Ollscoil Chathair
Bhaile Átha Cliath
Dublin City University

SEMESTER 2 EXAMINATIONS 2018/2019

MODULE: CA4012 - Statistical Machine Translation

PROGRAMME(S):

CASE	BSc in Computer Applications (Sft.Eng.)
CPSSD	BSc in Computational Problem Solv & SW Dev.
ECSA	Study Abroad (Engineering & Computing)
ECSAO	Study Abroad (Engineering & Computing)

YEAR OF STUDY: 4,O,X

EXAMINER(S):

Prof Andrew Way	(Internal)	(Ext: 5074)
Dr Mohammed Hasanuzzaman	(Internal)	(Ext: 6719)
Dr Dimitar Shterionov	(Internal)	(Ext: 6719)
Dr Hitesh Tewari	(External)	
Prof Brendan Tangney	(External)	

TIME ALLOWED: 3 Hours

INSTRUCTIONS: Answer **five** questions. You **must** attempt *at least one* question from *each* of Sections A, B and C. All questions carry equal marks.

PLEASE DO NOT TURN OVER THIS PAGE UNTIL YOU ARE INSTRUCTED TO DO SO.
The use of programmable or text storing calculators is expressly forbidden.
Please note that where a candidate answers more than the required number of questions, the examiner will mark all questions attempted and then select the highest scoring ones.

There are no additional requirements for this paper.

SECTION A

QUESTION 1 ***[TOTAL MARKS: 20]***

Q 1(a) **[6 Marks]**

Any statistical approach to MT requires the availability of aligned bilingual corpora which are (i) large, (ii) good-quality, and (iii) representative. Explain why all three requirements are important.

Q 1(b) **[7 Marks]**

Provide the fundamental equations of (i) the noisy channel model of SMT, and (ii) the log-linear model of SMT. With reference to these equations, name the different components in both models, and describe their basic function.

Q 1(c) **[4 Marks]**

Give **two** reasons why we can expect a customised client-specific MT engine to produce superior quality translations compared to a freely available web-based system such as Google Translate.

Q 1(d) **[3 Marks]**

With the switch from statistical to neural MT, should we be concerned that the field will be dominated more than ever by the large providers such as Google Translate and Bing Translator?

[End of Question 1]

QUESTION 2 *[TOTAL MARKS: 20]*

Q 2(a) **[5 Marks]**

Corpus-based models have outperformed rule-based MT systems for 30 years now. In your opinion, do rules play *no* part in today's state-of-the-art neural MT models, or is the way forward a hybrid combination of rules and data-driven systems? Provide **three** examples to support your argument.

Q 2(b) **[9 Marks]**

In 2016, Google announced that their neural models were "bridging the gap" between MT and human translation quality. In 2018, Microsoft claimed to have achieved "human parity" for Chinese-to-English neural MT. What do you think of such claims? How would you go about testing them? What advice would you give to translators who were concerned at the possibility of losing their job in light of such claims?

Q 2(c) **[6 Marks]**

While in most cases neural models of translation clearly outperform statistical MT (SMT), our ability to explain their outputs is less than it was for SMT. Why do you think this is the case? How would you suggest that neural models could become more inspectable?

[End of Question 2]

[END OF SECTION A]

SECTION B

QUESTION 3 ***[TOTAL MARKS: 20]***

Q 3(a) **[5 Marks]**

Explain the Markov assumption. Why do we need to take it into account when building *n*-gram language models?

Q 3(b) **[5 Marks]**

How is the Maximum Likelihood estimate of a *trigram* language model computed? Compute *P*(ate|Bukka) from the following unigram and bigram counts.

Bukka sandwich	10
ate Bukka	8
Bukka the	15
Bukka	35
Bukka ate	16
ate the	20
sandwich ate	2
sandwich Bukka	3

Q 3(c) **[4 Marks]**

Assume (i) a unigram and (ii) a bigram language model trained on standard English. Based on these two different language models, how would you expect the probability of the phrase “the sandwich Hakka ate” to compare to the probability of the sentence “Hakka ate the sandwich”?

Q 3(d) **[6 Marks]**

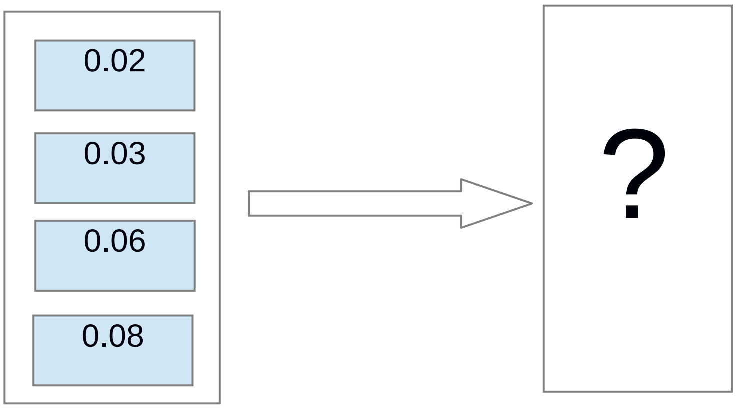
Why do *n*-gram language models need to be smoothed? Name **three** methods of smoothing, and explain one method in detail.

[End of Question 3]

QUESTION 4 **[*TOTAL MARKS: 20]***

Q 4(a) **[6 Marks]**

In decoding, what is pruning, and why it is important? How many of the following hypotheses will be pruned if we prune all hypotheses that are at least 0.4 times worse than the best hypothesis? Show your calculations.



Q 4(b) **[7 *Marks*]**

Assume the following partial phrase table:

se	she	0.4
bhalobase	likes	0.3
bhalobase	likes to	0.5
se bhalobase	he likes	0.3
se bhalobase	he likes to	0.6
khete	eat	0.6
khete	eating	0.7
khete bhalobase	likes eating	0.3
khete bhalobase	likes to eat	0.7

Assume that only monotone word order is permitted and that the language model is ignored.

Draw the search graph constructed during decoding for the sentence “se bhalobase khete”.

Q 4(c) **[7 Marks]**

Using the constructed search graph for Q 4(b), calculate all possible hypotheses, and indicate which hypothesis provides the optimal translation for the above sentence.

[End of Question 4]

[END OF SECTION B]

SECTION C

QUESTION 5 ***[TOTAL MARKS: 20]***

Q 5(a) **[4 Marks]**
Explain Word Error Rate (WER). How is the WER score computed?

Q 5(b) **[4 Marks]**
Given the following two candidate translations (from *MT system 1* and from *MT system 2*), compute their WER scores with respect to the reference provided.

- *Candidate MT system 1*: near the shore the ship sank
- *Candidate MT system 2*: the big ship sank close to the shore
- *Reference*: the large ship sank near the shore

Q 5(c) **[6 Marks]**
Why is the BLEU score typically calculated over the entire document, rather than on a sentence-by-sentence level? Give **two** examples of the relevant shortcomings of this metric.

Q 5(d) **[6 Marks]**
In BLEU we compute the *n*-gram clipped precision (typically for 1-, 2-, 3- and 4-grams). What is clipped precision and why do we need to perform clipping? To support your explanation, provide **two** examples.

[End of Question 5]

QUESTION 6 ***[TOTAL MARKS: 20]***

Q 6(a) **[6 Marks]**

The goal of MT is to find a sentence e that is the most likely translation of a source-language sentence f, p(e|f).

Why is the translation model concerned with the translation probabilities of *words* or *phrases*, rather than the sentence as a whole? Use the following corpus to provide **one** example that would support your answer.

Source (Dutch)	**Target (English)**
ze speelt tennis graag .	she likes to play tennis .
hij eet pizza graag .	he likes eating pizza .
ze eet spaghetti graag .	she likes eating spaghetti .
hij speelt voetbal graag .	he likes to play football .

Q 6(b) **[6 Marks]**

Expectation maximisation (EM) is an algorithm to iteratively estimate translation probabilities and alignments. Explain what the following three formulae compute. How are they employed in the EM algorithm?

1.	$\prod_{j=1}^{l_{\mathbf{e}}} \frac{t(e_j \mid f_{a(j)})}{\sum_{i=0}^{l_{\mathbf{f}}} t(e_j \mid f_i)}$
2.	$\sum_{a} p(a \mid \mathbf{e}, \mathbf{f}) \sum_{j=1}^{l_e} \delta(e, e_j)\delta(f, f_{a(j)})$
3.	$\frac{\sum_{(\mathbf{e},\mathbf{f})} c(e \mid f; \mathbf{e}, \mathbf{f})}{\sum_{e} \sum_{(\mathbf{e},\mathbf{f})} c(e \mid f; \mathbf{e}, \mathbf{f})}$

Q 6(c) **[4 Marks]**

In your own words, explain IBM model 1, IBM model 2, IBM model 3 and IBM Model 4, focusing in particular on their differences and similarities.

Q 6(d) **[4 Marks]**

Assume the following two sentence pairs and alignments:

Greek: i gata ekatse sto halaki
English: the cat sat on the mat
alignment: 1->1, 2->2, 3->3, 4->4, 5->4, 6->5

Greek: i gata ekatse sto halaki
English: sat on the mat the cat
alignment: 1->3, 2->4, 3->4, 4->5, 5->1, 6->2

Using the above examples, explain and illustrate the deficiencies of IBM Model 1. How might you propose a solution to this problem?

[End of Question 6]

QUESTION 7 ***[TOTAL MARKS: 20]***

Q 7(a) **[6 Marks]**

In neural machine translation, the most commonly used architecture is the encoder-decoder model. Explain the basic principles of the encoder-decoder model.
What type of neural networks are suitable for use in the encoder-decoder model, and why?

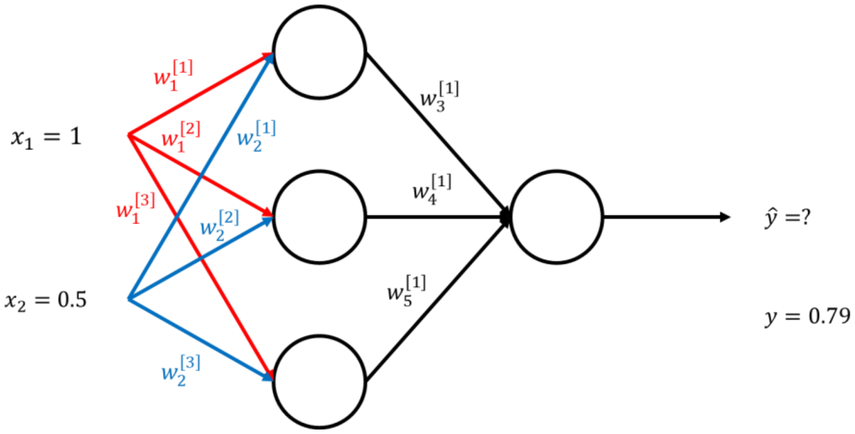
Q 7(b) **[7 Marks]**

In training neural networks, a common problem is overfitting. Explain what overfitting is in your own words. Provide **two** ways of dealing with this problem, together with a description of how they work.

Q 7(c) **[7 Marks]**

Below is given a simple neural network with two inputs, one output and one hidden layer. All the neurons in the hidden layer and the output layer use the **sigmoid** activation function and **all biases are equal to 1.**

Given the input x_1=1, x_2=0.5 and the expected output of y=0.79, perform **one** pass of a forward and backward propagation on this network. All initial weights are uniformly distributed (as is noted in the table). What are the missing values in the table below the network, corresponding to the **weights**, **network output** and **error** after the forward and the backward steps?



			Weights from input 1 to hidden layer			Weights from input 2 to hidden layer			Weights from hidden layer to output layer					
	Input 1	Input 2	$w_1^{[1]}$	$w_1^{[2]}$	$w_1^{[3]}$	$w_2^{[1]}$	$w_2^{[2]}$	$w_2^{[3]}$	$w_3^{[1]}$	$w_4^{[1]}$	$w_5^{[1]}$	Network Output ($\hat{y}$)	Expected Output (y)	Error
Forward	$x_1 = 1$	$x_2 = 0.5$	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3		0.79	
Backward	$x_1 = 1$	$x_2 = 0.5$											0.79	

[End of Question 7]

[END OF SECTION C]

[END OF EXAM]