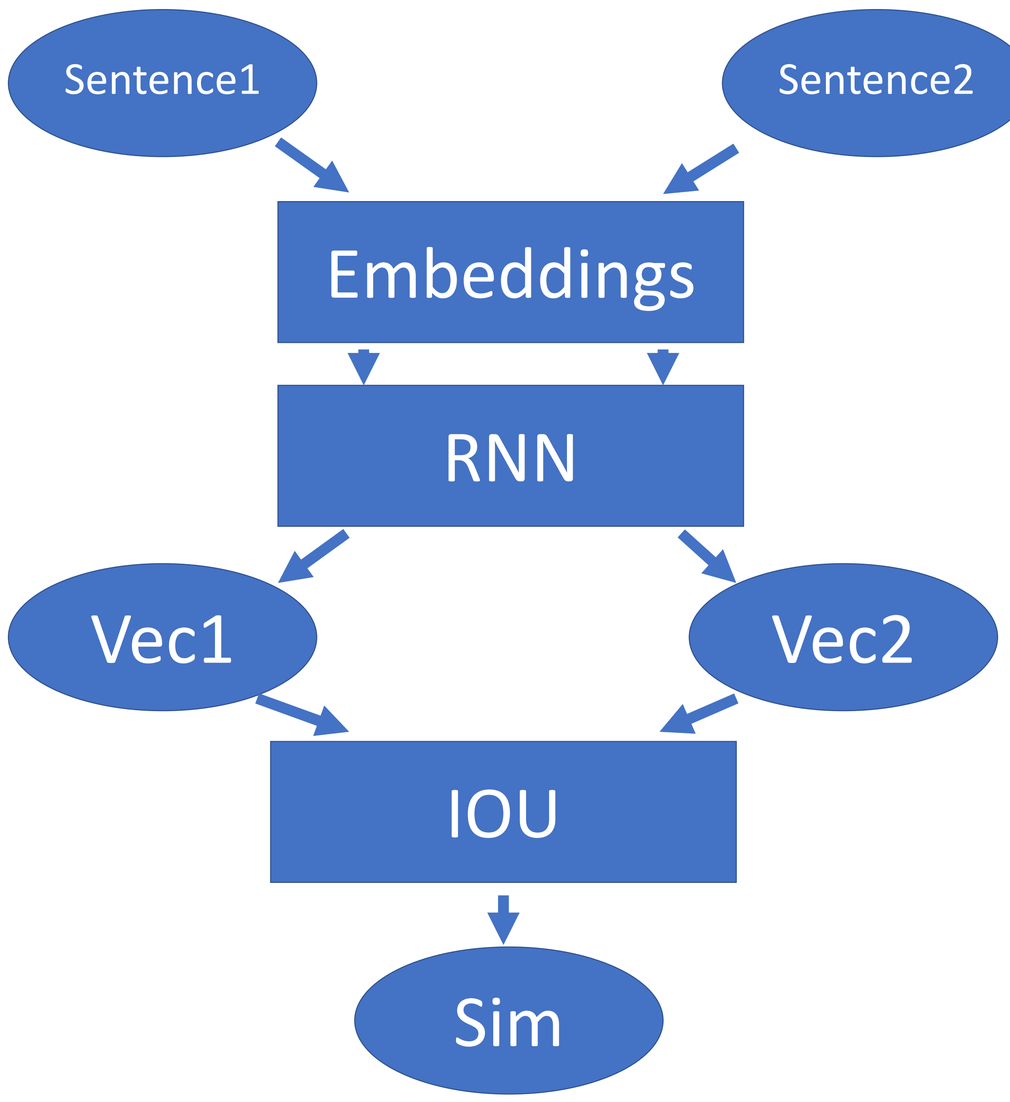
Unsupervised pretraining for text classification using siamese transfer learning

Maximilian Bryan, J. Nathanael Philipp



SIAMESE MODELS

When training
- Map input data onto vector representations
- Input data does not need to be reconstructed
- Thus, less weights are needed

Training pairs
- Data is shown in positive and negative pairs
- Positive pairs
  - have to be mapped onto the same vector representation
  - indicate semantic similarity
- Negative pairs
  - have to be mapped onto differing vector representations
  - indicate semantic distance
- Training pairs‘ vector representations are compared using distance functions like cosine or IOU

Labels
- Weak labels can be used
- We are using sentences from news articles
- Two sentences appears coherently in the same article
  - positive pair, i.e. similarity should be high
- Two randomly chosen sentences
  - negative pair, i.e. similarity should be low
- **Sentences with a similar semantic meaning or sentence structure result in a similar vector representation**

ABSTRACT

When training neural networks, huge amounts of training data typically lead to better results. When only a small amount of training data is available, it has been proven useful to initialize a network with pretrained layers. For NLP tasks, networks are usually only given pretrained word embeddings, the rest of the network is not pretrained since pretraining recurrent networks for NLP tasks is difficult. In our article we present a siamese architecture for pretraining recurrent networks on textual data. The network has to map pairs of sentences onto a vector representation. When a sentence pair is appearing coherently in our corpus, the vector representations should be similar, if not, the representations should be dissimilar. After having pretrained that network, we enhance it and train it on a smaller dataset in order to have it classify textual data. We show that using this kind of approach for pretraining results in better results comparing to doing no pretraining or only using pretrained embeddings when doing text classification for a task with only a small amount of training data.

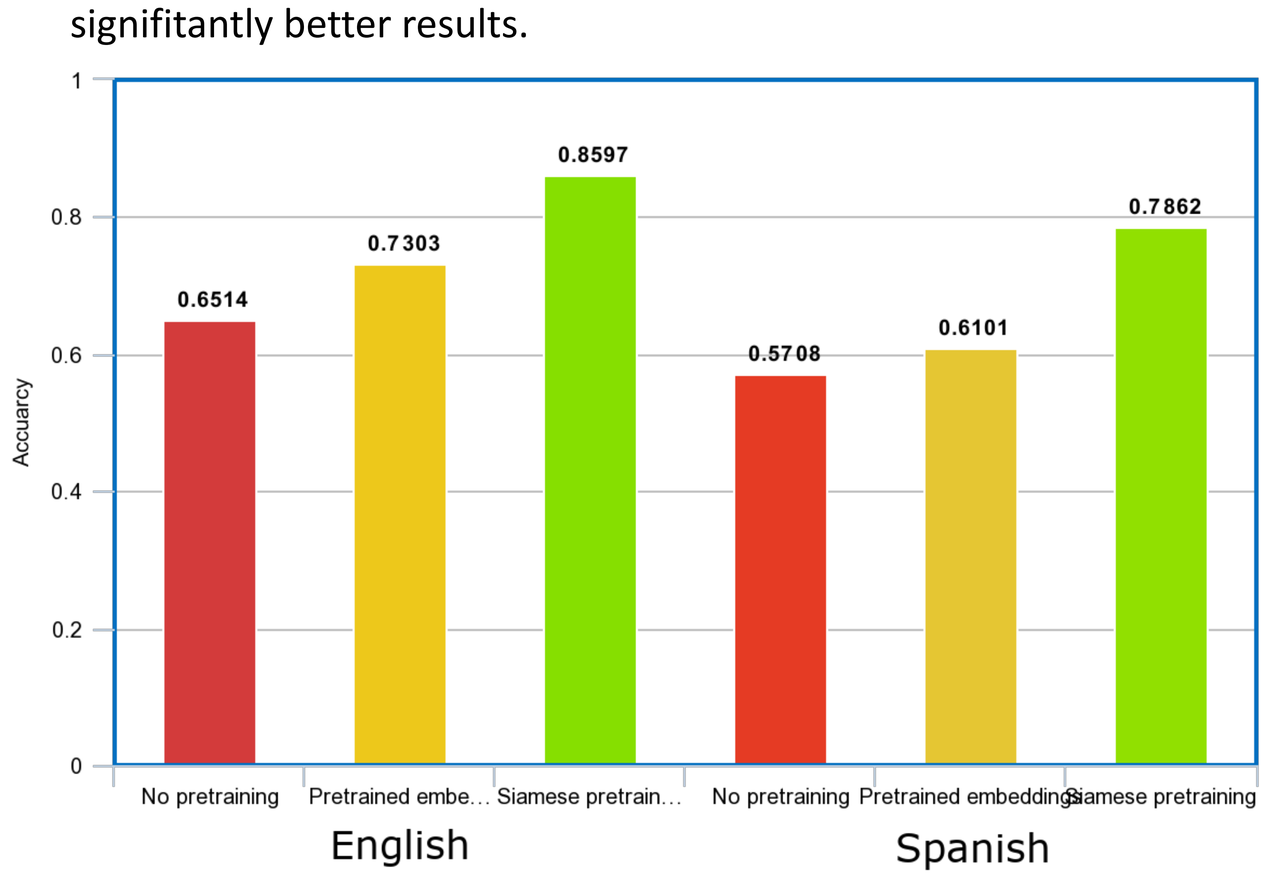
TASK

The bot/gender profiling challenge of PAN @ CLEF 2019 offered a dataset with tweets of 2880 different authors. It was the participants‘ tasks to classify the tweets, whereas the classes where referring to the tweet‘s author being male, female or a bot. We created a classifier that contained of an embedding an a recurrent layer and received tweets tokenwise. The output contained three softmax activated nodes indicating the probability for the corresponding classes.

RESULTS

We compared three different approaches:
- a not pretrained classifier
- a classifier with pretrained wordembeddings
- a classifier with pretrained wordembedding and recurrent layer having used the siamese architecture

Our results show that a more pretrained architecture leads to signifitantly better results.



Printed at Universitätsrechenzentrum Leipzig