THE UNIVERSITY OF WARWICK

MSc Examinations: Summer 2016

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 alphabetic strings. Can be either in lowercase or uppercase. [1]
   - ii. the set of all lowercase alphabetic strings starting with a. [1]
   - iii. all strings that start at the beginning of the line with an integer and end at the end of the line with a word. Assume that a word contains only alphabetic characters [2]

   (b) What are the (three) text pre-processing steps required by all NLP tasks? Give a brief description of each and the challenges each of these involves. [6]

   (c) i. What is the minimum edit distance algorithm and what is it used for? Why is it useful? [3]
   - ii. Compute the edit distance (using insertion cost 1, deletion cost 1, substitution cost 2) of "leda" to "deal". 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) What is the MaxMatch algorithm used for? Describe the algorithm.What are its advantages and disadvantages? [6]

2. (a) i. What is morphological analysis? What are morphemes and what are two main types of morphemes? [2]

ii. Give the expected output of a morphological analyser given the input "women" and "understands". [2]

iii. What is lemmatisation? What is stemming and how does it differ from lemmatisation? Give an example where the outcome of lemmatisation is different from the outcome of stemming. [3]

(b) i. What is a finite state automaton (FSA) and how can we formally specify an FSA? [5]

ii. Describe the algorithm for a Deterministic FSA. [5]

iii. What is the main difference between a FSA and a finite state transducer (FST)? [1]

iv. When would we want to use a FST in natural language processing? [2]

v. Given the FST in Figure 1 describe what steps it needs to go though to generate the intermediate morphological form of the surface form "buses#". [4]

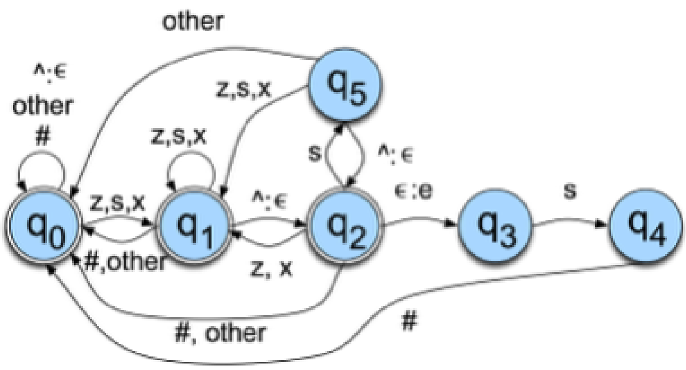


Figure 1

vi. What is the intermediate morphological form of the surface form "buses#"? [1]

3. (a) What is smoothing and why do we use it in language modelling? [2]

   (b) Describe Interpolated Kneser-Ney smoothing for bigrams. [6]

   (c) Consider the following corpus:
   <s>I am Will </s>
   <s>Will I am </s>
   <s>I will eat anything </s>

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

   ii. What is the continuation probability of the word “Will”? [1]

   iii. What is the interpolated Kneser-Ney smoothing of $P(Will|am)$? Assume the fixed discount is 0.5. [2]

   iv. Write all non-zero trigram probabilities for the above corpus. [5]

   v. What do you observe for the majority of them? [1]

   vi. How can you get more realistic estimates of probabilities for these trigrams? [1]

   vii. Calculate the probability of the sentence “i want chinese food”. Give two probabilities, one using Figure 2, and another using the add-1 smoothed table in Figure 3. Assume $P(i|<s>) = 0.25$ and $P(</s>|food) = 0.68$. [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

   viii. Which of the two probabilities you computed in the previous exercise is higher, unsmoothed or smoothed? Explain why. [2]

4. (a) i. What is a one-hot word vector? [1]
   ii. What are word embeddings? [1]
   iii. Give two reasons why word embeddings are useful. [2]
   iv. Name two groups of methods for obtaining word embeddings. Give one advantage and one disadvantage of each of them. [6]
   v. What is a continuous bag of words model? [1]
   vi. What is the input to word2vec and what is the output? [3]

   (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 man saw a cat."
      2. "the cat caught a bird".
   ii. Give two disadvantages of the one-hot representation. [2]

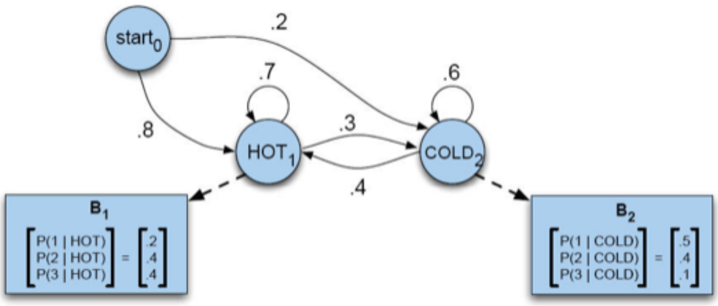
   (c) i. What do we mean when we say that Naïve Bayes is a generative model whereas the Maximum Entropy classifier is a discriminative model? Illustrate this with each of the algorithm's objectives. [4]
   ii. When would you use Naïve Bayes for text classification? [1]

5. (a) i. Give the definition of a Hidden Markov Model (HMM). [5]

ii. How does a HMM differ from a Markov chain? [2]

(b) Describe the Viterbi algorithm. [6]

(c) Given the 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 Viterbi algorithm the most likely sequence of hidden states that would produce the observation sequence [1,3,1]. [7]



(d) i. Describe the deleted interpolation algorithm. [4]

ii. What is the deleted interpolation algorithm used for? [1]

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. Name two different sequential approaches to POS tagging. [1]

iii. Describe the main differences between the two approaches and two implications stemming from this difference. [6]

iv. What is the label bias problem (use an example) and how can it be mitigated? [4]

(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 do/MD I/PRP get/V to/TO Warwick/NN ?
2. I/PRP hope/VBP you/PRP have/VB rooms/NN available/JJ.
3. This/DT room/NN is/VBZ too/JJ noisy/JJ.
4. Can/VB you/PRP give/VB me/PRP another/DT room/NN?

ii. Given the relations and rules provided in the appendix, create a dependency parse tree and a CFG parse tree for the sentence below. “I would like an evening flight to New York.” [4]

(c) i. What is information extraction? [1]

ii. What is named entity recognition? [1]

iii. What is relation extraction? [1]

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

Appendix

Figure 11: 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 12: CFG rules

```
S -> NP VP
NP -> PRN
NP -> NNP
NP -> Det Nominal
NP -> NP PP
Nominal -> Nominal NN
Nominal -> NN
VP -> VB
VP -> MD VB
VP -> VB NP
VP -> VB NP PP
VP -> VB PP
PP -> TO NP
```


End

Figure 13: Dependency relations

- *root* - root
- *dep* - dependent
    - *aux* - auxiliary
        - *auxpass* - passive auxiliary
        - *cop* - copula
    - *arg* - argument
        - *agent* - agent
        - *comp* - complement
            - *acomp* - adjectival complement
            - *ccomp* - clausal complement with internal subject
            - *xcomp* - clausal complement with external subject
            - *obj* - object
                - *dobj* - direct object
                - *iobj* - indirect object
                - *pobj* - object of preposition
        - *subj* - subject
            - *nsubj* - nominal subject
                - *nsubjpass* - passive nominal subject
            - *csubj* - clausal subject
                - *csubjpass* - passive clausal subject
    - *cc* - coordination
    - *conj* - conjunct
    - *expl* - expletive (expletive "there")
    - *mod* - modifier
        - *amod* - adjectival modifier
        - *appos* - appositional modifier
        - *advcl* - adverbial clause modifier
        - *det* - determiner
        - *predet* - predeterminer
        - *preconj* - preconjunct
        - *vmod* - reduced, non-finite verbal modifier
        - *mwe* - multi-word expression modifier
            - *mark* - marker (word introducing an *advcl* or *ccomp*
        - *advmod* - adverbial modifier
            - *neg* - negation modifier
        - *rcmod* - relative clause modifier
        - *quantmod* - quantifier modifier
        - *nn* - noun compound modifier
        - *npadvmod* - noun phrase adverbial modifier
            - *tmod* - temporal modifier
        - *num* - numeric modifier
        - *number* - element of compound number
        - *prep* - prepositional modifier
        - *poss* - possession modifier
        - *possessive* - possessive modifier (*'s*)
        - *prt* - phrasal verb particle
    - *parataxis* - parataxis
    - *goeswith* - goes with
    - *punct* - punctuation
    - *ref* - referent
    - *sdep* - semantic dependent
        - *xsubj* - controlling subject

End