    - No Late Submission!
- Format:
    - Source code : put them under the file named Final_<group ID>
    - Report file : Final_<group ID>.pdf
    - Zip file : Final_<group ID>.zip

Format Error: -5 points.

Thanks

If you have any question about final project, please feel free to contact with TA: GUAN-YI,HO through email joy861106.cs11@nycu.edu.tw or E3.