Question Classification based on LSTMs

Introduction

Question Answering (QA) systems enable users to retrieve exact answers for questions posed in natural language. The topic of Question Classification arises in the area of automated question-answering systems.

Question Classification (QC) is the task that, given a question, maps it to one of k classes, which provide a semantic constraint on the sought-after answer.

Dataset

This dataset contains 5500 labelled questions in the training set and another 500 for the test set. The dataset has six labels, including Abbreviation (**ABBR**), Entity (**ENTY**), Description (**DESC**), Human (**HUM**), Location (**LOC**) and Numeric (**NUM**). More information on the labels in the dataset is presented in [1]. Average length of each sentence is ten, with a vocabulary size of 8700. The training and test **txt** files can be downloaded from [2] and [3], respectively. The dataset is composed of two attributes (text, label).

Mission

The final objective is to build a text classifier that can distinguish questions in different categories. The output of the system can be an input to the other modules in a question answering system.

To do that, some tasks need to be accomplished:

1. Understand the content that is available in the dataset .
2. Pre-process the textual data to improve the final results.
3. LSTM model
    a. Build a LSTM (Long Short-term Memory) text classifier to classify questions into one of the categories.
    b. Analyze the outcomes.
4. SMV model
    a. Do the same categorization using a Support Vector Machine (SVM) based approach.
    b. Test different lengths of N-Grams and report the best N in which the classifier achieved the best performance
    c. Report the top 10 representative (most repeated) 1, 2 and 3-grams for each of the classes.
    d. Analyze the outcomes.
5. Use the different packages of visualization explained during the course to visualize findings from both approaches.
6. Compare the results from two models, using the F1 measure.
7. Obtain statistics that can corroborate the results achieved and explain the reason of these choices.

Deliverables

To carry out the assessment of the project, the group has to submit the following:

- A report using Google Doc and explaining the concept about the project solution and the expected division of the tasks regarding the components of the group. This document should not be longer than 10 pages (including cover, table of contents, etc).
- Collaborative work using Git with commit + push changes on a daily basis.
- A presentation explaining the thought process, your approaches and the reason for this choice, your findings and the real task division in your group at the end of the project.Every group has 15 minutes per presentation and there will be 5 minutes of questions.

[1] https://cogcomp.seas.upenn.edu/Data/QA/QC/definition.html
[2] https://cogcomp.seas.upenn.edu/Data/QA/QC/train_5500.label
[3] https://cogcomp.seas.upenn.edu/Data/QA/QC/TREC_10.label