DUBLIN CITY UNIVERSITY

AUGUST/RESIT 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 4 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		*Thermodynamic Tables*
	Graph Paper		*Actuarial Tables*
	Dictionaries		*MCQ Only – Do not publish*
	Statistical Tables		*Attached Answer Sheet*
	Bible		*Exam Paper to be returned with Booklet*

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

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

For a language pair of your choice, provide **three** examples of translational phenomena which demonstrate why MT is a difficult problem, no matter what type of system might be built.

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

Provide the fundamental equations of (i) the noisy channel model of SMT, and (ii) the log-linear model of SMT. Name the different components in (i), and describe their basic function. Demonstrate how the two equations might be equivalent.

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

Why do translation companies build customised solutions for their clients? Give **two** reasons why customised engines are likely to produce better output than a freely available web-based system such as Google Translate.

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

Give **two** reasons why human translators are essential cogs in the MT pipeline. In your opinion, to what extent should human translators be fearful of the impact of MT on their profession?

[End of Question 1]

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

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

Given two SMT systems that translate between English and Irish (in both directions), their performance can be improved by incorporating linguistic knowledge. Suggest **three** ideas to that end. For each idea:

- state its motivation (i.e. the linguistic problem that it is trying to solve),
- how it could be implemented, and
- for which direction (i.e. English-to-Irish or Irish-to-English).

There should be *at least one* idea for each of these phases: (a) pre-processing, (b) decoding and (c) post-processing. There should be *at least one* idea for each of the following types of linguistic knowledge: (i) morphology and (ii) syntax.

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

State the main disadvantage of adding linguistic knowledge to an SMT system. How is this exacerbated for a language like Irish, compared to (say) English?

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

State one pre-processing step that is commonly carried out when translating:

1. from Chinese, regardless of what the target language is;
2. between languages that follow different word orders;
3. from a language with a rich inflectional system into English.

[End of Question 2]

QUESTION 3 ***[TOTAL MARKS: 25]***

Q 3(a) **[4 Marks]**
What is the main reason to use sentence boundaries in *n*-**gram-**based (*n* >1) language models? How are sentence boundaries typically represented in language models?

Q 3(b) **[4 Marks]**
State how a bigram language model would decompose the sentence “They didn 't evaluate their SMT systems .” in order to calculate its probability, both with and without sentence boundaries.

Q 3(c) **[3 Marks]**
Given the sequence of words *x y z*, provide the formulae to calculate the probability of *z* using (i) a bigram language model, and (ii) a trigram language model.

Q 3(d) **[7 Marks]**
Calculate the probability of the last word in the sequence “the green witch” using bigram and trigram language models. The following sequences occur in the training data the number of times shown:

- “the”, 600 times.
- “green”, 10 times.
- “the green”, 5 times.
- “green witch”, twice.
- “the green witch”, twice.
- “witch”, 5 times.

Q 3(e) **[3 Marks]**
Why is it a good idea to use “smoothing” in the context of language modelling?

Q 3(f) **[4 Marks]**
What are the strengths and weaknesses of higher and lower order *n*-gram models?

[End of Question 3]

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

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

Regarding IBM Model 1, what algorithm do we usually use to obtain the word translation probabilities given a parallel corpus? Detail the steps involved in using this algorithm to calculate the IBM Model 1 score, using an example of your choice. .

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

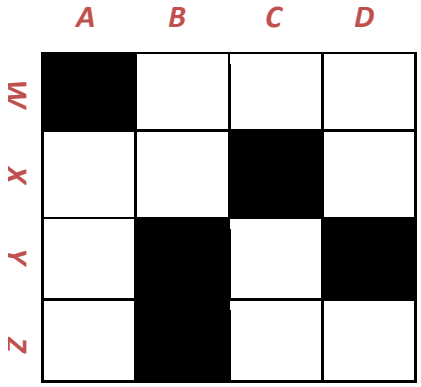
Regarding phrase-based SMT, how do we learn a phrase translation model from a parallel corpus? List the main steps involved, describe the main purpose of each step, and describe what happens at each step and the process involved at each step.

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

How is a "phrase" defined in phrase-based SMT? What are the main differences between a phrase-based translation model and a word-based translation model?

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

Describe the basic rule that we need to follow when we extract phrases**.** List all phrase pairs with the following word alignment based on the rule you have provided.



[End of Question 4]

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

Q 5(a) **[9 Marks]**

Assume the following partial phrase table:

ta	*she*	0.4

shanchang	*likes*	0.3
shanchang	*is good at*	0.7

paobu	*running*	0.6
paobu	*run*	0.4

ta shanchang	*she likes*	0.2
ta shanchang	*she is good at*	0.8

shanchang paobu	*likes running*	0.3
shanchang paobu	*is good at running*	0.7

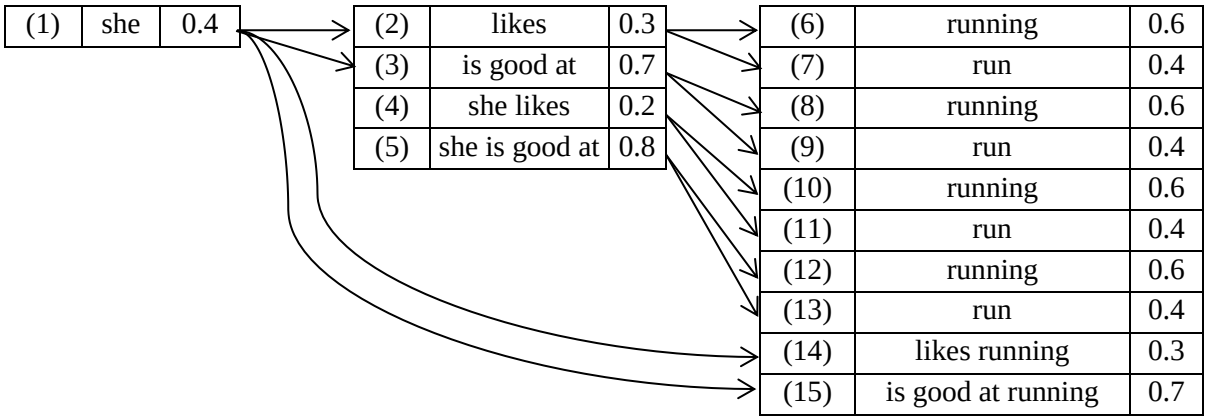
Consider the following input sentence:

ta shanchang paobu

Assume that:

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

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



Given this search diagram, calculate the probabilities for all possible hypotheses (search paths). Furthermore, indicate which hypothesis provides the optimal translation for the given input sentence.

Q 5(b) **[6 Marks]**

Given the above search diagram, indicate (i) which group of hypotheses can be recombined and (ii) which hypothesis should be selected to represent each group.

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

Assume histogram pruning after recombination, where the maximum number of hypotheses in each stack is 2. Indicate which hypotheses will be pruned.

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

Assume threshold pruning after recombination, where the threshold is 0.5. Indicate which hypotheses will be pruned.

[End of Question 5]

[END OF EXAM]