UNIVERSITY OF EDINBURGH
COLLEGE OF SCIENCE AND ENGINEERING
SCHOOL OF INFORMATICS

INFR11062 MACHINE TRANSLATION

Monday 30th April 2018

14:30 to 16:30

INSTRUCTIONS TO CANDIDATES

Answer any TWO of the three questions. If more than two questions are answered, only QUESTION 1 and QUESTION 2 will be marked.

All questions carry equal weight.

CALCULATORS MAY NOT BE USED IN THIS EXAMINATION

MSc Courses

Convener: G. Sanguinetti
External Examiners: W. Knottenbelt, M. Dunlop, M. Niranjan, E. Vasilaki

THIS EXAMINATION WILL BE MARKED ANONYMOUSLY

Answer the questions as concisely as possible. Typically, a paragraph is sufficient, and longer answers will not receive more credit. When a question consists of multiple sub-questions or asks for a list, a sentence or two per item is sufficient.

1. You have been asked to create a machine translation system to translate Turkish news articles into English. You decide to build a neural machine translation system.

   (a) How would you train your model? Which loss function and which training algorithm would you use? How would you decide your stopping criterion? [*3 marks*]

   (b) For efficiency and training stability, you want to perform every training step on a small mini-batch of sentences. How do you deal with sentences of varying length in one mini-batch? How can you reduce inefficiencies that result from this? [*3 marks*]

   (c) The client can provide 200,000 sentence pairs of Turkish–English news data, which is considered low-resource for neural machine translation. Apart from using more training data, list two techniques that you learned about that are especially helpful in training neural translation systems in low-data conditions. [*4 marks*]

   (d) The client notes that he has access to millions of sentences of monolingual English news texts. Describe three ways how this data could be used to improve your neural machine translation system. [*6 marks*]

   (e) The client also notes that he also has access to millions of sentences of German–English news data. Describe a way how this data could be used to improve your neural machine translation system, other than just using the English side of the bitext as monolingual data. [*3 marks*]

   (f) The client wants to compare translation quality of your system to that of a commercial translation system using BLEU. What resources do you need for this evaluation? Briefly explain how BLEU is computed (you do not need to provide the exact equation or an example). [*3 marks*]

   (g) Another client is interested in Turkish→English translation of legal documents. He can provide 20,000 sentence pairs of training data from this domain, and you are allowed to use the training data of the first client. How would you go about training a system that is adapted to the legal domain? [*3 marks*]

2. language representation and linguistic structure

(a) How are words represented internally in a word-level neural language or translation model? Discuss both input and output words (a short text discussion is sufficient, and no equations or diagrams are required). [*4 marks*]

(b) What are the consequences (in terms of computational complexity) of increasing the network vocabulary size in neural machine translation? [*3 marks*]

(c) *Byte-pair encoding* (BPE) has been adapted to the problem of word segmentation to achieve open-vocabulary translation with a fixed-size network vocabulary. Explain the algorithm, and show the algorithm's output, based on the following example dictionary. Compute the first 4 BPE merge operations, and in the end, apply them to all words in the dictionary. [*6 marks*]

word	frequency
cat	9
car	5
rat	4
ran	3

(d) For word-level systems, a popular technique for handling out-of-vocabulary words is to represent them with a special UNK token, and then copy or translate the word with a back-off mechanism based on the corresponding source word. What is a limitation of this approach? [*3 marks*]

(e) Phrase-based machine translation is poor at modelling grammatical agreement over long distances. Explain where this limitation comes from, and how most neural translation systems overcome this limitation. [*3 marks*]

(f) Homographs are words that share the same spelling, but have different meanings. An example is *bank*, which can refer to a financial institution or the land along a river. Discuss if neural machine translation (with raw text input) is technically able to disambiguate homographs. Depending on your answer, also explain how, or why not. [*3 marks*]

(g) Natural languages allow for the same meaning to be expressed in different ways, for example using synonyms and paraphrases. Discuss the consequences for the automatic evaluation of machine translation. [*3 marks*]

3. neural machine translation architectures

(a) What is the main difference in terms of network architecture between a neural language model and

i. the encoder in a neural machine translation model?

ii. the decoder in a neural machine translation model? *[4 marks]*

(b) An attention model produces a fixed-size representation of some input values, even though the number of input values is variable. Explain what mechanism different attention models have in common that guarantees this fixed-size representation. *[4 marks]*

(c) Why is applying dropout problematic for recurrent connections? Explain the problem, and discuss one method described in the research literature to use dropout in recurrent language and translation models that mitigates this problem. *[4 marks]*

(d) In language modelling, self-normalized networks, which use an identity activation function instead of a softmax activation function in the output layer, can deliver large speed improvements at test time when only the score of a hypothesis is queried. Explain why. *[4 marks]*

(e) Susan claims that in neural machine translation, self-normalized networks would not substantially improve the computational complexity of decoding. Say whether you agree with this statement, and explain why. *[4 marks]*

(f) Recently used architectures for neural machine translation, convolutional networks and self-attentional networks, use masking in the decoder to ignore any words or hidden states after the current time step. Peter experiments with this, and notices that his cross-entropy at training time improves when he removes masking. Why is masking used, and what are the consequences of removing it? *[5 marks]*