Prof. Dr. Andreas Hotho, Albin Zehe, Konstantin Kobs
Lehrstuhl für Data Science

13./14.07.2021

8. Assignment in "Machine Learning for Natural Language Processing"

Summer Term 2021

1 General Questions

❓ Something to think about

1. Given two sets of word embeddings E and E', how can you combine these to form a new embedding?

❓ Something to think about

2. Propose a neural network architecture that would be suited for generating image descriptions, that is, given an image as input, creates a sentence describing the content of the image! Provide reasoning why your architecture is a good choice for this task.

2 Conceptnet Numberbatch

In the lecture, ConceptNet Numberbatch, a set of word embeddings enriched with world knowledge, has been introduced. To optimise the matrix W representing these embeddings, a loss function L has been defined relative to a set of initial embeddings U, $|U| = m$, and a knowledge graph $G = (V, E)$:

$$L(W) = \sum_{i=1}^{m} \left[\alpha_i \|w_i - u_i\|^2 + \sum_{(i,j)\in E} \beta_{i,j} \|w_i - w_j\|^2 \right]$$

Instead of regular gradient descent, ConceptNet Numberbatch uses the following update step to minimise the loss function:

$$W^{k+1} = \text{normalize}\left(\left[\left(SW^k + AW^0 \right)^T (I + A)^{-1} \right]^T \right)$$

Note: There is an error in the original paper which I only noticed after building this assignment. In the paper, the transpose operations are missing, leading to a shape error. The two transpose operations are equivalent to switching the operands of the multiplication. This error probably occurred in the paper because putting the normalisation factor at the end seems more intuitive, but the authors forgot to add the transposes.

1. In the lecture, it was stated that the component $(I + A)^{-1}$ has a "normalising" effect on the formula by preventing words contained in the original embedding from having twice the influence of words not in the original embedding.

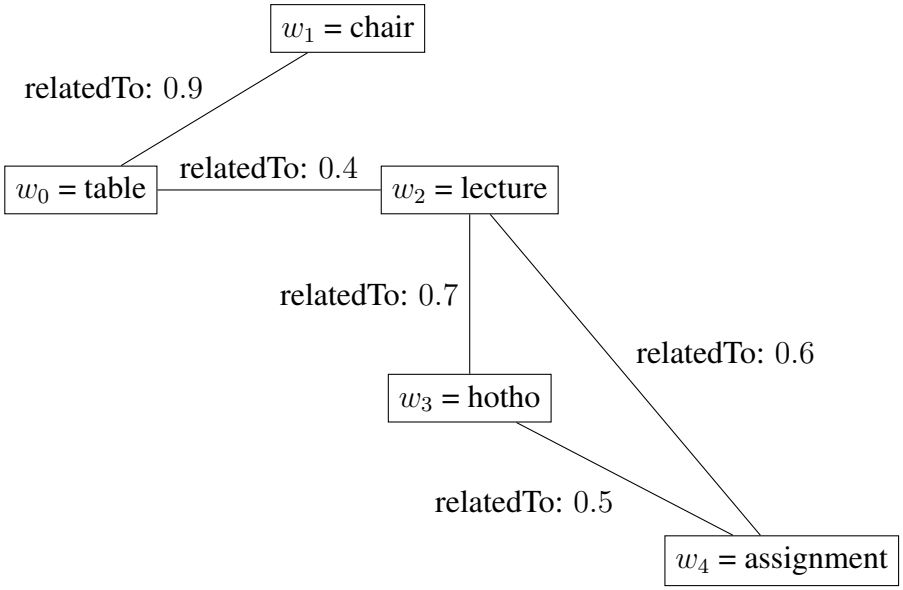
   Argue that this is indeed what happens!

   Some suggested steps:

   - What fixed properties does $(I + A)^{-1}$ have?
   - Knowing these properties, what happens when the first component gets multiplied with $(I + A)^{-1}$?

   You do not have to provide mathematical proof, just an understandable reasoning!

2. Assume the following knowledge graph:



   And the following set of initial word embeddings U:

   $u(\text{chair}) = (0.8, 0.6, 0.7)$, $u(\text{table}) = (0.74, 0.7, 0.6)$, $u(\text{lecture}) = (0.1, 0.5, 0.2)$, $u(\text{assignment}) = (0.2, 0.4, 0.18)$. Use this order of embeddings in your matrices to later comply with the sample solution.

   Note that there is no embedding for "hotho".

a) What are the values of U, W^0, A and S? You do not have to normalise matrix S.

b) Use the update formula to calculate W^1!

c) Calculate $L(W^0)$ and $L(W^1)$ to show that the update step has decreased the loss!