Paraphrase Identification by Multi-layer Perceptron

Revised 11/03/2022

Task:

given two sentences, build a multi-layer perceptron model to classify whether they express the same meaning or not. (note “multi-layer perceptron” is also called “fully-connected network” in some materials)

Data and format:

Please download the data from GitHub: https://github.com/yinwenpeng/MLFinalProject

- train_with_label.txt (7801 instances: pos/neg=0.32)
- dev_with_label.txt (4K instances; 1K positive, 3K negative)
- test_without_label.txt (4K instances; 1K positive, 3K negative)

Each row in the three files denotes one instance, i.e., the two sentences.
“train_with_label.txt” and “dev_with_label.txt” have four columns: instance_id, sentence_1, sentence_2, gold_label (0 or 1). Columns are separated by \tab.
“test_without_label.txt” has the same column order except that it has no the “gold_label” column

The required algorithms to use:

- Only multi-layer perceptron can be used (the final layer is often a logistic regression).
- Convolutional NN, recurrent NN, and Transformers are not allowed.
- No pretrained word representations should be used, such as word2vec, GloVe, etc.

The knowledge/skills this project looks for from you:

1) Implement a basic deep learning system
2) Know how to tune the deep learning system, such as learning rate, batch size, etc.

Timeline:

- Starting date: 11/03/2022
- Submission deadline: 12/06/2022, 11:59pm (EST)

What and how to submit:

You are expected to submit **three** things:

1) “**YourFullName_test_result.txt**”: Use your best model on “dev_with_label.txt” to test on the “test_without_label.txt” file and generate a new file named

“YourFullName_test_result.txt” with two columns separated by \tab: instance_id, predicted_label (0 or 1)
Each row in “YourFullName_test_result.txt” corresponds to one test instance.

2) “**YourFullName_system_description.pdf**”: A pdf file that describes what you did
3) **Code**: Commit the code to your github and let me know the repository URL

Commit “**YourFullName_test_result.txt**” and “**YourFullName_system_description.pdf**” to the same github repository with the code and send the **repository URL** to wenpeng.yin@temple.edu by the deadline.

How will your submission be scored?

The system will be evaluated by **F1** on the test file (the test file with gold labels is not given to you). Your score will be

$100\times\frac{your\ F1}{max\ F1\ in\ the\ class}$

Please note

- The accuracy of a random guess baseline in this task is 50%, so a system with an accuracy lower than 50% will be scored 0;
- You will only get the score if all three required submissions are available. A result file with no system description or code URL will be scored 0.

Some hints:

1) Pytorch has default multi-layer perceptron functions; you can call them or reimplement your own multi-layer perceptron layers
2) You can decide how many layers you want and how wide each layer is.
3) You can make use of the features you defined in the midterm project