Question Classification based on LSTMs

Michal Harakal, Petra Woßeng, Andreas Neutze, Sven Wontroba Project Presentation - Draft - 04.12.2020

Our Task

Question Classification: Our Task

Given train and test data from a Question Answering System, the main task is to predict the category the questions refers to.

After an Exploratory Data Analysis (EDA) we should use Neural Networks/LSTM and alternatively a Support Vector Machine.



DESC:manner How did serfdom develop in and then leave Russia?

ENTY:cremat What films featured the character Popeye Doyle?

DESC:manner How can I find a list of celebrities ' real names? ENTY:animal What fowl grabs the spotlight after the Chinese Year of the Monkey?

ABBR:exp What is the full form of .com?

HUM:ind What contemptible scoundrel stole the cork from my lunch?

HUM: or What team did baseball 's St. Louis Browns become?

HUM:title What is the oldest profession?

DESC:def What are liver enzymes?

HUM:ind Name the scar-faced bounty hunter of The Old West .

NUM:date When was Ozzy Osbourne born?

DESC:reason Why do heavier objects travel downhill faster?

HUM ind Who was The Pride of the Yankees?

HUM ind Who killed Gandhi?

ENTY:event What is considered the costliest disaster the insurance industry has ever faced?

LOC:state What sprawling U.S. state boasts the most airports?

DESC:desc What did the only repealed amendment to the U.S. Constitution deal with?

NUM:count How many Jews were executed in concentration camps during WWII?

DESC:def What is "Nine Inch Nails"?

DESC:def What is an annotated bibliography?

NUM:date What is the date of Boxing Day?

ENTY:other What articles of clothing are tokens in Monopoly?

HUM:ind Name 11 famous martyrs.

DESC:desc What 's the Olympic motto?

1 Exploratory Data Analysis (EDA)
2 Neural Networks: LSTMs
3 Support Vector Machines
4 Model comparison and scores (F1)
5 Outlook

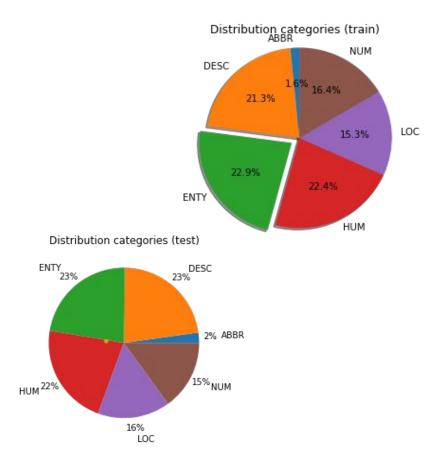
Exploratory Data Analysis (EDA)

The data sets: First insights into categories

The training data set consists of approx. 5500 labeled questions and the testing set of 500 data.

Some charts give an impression of the categories (and subcategories) sizes:

Note that the distribution in the test data is roughly the same for the main categories.

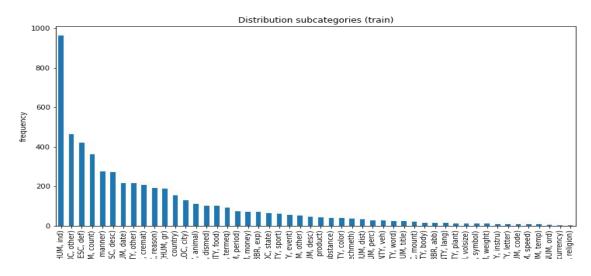


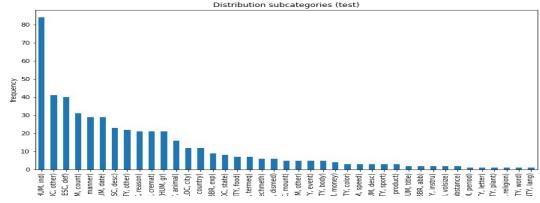
The distribution of subcategories

The distribution of subcategories is very unbalanced. The training data set shows 47 categories, the test data set still 38.

By far most frequent (18%) subcategory is 'Individual' (HUM), followed by 'Definition' (DESC) with 8%.

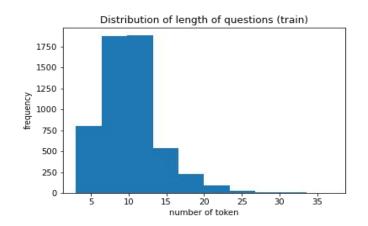
Note: 'Other' appears in three different categories.

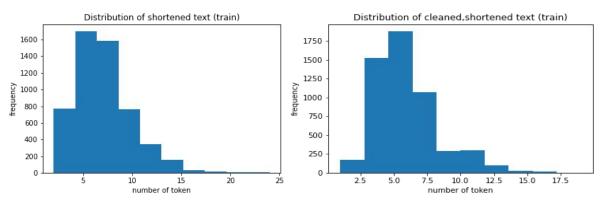




Some more EDA on the question texts

Analysis of the question texts cover the length of the question, i.e. number of tokens (words, punctuation marks, special characters).

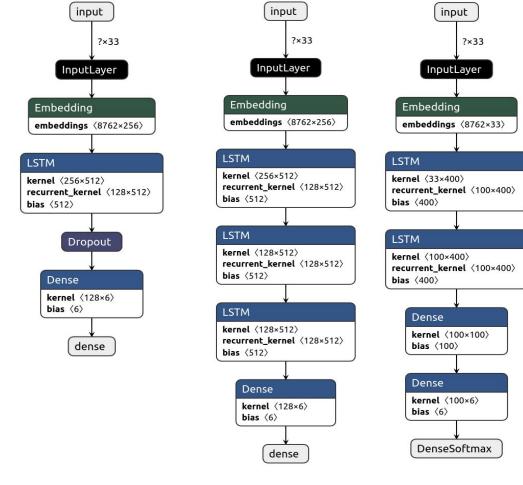




Neural Networks: LSTMs

The LSTM Architecture

The architecture of ...

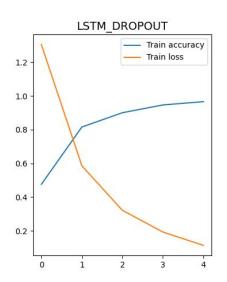


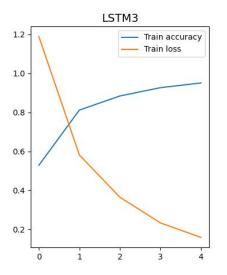
Model 1

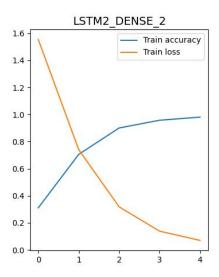
Model 2

Model 3

LSTM Training Accuracy







Support Vector Machines

The SVM Pipeline

```
from sklearn.pipeline import Pipeline
from sklearn.feature_extraction.text import
CountVectorizer
pipe_cv_ng12 = Pipeline(steps=[
   ('data_cv',
CountVectorizer(stop_words=[],
ngram_range=(1, 2)),
   ('model', svm.LinearSVC())
```

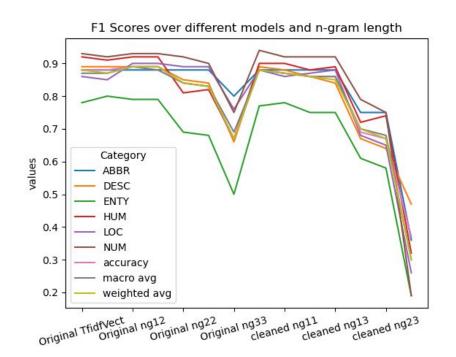


SVM

Accuracies on training set are tested with cross validation and finally after hyperparameter tuning we use the testset.

Different approaches:

Test with several text test sets and different n-gram length.



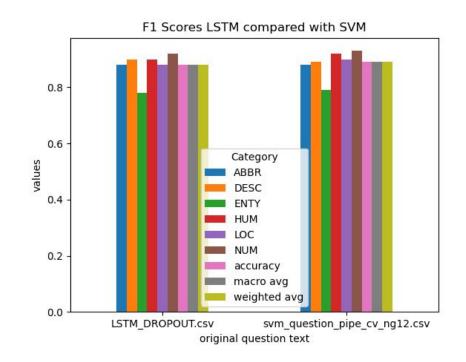
Model comparison and scores (F1)

Model comparison

Both models are similar in their scores - have only small differences in some of the categories

Both model to compare result processed the original text without cleaning

Interesting findings: The LSTM was better in categories where the SVM had problems with.

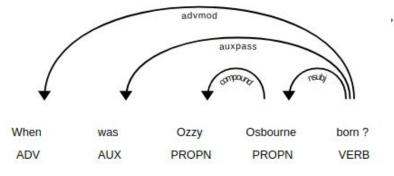


Outlook

Outlook

We also had some ideas to include spaCy to 'make more out of the question text'. For example on might also include POS or furthermore alphanumeric characters, lowercase etc. But this is postponed to another project.

When was Ozzy Osbourne born ?



```
token - lemma_ - lower_- pos_ - tag_- dep_ - sentiment - is_alpha - is_stop

0 When - when - when - ADV - WRB - advmod - 0.0 - True - True

1 was - be - was - AUX - VBD - auxpass - 0.0 - True - True

2 Ozzy - Ozzy - ozzy - PROPN - NNP - compound - 0.0 - True - False

3 Osbourne - Osbourne - osbourne - PROPN - NNP - nsubj - 0.0 - True - False

4 born - bear - born - VERB - VBN - ROOT - 0.0 - True - False

5 ? - ? - ? - PUNCT - . - punct - 0.0 - False - False
```