Natural Language

MP1

The greatest evaluation forum on Question Classification

Técnico, Alameda and Tagus,

2020

With this project we will simulate an evaluation forum. International evaluation forums are competitions in which participants test their systems in specific tasks and in the same conditions. Training/development sets are given in advance and, in a certain predefined date, a test set is released. Then, participants have a short period of time to return the output of their systems, which are straightforwardly compared, allowing to identify the state-of-the-art ones.

Goal

Build two models that classify questions according with a coarse and a fine-grained questions' taxonomy, from Li and Roth (see Table below).

Coarse	**Fine**
ABBREVIATION	abbreviation, expansion
DESCRIPTION	definition, description, manner, reason
ENTITY	animal, body, color, creative, currency, medical disease, event, food, instrument, language, letter, other, plant, product, religion, sport, substance, symbol, technique, term, vehicle, word
HUMAN	description, group, individual, title
LOCATION	city, country, mountain, other, state
NUMERIC	code, count, date, distance, money, order, other, percent, period, speed, temperature, size, weight

As an example, given the following questions:

What fowl grabs the spotlight after the Chinese Year of the Monkey?
What is the full form of .com?
What contemptible scoundrel stole the cork from my lunch?

Considering the coarse taxonomy (*COARSE* labels), your system should return:
ENTY
ABBR
HUM

Considering the fine taxonomy (*COARSE:fine* labels), your system should return:
ENTY:animal
ABBR:exp
HUM:ind

How

This project should be done in groups of 1 or 2.

As in an evaluation forum, you will be given "training" and development sets (TRAIN.txt and DEV.txt). The final test set (just questions) will be released in a specific date (see below).

Considering that TRAIN.txt and DEV.txt contain labels+questions, you should start by generating two files from DEV.txt:
a) DEV-questions.txt that includes questions only (by removing the labels)
b) DEV-labels.txt that contains only the labels.

During the development stage, you can assess the performance of your models as follows:

a) Predict the labels for the data.

```
python qc.py -coarse TRAIN.txt DEV-questions.txt > predicted-labels.txt # runs the coarse model
python qc.py -fine TRAIN.txt DEV-questions.txt > predicted-labels.txt # runs the fine model
```

b) Evaluate the performance of your model using the true labels:

```
python ./evaluate.py DEV-labels.txt predicted-labels.txt
```

You can use any technique that you have learned in LN (except deep learning or neural word embeddings). You should implement your model in Python3. You can use, for instance, measures already available (including the source code), as long as you identify the source. You are also free to create your own measures. You can use NLTK, NumPy and scikit-learn.

Evaluation

(1) Report (3 values):
NUM.pdf[1] with a maximum of 1 page, containing the following parts:

1. A description of your model (for instance, the evaluation metrics you have tested or the pre-processing you have done)
2. Accuracy resulting from evaluating your model in the development set
3. Short error analysis
4. Bibliography (if applicable)

(2) Automatic Evaluation (17 values):

When you receive the test file (TEST.txt), containing ONLY the questions (and not the labels), you should run your models (previously delivered – see below; ****no changes in your code should be done****) in the test set and save the predicted labels in two files:

- testNUM-coarse.txt should contain the labels of the coarse category. For instance:

ENTY
ABBR

[1] From now on, NUM is the number of the group.

HUM

- testNUM-fine.txt should contain the labels of the fine category. For instance:

ENTY:animal
ABBR:exp
HUM:ind

Notice that the position in which the question appears in the test file corresponds to the position of its label in the output file, and this is how it will be evaluated. A*ccuracy* will be the evaluation measure. Notice also that ****no manual editions should be done to these files****.

Submission (ATTENTION, PLEASE!!!)

Part 1 – on the 26/10/2020, **before 15h 30 (3:30 PM)**, you should deliver, via Fénix (MP1-Part1), a zip file (NOT a rar) containing the project, named after the group number (ex: 3.zip).

– the zip file should contain:
  - the file **NUM.pdf** with the report;
  - a file **qc.py** with the project code (you can have extra python files);
  - a file **developNUM-coarse.txt** with the results from the coarse development set, that is, a list of the coarse labels returned by your system when it ran on the development set);
  - a file **developNUM-fine.txt** with the results from the fine development set, that is, a list of the fine labels returned by your system when it ran on the development set.

The test set, TEST.txt, will be released between 15h 30 and 15h 35 (3:30PM – 3:35PM).

Part 2 – on the 26/10/2020, **between 15h 35 and 23h 59 (3:35 PM – 11:59 PM)**, you should deliver, also via Fénix (MP1-Part2):

  - a file **testNUM-coarse.txt** with the results from the coarse test set (a list of the coarse labels returned by your previously submitted system when it ran on the test set);
  - a file **testNUM-fine.txt** with the results from the fine test set (a list of the fine labels returned by your previously submitted system when it ran on the test set).

Note:

- 4 values will be taken if the names of the files are not correct;
- We will randomly select a set of projects (maybe all) and we will run them (previous commands) in the test sets. If any difference in results is found, the group will have a 0 in the project.

We will release FAQs about the project. Please, check them before sending questions to meic-ln@disciplinas.tecnico.ulisboa.pt (subject: MP1). Thank you!!