DCU

DUBLIN CITY UNIVERSITY

AUGUST/RESIT EXAMINATIONS 2014/2015

MODULE: CA446 – Statistical Machine Translation

PROGRAMME(S):

CASE	BSc in Computer Applications (Sft.Eng.)
MTT	MSc in Translation Technology

YEAR OF STUDY: 1,4

EXAMINERS:

Professor Qun Liu (Ext:5644)
Dr. Ian Pitt

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	Dictionaries		*MCQ Only – Do not publish*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		*Attached Answer Sheet*

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

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

1. In the distance-based reordering model, the distance for the i^{th} target phrase is: $start_i - end_i + 1$, please explain the meaning of $start_i$ and end_i?

2. Name and briefly explain the two techniques which are used to reduce the search space while decoding.

3. Brevity penalty does not punish candidates which are longer than references. Why?

4. It is said that in BLEU score, unigram precision accounts for adequacy, while other n-gram (n>1) precision represent fluency. Explain why.

5. Explain the idea of backoff smoothing for language models.

6. Assume that the size of the vocabulary is N. What is the number of parameters of an n-gram language model?

7. Draw the dependency trees of the following English sentences: (1) John bought a table with three legs (2) John bought a table with three dollars.

8. State and briefly explain the three components of a statistical machine translation system based on the noisy channel model.

9. Why are bi-directional word alignments necessary when we trained a phrase-based translation model?

10. State and briefly explain two submodels which are not included in IBM model 1 but are included in higher IBM models.

[End of Question 1]

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

Q 2(a) **[16 Marks]**

Given the following pairs:

S_1	S_2
fleur rouge *red flower*	*grande fleur* *big flower*

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:

$t(red|fleur)$
$t(flower|fleur)$
$t(big|fleur)$
$t(red|rouge)$
$t(flower|rouge)$
$t(big|rouge)$
$t(red|grande)$
$t(flower|grande)$
$t(big|grande)$

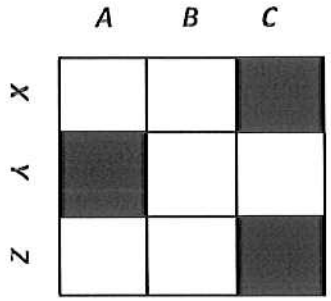
Assuming only one-to-one and one-to-zero patterns are allowed for source-to-target word alignments in these sentence pairs.

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

Assuming any source can be aligned with any number (including zero) of target words and vice versa, please list all the possible word alignments for the sentence pair S_1.

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

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



[End of Question 2]

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

Q 3(a) **[15 Marks]**

Given the following sentences:

<s> *Tom chases Jerry* </s>

<s> *Tom likes Jerry* </s>

<s> *Jerry hates Tom* </s>

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

Q 3(b) **[15 Marks]**

Calculate the probabilities of the following sentences:

<s> *Jerry likes Tome* </s>

<s> *Tome hates Jerry* </s>

Use interpolated smoothing where: $\lambda_{\text{unigram}} = \lambda_{\text{bigram}} = 0.5$.

[End of Question 3]

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

Q 4(a) **[14 Marks]**

Given a partial phrase table:

ta	*he*	0.4

xihuan	*likes*	0.4
xihuan	*likes to*	0.6

youyong	*swimming*	0.2
youyong	*swim*	0.8

ta xihuan	*he likes*	0.2
ta xihuan	*He likes to*	0.8

xihuan youyong	*likes swimming*	0.3
xihuan youyong	*likes to swim*	0.7

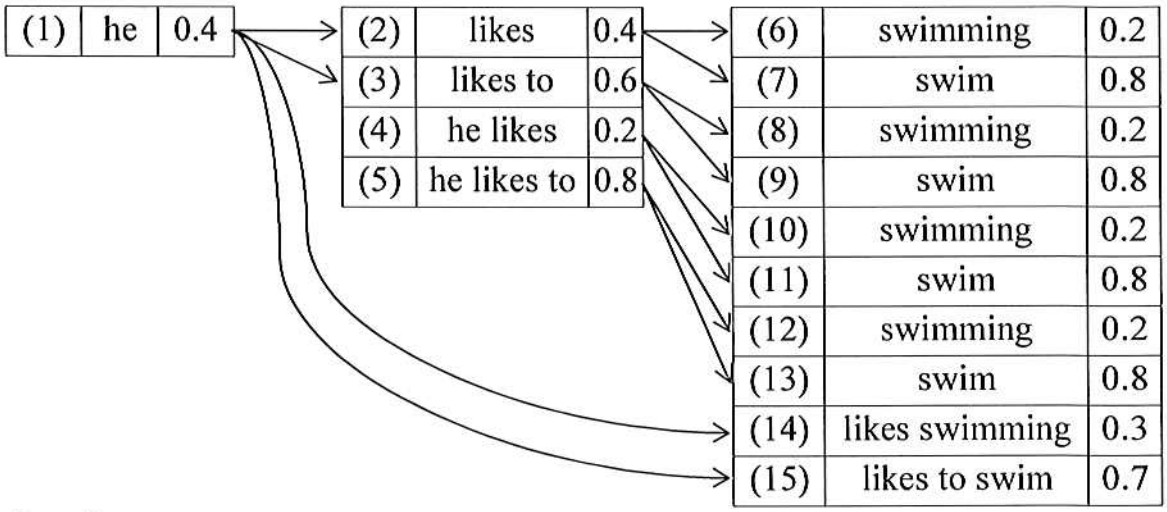
Considering we have the following input sentence:

Ta xihuan youyong

Assuming that:

- Only monotone word order is permitted;
- Language model is ignored.

Then we have the following searching diagram:



Questions:

1) Give the full texts and calculate the probabilities for all the candidates. Indicate which hypothesis provides the optimal translation for the input sentence.
2) Give all the groups of hypotheses which can be recombined and indicate which hypothesis should be selected to represent each group;

3) Assume histogram pruning after recombination, where the maximum number of hypotheses in each stack is 2. Indicate which hypotheses will be pruned;
4) Assuming threshold pruning after recombination, where the threshold is 0.5, please indicate which hypotheses will be pruned.

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

Provide the fundamental equation of the log-linear model of SMT and list three frequently used features.

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

Explain the terms hypothesis and future cost in the context of SMT decoding.

[End of Question 4]

[END OF EXAM]