DUBLIN CITY UNIVERSITY

SEMESTER 2 EXAMINATIONS 2015/2016

MODULE: CA4012 – Statistical Machine Translation

PROGRAMME(S):
CASE BSc in Computer Applications (Sft.Eng.)

YEAR OF STUDY: 4

EXAMINERS:

Prof. Andy Way	(Ext: 5074)
Dr. Jinhua Du	(Ext: 6716)
Dr. Antonio Toral	(Ext: 8712)
Dr. Ian Pitt	

TIME ALLOWED: 2 Hours

INSTRUCTIONS: Answer any **four** questions.
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.

Requirements for this paper (Please mark (X) as appropriate)

- □ *Log Tables*
- □ *Graph Paper*
- □ *Dictionaries*
- □ *Statistical Tables*
- □ *Bible*
- □ *Thermodynamic Tables*
- □ *Actuarial Tables*
- □ *MCQ Only – Do not publish*
- □ *Attached Answer Sheet*
- □ *Exam Paper to be returned with Booklet*

QUESTION 1 ***[TOTAL MARKS: 25]***

Q 1(a) **[8 Marks]**

How might you argue that “The time for MT is now”? Justify your answer by providing *three* use-cases where you claim that MT is the only solution, i.e. that there is no place for human intervention in the translation pipeline for such use-cases.

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

Study these sentences in Abma, an Austronesian language spoken in the South Pacific island of Vanuatu, and their English translations:

1. Mwamni sileng. ⬄ He drinks water.
2. Nutsu mwatbo mwamni sileng. ⬄ The child keeps drinking water.
3. Nutsu mwatbo mwegalgal. ⬄ The child keeps crawling.
4. Mwerava Mabontare mwisib. ⬄ He pulls Mabontare down.
5. Mabontare mwisib ⬄ Mabontare goes down.
6. Mweselkani tela mwesak. ⬄ He carries the axe up.
7. Mwelebte sileng mwabma. ⬄ He brings water.
8. Mabontare mworob mwesak. ⬄ Mabontare runs up.

Assume the following additional lexical entries:

- sesesrakan ⬄ teacher
- mwegani ⬄ eat

Translate the following sentences into Abma:

i. The teacher carries the water down.
ii. The child keeps eating.

Q 1(c) **[5 Marks]**

Describe in your own words how you produced these translations, focussing in particular on the particular types of inferences you made from the parallel data provided. How is this analogous to how SMT works?

Q 1(d) **[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. What are some of the potential problems if any one of these requirements are not met?

[End of Question 1]

QUESTION 2 ***[TOTAL MARKS: 25]***

Q 2(a) **[15 Marks]**

For the source-language sentence (A):

(A) *Kuopion kaupunginvaltuusto hyväksyi liitoksen yksimielisesti maanantaina .*

Assume that the outputs (B)(i) and (B)(ii) were produced by an MT system:

(B)(i) The city council unanimously approved the joint Niiralan on Monday .
(B)(ii) The Kuopion liitoksen City council approved unanimously on Monday .

Assume also that the 'gold standard' reference translation is (C):

(C) The city council of Kuopio accepted the annexation unanimously on Monday .

Calculate the BLEU scores of the two candidate translations using maximum *n*-gram length of 3.

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

Why is *standard* BLEU unsuitable for sentence-level evaluation of MT quality, especially when sentences are short? How would you modify BLEU to make it suitable to be used at sentence level?

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

Explain the concepts of "fluency" and "adequacy" as they apply to the evaluation of MT output. To support your answer, give example translations (in English) which are:

- fluent and adequate,
- fluent but inadequate,
- disfluent but adequate,
- disfluent and inadequate.

[End of Question 2]

QUESTION 3 *[TOTAL MARKS: 25]*

Q 3(a) **[10 Marks]**

Given the following sentences:

- <s> Denis likes Ada </s>
- <s> Ada likes Richard </s>
- <s> Ada hates Java </s>

List all the parameters of the unigram and bigram language models trained with these sentences without smoothing.

Q 3(b) **[10 Marks]**

Given the language models you built in Q3(a), but now with add-alpha smoothing where α=0.3, calculate the probabilities of the following sentences:

- <s> Richard likes Ada </s>
- <s> Richard hates Ada </s>

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

Explain the Markov assumption. Why do you need to take it into account when building *n*-gram language models? How can language models based on neural networks be non-Markovian?

[End of Question 3]

QUESTION 4 ***[TOTAL MARKS: 25]***

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

Assume the following Chinese—English segment-pairs:

S_1	S_2
yuan	*hen yuan*
far	*far away*

The source side is Chinese, and the target side is English. In this question, the *NULL* token is ignored.

Assuming each target word is exactly aligned with one source word, list all possible word alignments for the two segment-pairs.

Q 4(b) **[10 Marks]**

For all the word alignments you computed above, state what the following translation probabilities will be after two iterations of the Expectation Maximisation algorithm, and show all the interim steps by which you arrived at these values:

- $t(far|yuan)$
- $t(away|yuan)$
- $t(far|hen)$
- $t(away|hen)$

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

Explain the term "consistency" as used in phrase extraction.

Q 4(d) **[5 Marks]**

List all phrase pairs that are consistent with the following word alignment:

	A	*B*	*C*
X			■
Y	■		
Z		■	

[End of Question 4]

QUESTION 5 ***[TOTAL MARKS: 25]***

Assume the following partial phrase table:

wo	*I*	0.7
wo	*me*	*0.3*

xihuan	*like*	0.4
xihuan	*like to*	0.5
xihuan	*likes to*	0.1

kaiche	*driving*	0.3
kaiche	*drive*	0.5
kaiche	*drive a car*	0.2

wo xihuan	*I like*	0.3
wo xihuan	*I like to*	0.5
wo xihuan	*I likes to*	0.1
wo xihuan	*me likes to*	0.1

xihuan kaiche	*like driving*	0.3
xihuan kaiche	*like to drive*	0.5
xihuan kaiche	*like to drive a car*	0.1
xihuan kaiche	*likes to drive*	0.1

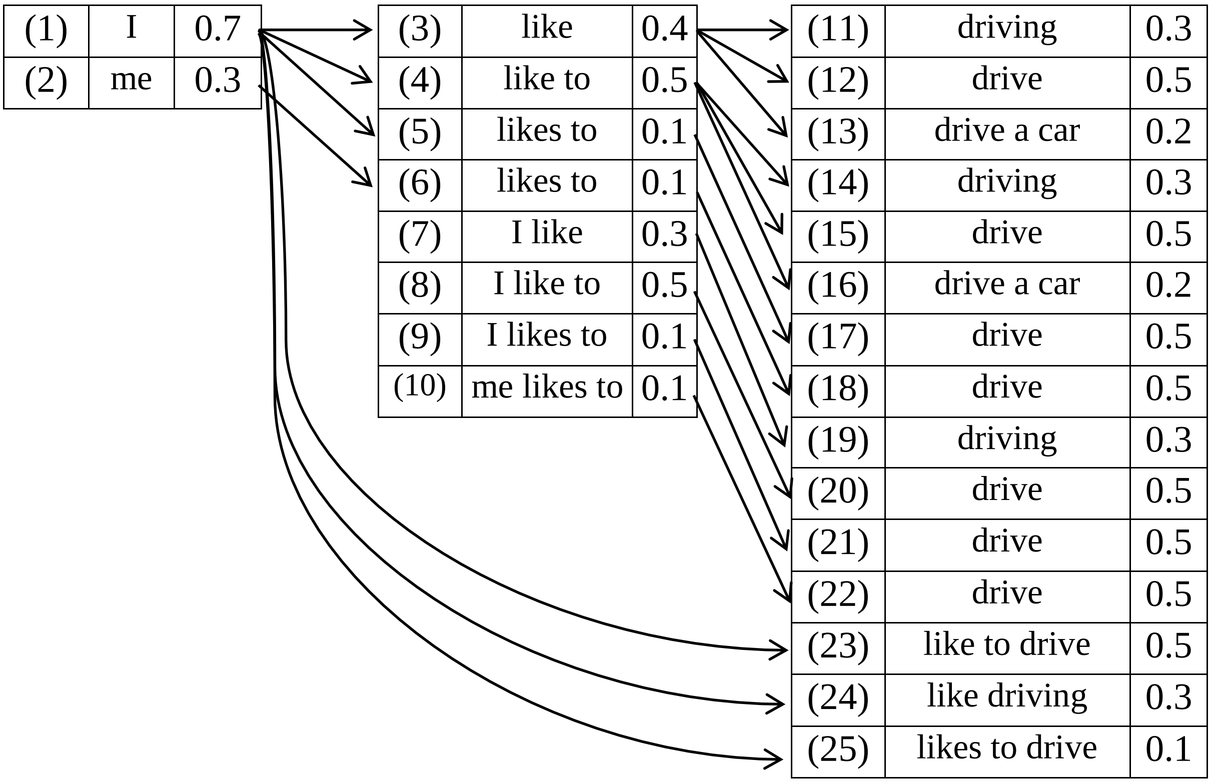
Consider the following input sentence:

wo xihuan kaiche

Assume that:

- Only monotone word order is permitted;
- The language model is ignored.

Then we have the following (partial) search space diagram:



(1)	I	0.7
(2)	me	0.3

(3)	like	0.4
(4)	like to	0.5
(5)	likes to	0.1
(6)	likes to	0.1
(7)	I like	0.3
(8)	I like to	0.5
(9)	I likes to	0.1
(10)	me likes to	0.1

(11)	driving	0.3
(12)	drive	0.5
(13)	drive a car	0.2
(14)	driving	0.3
(15)	drive	0.5
(16)	drive a car	0.2
(17)	drive	0.5
(18)	drive	0.5
(19)	driving	0.3
(20)	drive	0.5
(21)	drive	0.5
(22)	drive	0.5
(23)	like to drive	0.5
(24)	like driving	0.3
(25)	likes to drive	0.1

Q 5(a) **[9 Marks]**
Given this search diagram, calculate the probabilities for all possible hypotheses (search paths). Furthermore, indicate which hypothesis provides the most likely translation for the given input sentence.

Q 5(b) **[6 Marks]**
Given the search diagram, indicate which group of hypotheses can be recombined, and indicate which hypothesis should be selected to represent each group.

Q 5(c) **[6 Marks]**
Assuming histogram pruning after recombination, where the maximum number of hypotheses in each stack is 4, indicate which hypotheses will be pruned.

Q 5(d) **[4 Marks]**
Provide the fundamental equations of (i) the noisy channel model of SMT, and (ii) the log-linear model of SMT. List *three* frequently used features in log-linear models of SMT.

[End of Question 5]

[END OF EXAM]