**Human Language Engineering (HLE). Laboratory cases. 2019-2020**

This labo case covers some of the functionalities of BioAsQ challenge task 6b

(http://bioasq.org/participate/challenges_year_6)

We will provide a dataset of 2,252 questions annotated with their type. We will consider four types (Question Type, QT):

- Factoid
- List
- Summary
- Yes/no

See a couple of examples:

| | |
|---|---|
| Is there any role of TBR1 in autism? | Yes/no |
| What is the role of the MCM2-7 complex? | Summary |

You should split this dataset into learning and test datasets.

**Our first task** is to write a multiclass classifier for classifying the questions into one of the 4 QT. The program has to be written in Python. No restrictions about the classifier are set. You are free to use linear classifiers (using scikit learn python library, http://scikit-learn.org/stable/) or neural models (using Keras, https://keras.io/, TensorFlow, https://www.tensorflow.org/, or Pytorch, https://pytorch.org/).

***Comparing** two or more classifiers would have a bonus.*

For parsing you could use Stanford CoreNLP tools.

Download from https://stanfordnlp.github.io/CoreNLP/

From NLTK you can also access Stanford parser because there is a wrapper.

https://www.nltk.org/

Other options: using the parsers of NLTK (maybe too much simple), using python and freeling tool suite (installation of freeling required).

http://nlp.lsi.upc.edu/freeling/index.php/node/1

Another useful tool would be NCBO annotator. You need to register to BioPortal

https://www.bioontology.org/

Probably, due to the small size of the training dataset, the results obtained are not as good as expected.

For trying to alleviate this issue, the **second task** consists on extending the training dataset automatically using a kind of co-training.

From https://data.quora.com/First-Quora-Dataset-Release-Question-Pairs you can download a huge dataset of Question Pairs. The dataset contains 404,302 pairs. Each pair contains two questions assumed to be roughly semantically equivalent (paraphrase of each other).  An example of pair is the following:

- What is the step by step guide to invest in share market in india?
- What is the step by step guide to invest in share market?

The procedure for getting more examples would be the following:

1) Apply the classifier from your first task to all the pairs of questions in the quora dataset. Collect all the pairs whose two questions have been classified into the same class.
2) Consider the two questions of these pairs as examples tagged with the corresponding class.
3) Add the new set of examples to the original dataset and repeat the learning process.

Many variations of this schema can be tried:

- Repeat the process iteratively for increasing the global accuracy of the extended dataset.
- Perform some filtering process over the pairs of the quora dataset in order to get questions close to the original ones (for instance, filter out pairs not belonging to the medical/biological domain).

The HLE lab work can be done individually or in pairs.

**Calendar**:

| | |
|---|---|
| 7 November | Preliminary deliverable and short presentation. |
| 7 January | Final deliverable (report and slides of presentation) |
| 9 & 16 January | Students presentations |

This lab work could be the first step for a master thesis to be developed in the frame of GRAPH-MED project:

http://www.talp.upc.edu/project-detail/497/GRAPH-MED%20

This is the UPC subproject of a coordinated project (Prosa-med).

For research projects that could offer a master thesis proposal in the area of Human Language Engineering you could check TALP (Language and Speech Techonologies) research center:

http://www.talp.upc.edu/