Part 1 – Understanding the Challenge

Name

Simon Raviv

CBOW Approach

The two languages can't be distinguished using bag of words approach. In this approach, we lose the information regarding the order, due to the summation of the word vectors, which don't encode information regarding the order of the sequence.

Markovian bigram/trigram approach

The two languages can't be distinguished using bigrams approach and trigram approach. In this approach, we use the statistics over the bigram/trigram of the letters in the dataset.

Due to the nature of bigram/trigrams, which are limited in history of 2/3 words, this kind of sequence can't be distinguished in this language, since in this language each subsequence, e.g., [1-9]+, a+, etc. is not bounded by in length.

Note: For sub languages of this languages, it will be possible to use trigrams to distinguish (thought probably not too accurately), for example, when the number of digits in each digit's part, is limited to 1 digit.

CNN Approach

The two languages can't distinguish using CNN approach. In this approach, we use spatial information from the input, since the spatiality in this kind of languages is not bounded, i.e. in this language each subsequence, e.g., [1-9]+, a+, etc. is not bounded by length, CNN will not work. Each filter size is fixed; thus, we will not be able to learn the connections if they span over bigger number of words than the filter size. Even with sliding windows, the situation is the same, since the window size is limited.

Note: For sub languages of this languages, it will be possible to use CNN to distinguish (thought probably not too accurately), for example, when the subsequences are bound by length and the filter length is bigger than this bound.