MANIPAL INSTITUTE OF TECHNOLOGY
MANIPAL
(A constituent unit of MAHE, Manipal)

M. Tech (CSE) INTERNAL EXAMINATIONS SEP 2019

TEST – 1

Program Elective II	**Natural Language Processing**	**CSE 5010**
17/02/2020	2.30:3.30 PM	Max. Marks: 15

1) Considering Autocomplete application of our smart phone, distinguish natural language processing applications from data processing systems and describe the processing of knowledge. (2)

2) Write true or false for each of the sentences. If true, justify with an example for each.

   i) Derivation changes the category but does not change the meaning (0.5)
   ii) Derivation does not change the lexical category but changes the meaning (0.5)
   iii) The affix *–ing* in inflectional form and two different derivational forms changing the lexical category after attaching derivational affix *–ing*. (1.5)
   iv) Two morphological parsed output form for word *book* (0.5)

3) For following orthographic rules, write Chomsky-Halle rule notation. Draw three level tapes and mark items c and d of the rule clearly on the intermediate tape. (1*3=3)
   i) E-insertion rule for word *foxes*
   ii) E-insertion rule for word *watches*
   iii) K- insertion rule for word *panicked*

4) Justify giving pictorial representations and examples: (1+1)
   i) Time complexity of Minimum distance algorithm - dynamic programming solution vs. brute force recursive algorithm
   ii) Applying Minimum edit Distance algorithm to a source string X[1…m] and a target string Y[1…n] but reached the end of X substring and there are yet k characters left in Y.

5) For a corpus containing below text
   a b c d e e d c b a a b c d c e e e with vocabulary size=5 containing words {a, b, c, d, e}, show all the processing steps and display the results in the appropriate table format (round decimal values to two decimal places) (5)
   i) Calculate unigram count for vocabulary size.
   ii) Calculate MLE(Maximum Likelihood Estimation) count and MLE probabilities for the bigrams
   iii) Calculate adjusted count C* and Add-one probabilities for bigrams

1) Natural language processing applications vs. data processing systems: use of knowledge of language

Predictive text- Autocorrect, autocomplete, and predictive text are common place on our smartphones. Autocomplete and predictive text are similar to search engines in that they predict things to say based on what we type, finishing the word or suggesting a relevant one. And autocorrect will sometimes even change words so that the overall message makes more sense.

1. Phonetics and Phonology — The study of linguistic sounds
2. Morphology —The study of the meaningful components of words
3. Syntax —The study of the structural relationships between words
4. Semantics — The study of meaning
5. Pragmatics — The study of how language is used to accomplish goals
6. Discourse—The study of linguistic units larger than a single utterance

2)

i. Derivation changes the category but does not change the meaning (0.5)
Derivational suffix *–al* : The noun *Person* changes to *personal* adjective
Derivational suffix *–ly:* adjective *quick* into an adverb *quickly*

ii. Derivation does not change the lexical category but changes the meaning (0.5)
All English prefixes are derivation and they do not change the lexical category of the word but does change the meaning.

iii) Derivational: Verb + *-ing* = Noun - I watched the dancing in the room.
Derivational: Verb + *-ing* = Adjective - The dancing peacock
Inflectional: Verb + *-ing* = Verb - The peacock is dancing

iv) book: (book +N +SG) or (book +V)

3)

i) E-insertion rule for word *foxes*
e added after *-s*, *-z*, *-x*, *-ch*, *-sh*, before *–s*

- *a -> b / c___d*
  rewrite a as b when it occurs between c and d.
- *ε -> e / {x, s, z} ^___s#*
  Insert e on the surface tape when the lexical tape has a morpheme ending in x, s or z and the next morpheme is –s.

The E-insertion rule

Lexical	{	f	o	x	+N	+Pl			}
Intermediate	{	f	o	x	^	s	#		}
Surface	{	f	o	x	e	s			}

ii) E-insertion rule for word *watches*

e added after -s, -z, -x, -ch, -sh, before -s

$\epsilon \rightarrow e / \{ch, sh\}^{\wedge} \_ s\#$

watch +V +PL

watch ^ s#

watches

iii) K- insertion rule for word *panicked*
Verb ending with vowel + -c add -k

panic +V +PAST

panic ^ ed#

panicked

$\epsilon \rightarrow k / \{Vc\}^{\wedge} \_ ed\#$

where V: {a, e, i, o, u}.

4)

i) Minimum Edit Distance: dynamic programming solution Time: O(nm)
Minimum Edit Distance: Bruteforce Recursive : $O(3^m)$

ii)

source → MAN ****

target → MANIPAL

K = 4

Insert k = 4 characters.

CASE 1: We have reached the end of
either x or y

If substring x is empty, then we return all
remaining characters of substring y as y
and the cost of this operation is equal to
number of characters left in substring y.

('', 'ABC') ==> ('ABC', 'ABC')
(cost = 3)

If substring y is empty, then we delete all
remaining characters of x to convert it
into substring y. The cost of this
operation is equal to number of
characters left in substring x.

('ABC', '') ==> ('', '') (cost = 3)

5)

a b c d e e d c b a a b c d e e e e vocabulary size=5 containing words {a, b, c, d, e}

i) Calculate unigram count for vocabulary size and N=18.

a	b	c	d	e
3	3	3	3	6

ii)

Calculate MLE(Maximum Likelihood Estimation) count and MLE probabilities for the bigrams

	a	b	c	d	e
a	1	2	0	0	0
b	1	0	2	0	0
c	0	1	0	2	0
d	0	0	1	0	2
e	0	0	0	1	4

	a	b	c	d	e
a	1/3=0.3	2/3=0.67	0	0	0
b	1/3=0.3	0	2/3=0.67	0	0
c	0	1/3=0.3	0	2/3=0.67	0
d	0	0	1/3=0.3	0	2/3=0.67
e	0	0	0	1/6=0.17	4/6=0.67

iii) Calculate adjusted count C* and Add-one probabilities for bigrams

Add-one count

	a	b	c	d	e
a	2	3	1	1	1
b	2	1	3	1	1
c	1	2	1	3	1
d	1	1	2	1	3
e	1	1	1	2	5

Adjusted count C*

	a	b	c	d	e
a	2*3/(3+5) =0.75	3*3/(3+5) =1.13	1*3/(3+5) =0.38	1*3/(3+5) =0.38	1*3/(3+5) =0.38
b	2*3/(3+5) =0.75	1*3/(3+5) =0.38	3*3/(3+5) =1.13	1*3/(3+5) =0.38	1*3/(3+5) =0.38
c	1*3/(3+5) =0.38	2*3/(3+5) =0.75	1*3/(3+5) =0.38	3*3/(3+5) =1.13	1*3/(3+5) =0.38
d	1*3/(3+5) =0.38	1*3/(3+5) =0.38	2*3/(3+5) =0.75	1*3/(3+5) =0.38	3*3/(3+5) =1.13
e	1*6/(6+5) =0.55	1*6/(6+5) =0.55	1*6/(6+5) =0.55	2*6/(6+5) =1.09	5*6/(6+5) =2.73

Add-one probabilities

	a	b	c	d	e
a	2/(3+5) =0.25	3/(3+5) =0.38	1/(3+5) =0.13	1/(3+5) =0.13	1/(3+5) =0.13
b	2/(3+5) =0.25	1/(3+5) =0.13	3/(3+5) =0.38	1/(3+5) =0.13	1/(3+5) =0.13
c	1/(3+5) =0.13	2/(3+5) =0.25	1/(3+5) =0.13	3/(3+5) =0.38	1/(3+5) =0.13
d	1/(3+5) =0.13	1/(3+5) =0.13	2/(3+5) =0.25	1/(3+5) =0.38	3/(3+5) =0.38
e	1/(6+5) =0.09	1/(6+5) =0.09	1/(6+5) =0.09	2/(6+5) =0.18	5/(6+5) =0.45