PAGE NO.:
DATE: 8th April

Linguistic terminologies

Morphology (structured) (word)
↓
Syntax (phrase/sentence)
↓
Semantics (meaning)
↓
Pragmatics (contextual meaning/information)
↓
Discourse (higher than the paragraph level)

- Turin test - 3 actors (2 humans, 1 machine)
- Eliza

False positive (Type 1) - Precision / Accuracy
False negative (Type 2) - Recall / coverage

- Antagonistic efforts
1. minimizing FP
2. minimizing FN

So, we measure F1-score (HM of Precision & Recall)

Every word is token but every token is not a word
San Francisco - 2 tokens, but 1 word

- UNIX for poets

PAGE NO.:
DATE:

Tokenization - 1 word per line

state-of-the-art → 1 word

Edit Distance

Minimum edit distance

I	N	T	E	*	N	T	I	O	N
|	|	|	|		|	|	|	|	|
*	E	X	E	C	U	T	I	O	N
1	2	2	-	1	2	-	-	-	-

(Deletion) (Substitution) (Insertion)

Total cost = 8 units

Alignment in Computational Biology

Named entity extraction - extract proper nouns from text

How to find Min Edit Distance?

Word type	Tokens
- Whatever that goes into vocabulary	- all the occurences of a word in a text

Minimum Edit as Search

for 2 strings
X of length n
Y of length m

We define D(i,j)

Dynamic Programming
→ Tabular computation of D(n,m)
→ solving problems by combining solution to sub proble

Bottom - up
→ Compute D(i,j) for small i,j
→ And Compute larger D(i,j)

~~Lev~~ Levenshtein
→ Initialization
D(i,0) = i
D(0,j) = j

→ Recurrence Relation

PAGE NO.:
DATE:

for each $i = 1 \ldots M$
  for each $j = 1 \ldots N$

$$D(i,j) = \min \begin{cases} D(i-1,j) + 1 \\ D(i,j-1) + 1 \\ D(i-1,j-1) + 2; & x(i) \neq y(j) \\ 0; & x(1) = y(1) \end{cases}$$

Page 107

Assignment - 1 Deadline : Friday

1) fill in the table with min edit distance values + back trace pointers (arrows)

2) Explain the process (also back trace pointers)

Performance

Time $O(nm)$

Space $O(nm)$

Back trace $O(n+m)$

www.unicode.org

PAGE NO.:
DATE:

Weighted Edit Distance

→ Some words are more likely to be misspelled.

Unicode and Multilingual computing

Software Localization

Locale — Date
Currency
Days of a month

Internationalization
Localization
Globalization

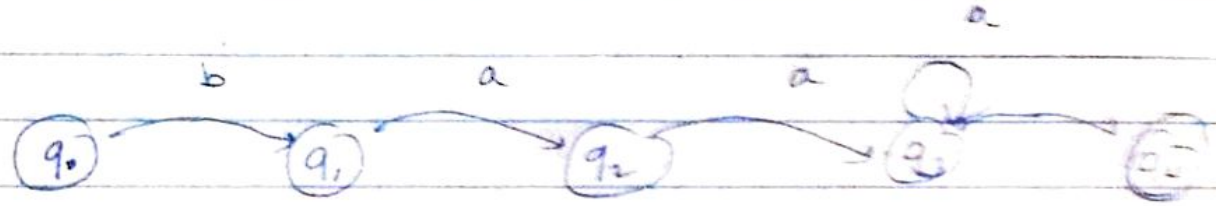
Text processing not possible - Preeti, Kantipur, Annapurna

Glyph - graphical representation of a character

Nepali NLP

Information Retrieval
Extraction

Finite State Automata



DFSA
determined path for an input

NFSA
multiple paths

Morphology and Finite State Transducer

Study of word forms

Surface form

Lexical form

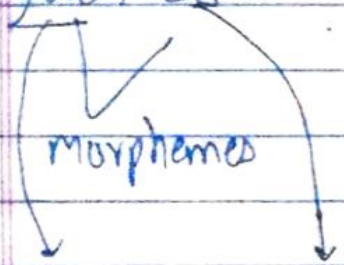
nds

PAGE NO.
DATE:

Smallest form of a word is morpheme

foxes

= fox + es



free morpheme — bound morpheme

Morphological parsing vs Stemming

- break down a particular word into free morpheme and bound morpheme

foxes
→ fox + es

- IR
- Machine translation
- Spell checking

- only concerned about the stem (not about bothers)
- breaking down a particular word into its corresponding stem or root word

eg sing, sung, sang, singing, sings → sing

stem - free/main morpheme
affixes - bound morpheme

prefix, suffix, infix, circumfix

PAGE NO
DATE:

1. Inflectional morphology

N

fox + es → Noun to noun
↓ → no change in class/parts of speech

foxes
N

2. Derivational morphology
→ result in different class

V

compute + ing → eg. Verb to noun
↓
computing (N)

GNP - (Nepali)
Gender Number Person

English - NP only
He eats
She eats

HMM Hidden Markov Model POS tagging

maximize $P(\hat{t_1} | \hat{w_L})$

Window size.

NN VB

Sequence Labelling.

POS tagging approaches

Classification learning

and most frequent word in English Language

PAGE NO.:
DATE:

Chapter 8 Word Classes and Part-of-Speech Tagging

is lowest level of syntactic analysis

Parts of speech (word class / morphological class / lexical tag)

English Word Classes

1. Closed class
   - relatively fixed membership eg: prepositions
2. Open class
   noun, verb, adjectives, adverbs

Parts of speech tagging

1) Rule-based taggers
2) Stochastic taggers
   - used a training corpus to count the probability

Brill tagger - hybrid of both