

Name:

Student ID number:

#### Question 4

- (a) Is a feedforward network a suitable model for language modelling? Also say why or why not (in 1-2 sentences). 2 points
- (b) Do you agree with the statement "Sequence models are complete models of sentences, as a sentence is a sequence of words"? Also say why or why not (in 1-2 sentences). 2 points
- (c) Suppose we use a bigram-based hidden markov model (HMM) to tag the words  $w_i$  in a sentence with their parts-of-speech  $t_i$ . The most probable tag sequence  $\hat{t}_1^n$  for the sentence  $w_1^n$  in this model is given by :

$$\hat{t}_1^n = \arg \max_{t_1^n} \prod_{i=1}^n P(t_i | t_{i-1}) P(w_i | t_i)$$

What simplifying assumptions (*hint: "independence" assumptions*) are we making in this model? Why are we required to make them?

#### Question 5

- (a) In the original Skip-gram model, if  $v_j$  is the vector for the target word  $w_j$  and  $c_k$  is the vector for the context word  $w_k$ , the softmax is used to convert their dot product into probabilities. 6 points

$$p(w_k | w_j) = \frac{e^{c_k \cdot v_j}}{\sum_{c_i \in V} e^{c_i \cdot v_j}}$$

Here, the normalisation term in the denominator is expensive to compute (as for every word it has to be computed over the entire vocabulary  $V$ ). The solution to this problem in Word2Vec is "negative sampling".

- (i) Describe in a few sentences (using examples if you like) how negative sampling works in the Skip-gram model. Please make sure to say precisely what the negative samples are in this case.
- (ii) Is the following statement then true or false: **Skip-gram with negative sampling trains a classifier on a binary prediction task.**
- (b) Suppose you are making distributional word vectors from a corpus. Describe at least two things you might do differently if you are constructing semantic vectors (i.e. vectors capturing word meaning) versus syntactic vectors (i.e. vectors capturing syntactic behaviour). 4 points

#### Question 6

- (a) 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? Justify your answer, and suggest improvements if necessary. *Hint: There is not necessarily only one correct answer here; you will be judged on your reasoning and knowledge of the task.* 5 points

Name:

Student ID number:

(b) 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*)

2 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  $t$ .

(c) In Mitchell et al. (2008) "Predicting Human Brain Activity Associated with the Meanings of Nouns",

3 points

(i) what imaging technique is used to measure neural activity in subjects, and what stimulus is presented?

(ii) what technique is used to learn and represent the meaning of nouns in the computational model?

(iii) summarise the main findings in 1-2 sentences.