Deep Learning Course

Native Language Identification using RNN

Gilad Eini and Liat Nativ

**Introduction**

Native language identification is the task of determining the native language of the author, given only a text in a foreign language. The task was introduced by Koppel et al. (2005) has gained much popularity, typically aimed to the language of learners. However, this task is relevant in the much more challenging context of advanced non-native speakers, especially since while English language dominates the internet, there are far more non-native English speakers than native speakers.   
This Project focuses on the task of native language identification of highly fluent speakers, using a corpus of *Reddit* posts in which the native language of the author has been accurately annotated (Rabinovich et al. 2018). Based upon this corpus our task is to identify the native language of non-native authors from 5 different counties of origin.  
We obtained these results:  
Accuracy XXXXXXXXXXXXX  
Recall XXXXXXXXXXXX  
…

**Dataset**

We used the Reddit dataset released by Rabinovich et al. (2018)   
Reddit is a popular online community consisting of thousands of forums in a wide range of topics. The dataset includes Reddit posts whose content is generated by users specifying their country as a flair (metadata attribute). We selected 5 native languages – represent here as the countries of origin - with a large amount of data for our dataset: USA, Germany, Turkey, France, Russia. We sampled an equal portion of sentences for each class (100K sentences), dividing each to (90%) training set and 10% test set. We further divided the training set to (90%) training set and 10% dev (validation) set. Our resulting dataset is 81000, 10000 and 9000 sentences for the training test and dev sets respectively.

**Pre-processing** & data representation  
We performed some clean-ups and pre-processing aimed at getting better classification results. We removed any non-alphanumeric characters from the text, except for punctuation that might be meaningful for determining the author's native language.   
We also restricted the length of the sentences, getting best results with sentences from XXXX to XXXX length.  
We shuffled the sentences in all 3 groups to prevent any bias.  
To create a suitable machine-learning representation we converted each word to a vector of real numbers using word embeddings. We used GloVe pre-trained word embeddings, based on Wikipedia (2014) containing 6B tokens, 400k vocabulary[[1]](#footnote-1). We got best results when using 100-dimension word vectors.

**Model**  
We used bi-directional RNN with LSTM.   
  
**Experiments and improvements**   
Loss function  
LSTM cell (basic, gru)  
# of hidden units  
Optimizers  
Regularization  
Learning Rate  
Batch Size  
Embedding dim size  
dataset size  
sentence length

**Results**Accuracy & other measures   
Learning Graph

1. https://nlp.stanford.edu/projects/glove/ [↑](#footnote-ref-1)