Name: **Student ID number:**

(1) Given co-occurrence probabilities $P(c|t)$ between target words t and context words c , the crucial insight in creating GloVe word embeddings (Pennington et al. 2014) is: *(choose one)* 1 points

(a) using the difference between co-occurrence probabilities $P(c|t_1) - P(c|t_2)$

(b) using the ratio of co-occurrence probabilities $P(c|t_1)/P(c|t_2)$

(c) using co-occurrence probabilities $P(c|t)$ instead of counts of c and t.

b

(2) Is the following statement true or false: In Skip-gram with negative sampling, a classifier is trained on a multi-class prediction task. 1 points

False

(3) Suppose that you are using an RNN for POS tagging and that the output at each time step is a distribution over the POS tagset, as generated by a softmax layer. Is this output sufficient to determine the optimal POS tag sequence for the whole sentence? Say why or why not, and make some suggestions for improvement if necessary. (3-5 sentences) 3 points

No, the RNN will not be able to find the optimal tag sequence for the whole sentence (1). This is because an RNN will tag each word based only on the *prior context*, and *does not have the ability to consider POS-tags of following words*, nor back-track based on them. Outputting the best POS tag from a distribution of POS tags at each time step will thus not necessarily lead to the optimal POS tag sequence for the entire sentence (1).
Suggestions for improvement can involve either using the Viterbi algorithm on top, which can find highest probability sequence exactly based on the POS distributions of all words in the sentence, or else use a bi-directional RNN to mitigate the above problem. (Any other reasonable suggestion for an improved model is okay too (1).

(4) What is the difference between "similarity" and "relatedness" of words? Illustrate using examples. 1 points

Similar words are substitutable in many contexts and share properties (cat and dog). Related words occur in the same contexts (cat and purr)

(5) Describe in 1-2 sentences one improvement made in Abnar et al. (2018) "Experiential, Distributional and Dependency based Word Embeddings have Complementary Roles in Decoding Brain Activity" over the original experiments in Mitchell et al. (2008) " Predicting Human Brain Activity Associated with the Meanings of Nouns"? 2 points

Any one difference between the two papers: Abnar et al. (2018) (i) used and compared 8 different word embedding models on their usefulness for predicting the neural activation patterns associated with concrete nouns (ii) included an experiential model, based on crowd-sourced association data. (iii) also predicting in the reverse direction (words vectors from brain images), etc.

(6) Name at least three possible applications of Recurrent Neural Networks for language processing (1.5 points). For one of them, describe the input and the output of the network at one given timestep (2 points). 3.5 points

Name: **Student ID number:**

Applications can include:

- Language modelling. Input: the current word in the sequence (its embedding), the previous hidden state of the network. Output: the probability distribution over the vocabulary words $\rightarrow$ the most probable word becomes the next word in the sequence.
- POS tagging. Input: the current word in the sequence (its embedding), the previous hidden state of the network. Output: the probability distribution over the possible POS tags $\rightarrow$ the most probable POS tag is assigned to the current word.
- Machine translation. Input (encoder): the current word in the source text (its embedding), the previous hidden state of the network. Output (encoder): the updated hidden state of the network, with the information about the current word added. Input (decoder): the current word in the translation (its embedding), the previous hidden state of the network, optionally the encoding of the source text. Output (decoder): the probability distribution over the vocabulary words in the target language $\rightarrow$ the most probable word becomes the next word in the translation sequence.
- Image captioning. Input: the current word in the image description, the previous hidden state of the network, optionally the encoding of the image. Output: the probability distribution over the vocabulary words $\rightarrow$ the most probable word becomes the next word in the image description.

Any other reasonable applications and input-output descriptions are accepted.