CS918

THE UNIVERSITY OF WARWICK

MSc Examinations Summer 2017

CS918 Natural Language Processing

Time allowed: 2 hours.

Answer ANY FOUR questions.
Read carefully the instructions on the answer book and make sure the particulars required are entered on each answer book.
Approved calculators are allowed

---

1. (a) Write regular expressions for the following languages:
   i. the set of all alphanumeric strings. Can be either in lowercase or uppercase. [1]
   ii. the set of all uppercase alphabetic strings starting with M. [1]
   iii. the set of all strings that denote prices in dollars. Assume these money expressions are numeric, start with a dollar sign and capture up to two decimal places. [2]

   (b) What is tokenisation? What are some of the issues pertaining to tokenisation? [3]

   (c) i. Describe the minimum edit distance algorithm. [6]
   ii. Compute the edit distance (using insertion cost 1, deletion cost 1, substitution cost 2) of "clear" to "cellar". Show your work (using the edit distance grid). Use the backtrace to obtain an alignment between the two words, corresponding to the minimum cost. [6]

   (d) i. What is the main challenge of word tokenisation in Chinese? [1]
   ii. Describe a baseline algorithm for word tokenisation in Chinese. [2]
   iii. Use the algorithm in ii to tokenise the phrase "Thetabledownthere". What do you observe? [3]

Continued

2. (a) i. What are stems and affixes? [2]

ii. Assuming a morphological analyser, how would you want it to analyse the input "sheep" and "addresses"? [2]

iii. For the following three words provide the lemmas and stems: [3]

- computation
- computers
- living

(b) i. Give a deterministic FSA and a non deterministic FSA for the sheep language \baa+! \ [4]

ii. Describe the recognition algorithm of a non-deterministic FSA. [6]

iii. Give two examples of applications where we would use a FST. [2]

iv. Given the FST in Figure 1 describe what steps it needs to go though to generate the intermediate morphological form corresponding to the surface form "taxes#"? What is the resulting intermediate form? [4]

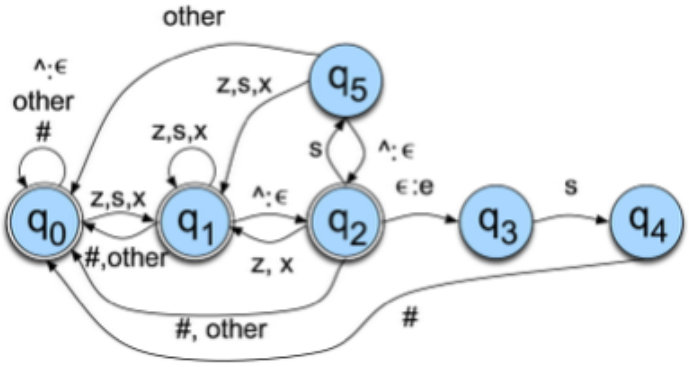

Figure 1

v. Would "taxs#" be accepted as the surface form for "taxˆs#" by this particular FST? Justify your answer. [2]

Continued

3. (a) What are N-gram language models? [2]

   (b) i. Why and how does one perform Laplace smoothing? Show the Maximum likelihood estimates for bigram probabilities before and after Laplace smoothing. [3]

   ii. What are the disadvantages of Laplace smoothing for language modelling? [1]

   (c) Consider the following corpus:
   <s>I am Sam </s>
   <s>Sam I am </s>
   <s>I do not like green eggs and Sam </s>

   i. What is the probability $P(Sam|am)$? [1]

   ii. Write all non-zero bigram probabilities for the above corpus. [5]

   iii. Express Kneser-Ney smoothing for calculating bigram probabilities and calculate the probabilities $P_{KN}(am|I)$, $P_{KN}(Sam|am)$, $P_{KN}(not|do)$ and $P_{KN}(Sam|and)$. Assume a discount factor of 0.5 [7]

   iv. Calculate the probability of the sentence "i want to eat lunch". Give two probabilities, one using Figure 2, and another using the add-1 smoothed table in Figure 3. You can disregard the probabilities $P(i|<s>)$ and $P(</s>|lunch)$ for the purpose of this exercise. [4]

Figure 2

	i	want	to	eat	chinese	food	lunch	spend
i	0.002	0.33	0	0.0036	0	0	0	0.00079
want	0.0022	0	0.66	0.0011	0.0065	0.0065	0.0054	0.0011
to	0.00083	0	0.0017	0.28	0.00083	0	0.0025	0.087
eat	0	0	0.0027	0	0.021	0.0027	0.056	0
chinese	0.0063	0	0	0	0	0.52	0.0063	0
food	0.014	0	0.014	0	0.00092	0.0037	0	0
lunch	0.0059	0	0	0	0	0.0029	0	0
spend	0.0036	0	0.0036	0	0	0	0	0

Figure 3

	i	want	to	eat	chinese	food	lunch	spend
i	0.0015	0.21	0.00025	0.0025	0.00025	0.00025	0.00025	0.00075
want	0.0013	0.00042	0.26	0.00084	0.0029	0.0029	0.0025	0.00084
to	0.00078	0.00026	0.0013	0.18	0.00078	0.00026	0.0018	0.055
eat	0.00046	0.00046	0.0014	0.00046	0.0078	0.0014	0.02	0.00046
chinese	0.0012	0.00062	0.00062	0.00062	0.00062	0.052	0.0012	0.00062
food	0.0063	0.00039	0.0063	0.00039	0.00079	0.002	0.00039	0.00039
lunch	0.0017	0.00056	0.00056	0.00056	0.00056	0.0011	0.00056	0.00056
spend	0.0012	0.00058	0.0012	0.00058	0.00058	0.00058	0.00058	0.00058

   v. What are the common aspects of smoothing on the bigram probabilities in part (c) iii and part (c) iv respectively. [2]

Continued

4. (a) i. What is a vector representation for language and give an example of such a vector representation. [4]

   ii. What are word embeddings and why are they useful? [3]

   iii. What are two principal methods for obtaining word embeddings? Give a short description for each. [5]

   iv. Give one advantage and one disadvantage for each of the above methods. [4]

   v. What is a skipgram model? [1]

   (b) i. Give the one-hot representation of the following two sentences. Assume they are the only sentences in your corpus. No linguistic pre-processing required. [4]

      1. "the world is your oyster."
      2. "enthusiasm moves the world."

   (c) How can one use and evaluate word embeddings? [4]

5. (a) i. Give the definition of a First Order Markov Chain. [5]

   ii. In what context would you use an Hidden Markov Model (HMM) as opposed to a Markov Chain? [2]

   (b) Describe the Forward algorithm. What do we use it for? [6]

   (c) Given the Hidden Markov Model (HMM) in the diagram below, with weather conditions H and C as hidden states, and numbers of ice creams consumed by an individual as observations, compute using the Forward algorithm the likelihood of the observation sequence [3,1,1]. Assume it's equally likely to end the sequence in either H or C. [7]

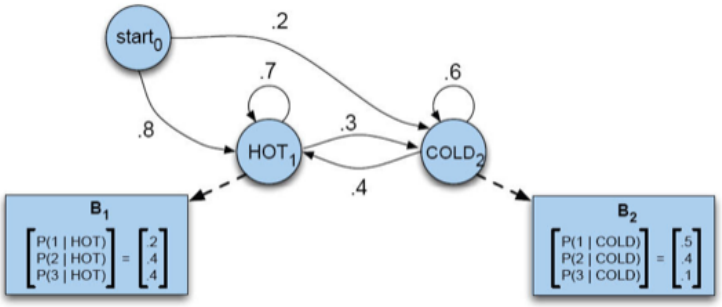

Figure 4

   (d) i. When would one use the Viterbi algorithm? [1]

   ii. When would one use the Forward-Backward algorithm? [1]

   iii. Describe a natural language processing application that has made use of Hidden Markov Models (HMMs) and how. [3]

Continued

6. (a) i. What is Part-of-Speech (POS) tagging and why is it considered as a sequential problem? Give examples to illustrate your answer. [4]

   ii. How do Maximum Entropy Markov Models (MEMMs) differ from Hidden Markov Models (HMMs)? [6]

   iii. What is the problem of bidirectionality in the context of the Viterbi algorithm and Maximum Entropy Markov Models (MEMM)? [1]

   (b) i. Given the Penn Treebank tagset in the appendix, correct the tagging errors in the following POS tagged sentences (one error per sentence): [2]
      1. How/WRB many/JJ students/NNS study/VBP at/IN Warwick/NN ?
      2. The/DT truth/NN is/VBZ more/RBR important/RB than/IN the/DT facts/NNS.
      3. If/IN you/PRP can/VB dream/VB it/PRP, you/PRP can/VB do/VB it/PRP.
      4. Few/JJ people/NN arrived/VBD late/RB for/IN the/DT exam/NN.

   ii. What is parsing? [1]

   iii. What are grammatical constituents? [1]

   iv. What are context free grammars? [1]

   v. What do context free grammars consist of? [1]

   vi. Given the tagset and grammar rules provided in the appendix, create a CFG parse tree for the sentence below. “I will have breakfast on the flight to London.” [4]

   vii. Describe the algorithm for semi-supervised relation extraction using bootstrapping. [3]

   (c) Identify the named entities in the sentences in questions (b) i & vi and identify their type. [1]

Appendix

Figure 5: The Penn Treebank tagset

Tag	Description	Example	Tag	Description	Example
CC	coordin. conjunction	*and, but, or*	SYM	symbol	+,%, &
CD	cardinal number	*one, two, three*	TO	"to"	*to*
DT	determiner	*a, the*	UH	interjection	*ah, oops*
EX	existential 'there'	*there*	VB	verb, base form	*eat*
FW	foreign word	*mea culpa*	VBD	verb, past tense	*ate*
IN	preposition/sub-conj	*of, in, by*	VBG	verb, gerund	*eating*
JJ	adjective	*yellow*	VBN	verb, past participle	*eaten*
JJR	adj., comparative	*bigger*	VBP	verb, non-3sg pres	*eat*
JJS	adj., superlative	*wildest*	VBZ	verb, 3sg pres	*eats*
LS	list item marker	*1, 2, One*	WDT	wh-determiner	*which, that*
MD	modal	*can, should*	WP	wh-pronoun	*what, who*
NN	noun, sing. or mass	*llama*	WP$	possessive wh-	*whose*
NNS	noun, plural	*llamas*	WRB	wh-adverb	*how, where*
NNP	proper noun, singular	*IBM*	$	dollar sign	*$*
NNPS	proper noun, plural	*Carolinas*	#	pound sign	*#*
PDT	predeterminer	*all, both*	"	left quote	' or "
POS	possessive ending	*'s*	"	right quote	' or "
PRP	personal pronoun	*I, you, he*	(	left parenthesis	[, (, {, <
PRP$	possessive pronoun	*your, one's*	)	right parenthesis	],), }, >
RB	adverb	*quickly, never*	,	comma	,
RBR	adverb, comparative	*faster*	.	sentence-final punc	. ! ?
RBS	adverb, superlative	*fastest*	:	mid-sentence punc	: ; ... – -
RP	particle	*up, off*			

Figure 6: CFG rules

S -> NP VP
NP -> PRP
NP -> NNP
NP -> DT Nominal
NP -> Nominal
NP -> NP PP
Nominal -> Nominal NN
Nominal -> NN
VP -> VB
VP -> MD VP
VP -> VB NP
VP -> VP NP PP
VP -> VP PP
PP -> TO NP

End