DCU

DUBLIN CITY UNIVERSITY

SEMESTER 2 EXAMINATIONS 2012/2013

MODULE: CA446 – Statistical Machine Translation

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

YEAR OF STUDY: 4

EXAMINERS:
Dr Jennifer Foster (Ext:5263)
Dr. James Power
Dr. Michael Manzke

TIME ALLOWED: 2 Hours

INSTRUCTIONS: Answer Question One and two other questions.

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		*Thermodynamic Tables*
	Graph Paper		*Actuarial Tables*
	Dictionaries		*MCQ Only – Do not publish*
	Statistical Tables		

QUESTION 1 ***[TOTAL MARKS: 40]***

Answer EIGHT of the following ten short questions. Each question is worth 5 marks.

1. State three reasons why machine translation (MT) is a difficult problem.

2. Derive Bayes' theorem from the definition of conditional probability.

3. What roles do a *language model* and *translation model* play in statistical machine translation (SMT)?

4. Why is it not feasible to directly compute the probability of an entire sentence in SMT? How do we approximate the probability of a sentence?

5. Briefly explain how the *perplexity* measure is used in language modelling in SMT.

6. Define what is meant by a *parallel corpus* and briefly explain its role in an SMT system.

7. State the rules that are used when building a translation model to decide whether a phrase pair is consistent with a word alignment.

8. Explain what is meant by the term *pruning* and briefly discuss its role in the SMT decoding process.

9. Briefly explain two advantages that human evaluation has over automatic evaluation when assessing the quality of the translations produced by an MT system.

10. Apart from translation *quality*, name two other attributes that are desirable in an MT system and explain why these are important.

[End of Question 1]

QUESTION 2 ***[TOTAL MARKS: 30]***

Answer all parts of the question.

Q 2(a) **[12 Marks]**

Given the following pairs:

The table *table*

An bord *bord*

State what the following translation probabilities will be after *two* iterations of the Expectation Maximisation algorithm and show all the steps followed to arrive at these values:

i. $t(bord|The)$
ii. $t(An|The)$
iii. $t(bord|table)$
iv. $t(An|table)$

Q 2(b) **[10 Marks]**

Explain how the factoring-out trick is used to more efficiently compute word alignments.

Q 2(c) **[8 Marks]**

Explain why word-based translation models in SMT were augmented by phrase-based models. To what extent are the word-based models still useful?

[End of Question 2]

QUESTION 3 ***[TOTAL MARKS: 30]***

Answer all parts of the question.

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

Explain the difference between a unigram, bigram and trigram language model and provide examples which demonstrate the superior modelling power of a bigram language model over a unigram model and a trigram language model over a bigram model.

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

Consider the following data:

Count	Count of Counts
1	4000
2	2200
3	1600
4	800
5	500

Readjust the following three bigram counts using *Good-Turing* smoothing.

Count	Bigram
4	Tea drinker
4	Tea time
3	Tea cup

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

Show, using an example, how the *stack decoding* algorithm works.

[End of Question 3]

QUESTION 4 ***[TOTAL MARKS: 30]***

Answer all parts of the question.

Q 4(a) **[12 Marks]**

Show how you would calculate the BLEU score for the following two translations:

Translation 1: *Salmons swim in river.*
Translation 2: *Fish swim in the river.*
Translation 3: *The salmon swam in the river.*
Reference: *Salmon swim in the river.*

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

Perform a "human evaluation" on Translation 1, Translation 2 and Translation 3 above.

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

Briefly explain why it is better to use more than one reference when carrying out automatic machine translation evaluation.

Q 4(d) **[8 Marks]**

Discuss the limitations of the BLEU metric using the example in (a) above.

[End of Question 4]

[END OF EXAM]