Question Classification

Machine Learning for Language Processing Project 2018/2019

1 Project Description

The task of question classification (QC) is to predict the entity type of a question which is written in natural language.

In this project we would like to implement a machine learning model for a Question Classification task. The model will be trained and evaluated using the TREC dataset. TREC dataset is a collection of an annotated questions into 6 classes :

1. ABBREVIATION → abbreviation
2. ENTITY → entities
3. DESCRIPTION → description and abstract concepts
4. HUMAN → human beings
5. LOCATION → locations
6. NUMERIC → numeric values

2 Data Description

TREC dataset is a collection of 6000 labeled questions. It consists of two separate set of 5500 and 500 questions in which the first is used as training set and the second is used as an independent test set. This dataset was first published in University of Illinois Urbana-Champaign (uiuc) usually referred as the UIUC dataset and sometimes referred as the TREC dataset since it is widely use in the Text REtrieval Conference (TREC).

Sentence	Class
What team did baseball 's St. Louis Browns become ?	HUMAN
What are liver enzymes ?	DESCRIPTION
When was Ozzy Osbourne born ?	NUMERIC
Who was The Pride of the Yankees ?	HUMAN
What sprawling U.S. state boasts the most airports ?	LOCATION
What is an annotated bibliography ?	DESCRIPTION

TABLE 1 – Example of sentences with their corresponding classes from the training set

Data are available at :

train : `http://perso.univ-lemans.fr/~fbouga/train_all.label`
test : `http://perso.univ-lemans.fr/~fbouga/TREC_test.label`

3 Evaluation

Systems are evaluated on classification accuracy (the percent of labels that are predicted correctly) for every parsed phrase. We Would like to have also the precision/recall scores for each class.

4 Project Roadmap

1. Preprocess and prepare the training data
2. Train and evaluate a vanilla deep recurrent neural network (RNN)
3. Use the Pytorch framework to train the RNN network.
4. Optimize the model and propose enhancement (Regularization, network init, new architecture)
5. Prepare the final defense
6. Present your model and the obtained results

5 Reference

Xin Li, Dan Roth, Learning Question Classifiers. COLING'02, Aug., 2002.