DL for NLP - Ass. 3 - challenge

Itamar Trainin, Guy Gispan

December 2019

a.

The two given language descriptions given in this assignment cannot be distinguished using a simple CBOW language model since the difference between them inherits from the ordering of the tokens in the text. CBOW computes the sum of the token representations and therefore dis-regards the order between them. Given a sentence in the first language and a sentence in the second language which has the same structure as the first sentence, only the b's and the c's are switch so it is in the format of the second language, the two sentences will have the **same** CBOW value and therefore will not be distinguishable.

b.

A bi/tri-gram approach will still not be enough to solve this problem since we can have arbitrary size digits between the letter tokens which won't allow the model to know whether the c's are next to the a's or the d's or alternatively, whether the b's are next to the a's or to the d's. Or actually any ordering the letters next to each other, we again must have a notion of order in the letters which cannot be achieved using bi/tri-grams.

c.

I this case again we will not be able to distinguished between the languages because here as well we can have an arbitrary repetitions of numbers between the letters which will not allow us to have a window that can catch ordering between the letter groups. Here again we need a model that can accept an arbitrary amount of tokens (that can also represent ordering).