Cohesion

* high level text docs.
* Eg: Academic articles.
    1) Abstract
    2) Intro
    3) Methodology
    4) Result
    5) Conclusion.
* Automatic detection ✗.
* Sol: algos for discourse Segmentation is used.
* The unsupervised discourse Segmentation algos are based on concept of cohesion.
* Lexical cohesion, 2 units unrelated by relations btw words.
  For ex: Peel, core and slice — pears & apples.
* sep by hyphen

Reference Resolution

→ " Lami is 10th std. She is a very good singer and participated in many music programs. The performance of this 16 yr old is also excellent in academics

__/__/____

→ * Mention – one person – Laxmi
* Linguistic expressions She, her denote one individual – Referene.

→ * defination : Ref resolution is a task to determine what entities are refered to by which linguistic expressions.
* The refered entity is called as Referene Eg: Laxmi

→ Anaphora : Reference to an entity which is prev introduced in discourse.

→ Anaphora : The referening expressing Eg: Pronoun, She and 16 years old

→ Reference Resolution – 2 tasks

a) Co reference resolution – It is the task to find out referening expressions refering to same entity. RE = RE

b) Pronominal Anaphora Resolution : It is a task to find out antecedent for single pronoun. Eg : antecedent of She is Laxmi.

→ (a) is subtask of (b).

→ Algos used.
Viterbi algo, viterbi algo, Log Linear

N Gram model

→ Please turn your homework.

Predict next word — Over, In.
→ known as word prediction — n gram models.
→ next word predicted from n-1 words.
→ → word seq probability is X.
→ ⏟ difficult to compute.

So decompose: chain rule probability.

~~P(w) = P(w...w~~

Markov's Assumption: n-1 preceding words.
(n-1)th order Markov model

n=2, bigram → 2 words → please turn (2) turn your.
n=3, trigram → 3 words seq. → please turn your

Language Model Eval.

Evaluate performance: 1) Include it in apps.
2) check performance of app.
↓
extrinsic evaluation

Intrinsic : quickly evaluate the improvement.

Coverage rate : % measure of n-gram in test set.
OOV can't be handled.

Perplexity : fits the data - best - among 2 models

MAXIMUM LIKELIHOOD ESTIMATION AND SMOOTHING.

max likelihood criteria + parameter smoothing :=
ngram prob.

* The max likelihood estimate

$$P(w_i | w_{i-1}, w_{i-2}) = \frac{C(w_i, w_{i-1}, w_{i-2})}{C(w_{i-1}, w_{i-2})}$$

$C(w_i, w_{i-1}, w_{i-2})$ is count of trigram of w_{i-1}, w_{i-2}, w_i in TD.

Smoothing - flatten peaks in n-gram predict by redistributing prob mass.
'0' estimate replaced by non zero.

Backoff - common smoothing tech.

P_{BO} for w_i given w_{i-1} w_{i-2}.

$$P_{BO}(w_i | w_{i-1}, w_{i-2}) =$$

$$\begin{cases} d_c P(w_i | w_{i-1}, w_{i-2}) & \text{if } C > T. \\ \alpha(w_{i-1}, w_{i-2}) P_{BO}(w_i | w_{i-1}) & \text{other} \end{cases}$$

$C \rightarrow$ count of (w_i, w_{i-1}, w_{i-2}).
$d_c \rightarrow$ discounting factor applied to higher Order distribution.

Language Model Adaption

* certain domains → data req to train a model — insufficient
* less data available.
* so tune a lang model, ported to new domain

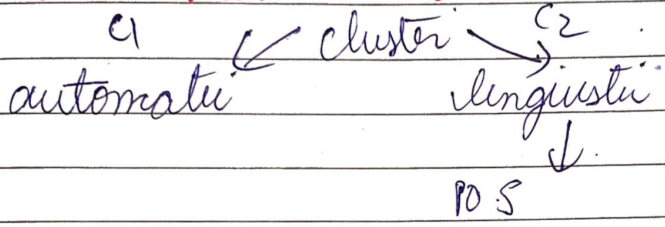
{ model interpolation — small data, 1 domain, 1 generic model.
  cluster based.
  ↳ Seymour Rosenfeld.

1) dynamic self adaption → triggers models.
2) unsupervised adaption → speech recognition apps.
   ↳ MAP.
   ↳ LSA PLSA.

prob distribution over word sequence.

Ex: Speech recog, OCR, HWR, MT.

Class Based Model



assumption :- words are conditionally independent of other words in a current word class.

Class Based Bigram Model.

$$P(w_i / w_{i-1}) \approx \sum_{c_i, c_{i-1}} P(w_i/c_i)\ P(c_i/c_{i-1}, w_{i-1})\ P(c_{i-1}/w_{i-1})$$

$$= \sum_{c_i, c_{i-1}} P(w_i/c_i)\ P(c_i/c_i)\ P(c_{i-1}/w_{i-1})$$

c_i = class
w_i = word.

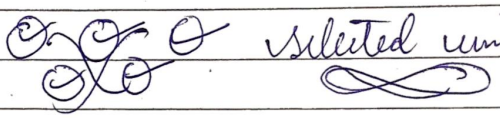
Use :- perplexity can be reduced.

c_i is independent of w_{i-1} given c_{i-1}.

class - contains - more than one word.

$$P(w_i / w_{i-1}) = P(w_i/c_i)\ P(c_i/c_{i-1})$$

Variable length lang Model.

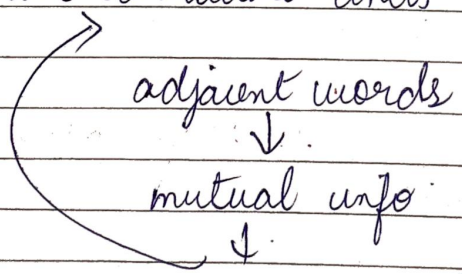
* words - seperated by - white space.
* SLM - vocabulary units - predict next word → invariable. fixed length based on history.
* units merge → generate → Variable length lang model.
* unit.  selected units = merged.

merged units → correspond to → freq observed. short phrases.

→ added to lang Model Vocab.

Eg: write off is diff from write-off.

Probable candidate units selection from.



actual candidate selection — Greedy Iterative algo.

NLP:-

Cross disciplinary → linguistics, CS, AI
Relation → digitalized computing devices & human lang.
NLP field → deals with — designing & prog computational devices.
Reason → to process & analyse large amts of data.
Forms → writing, speech, singing.
data → highly unstructured.
Use → read text, hear speak, interpret

Spectrum of NLP:- / why challenging.

Spectrum → wide.
Linguistic diversity → 1000+ spoken lang.
3 main lang → a) Indo European. (Eng).
b) Sino Tibetan (Chinese)
c) Afro Asiatic (Arabic).
Delimited → In Orthography → whitespace & punctuation.
Word forms → langs use, need not change much
Other hand → highly sensitive abt word forms
Gender → for some lang → content impacts Gender of noun.
NLP show structure → diff kinds & complexity
word blocks → not viable approach.

✡ Morphology in / imp to understand words

→ Indicative blocks :- words.
Words ⟶ tricky to define
due to → a) ambiguity
b) contextual meaning

→ Knowing how to work ⟶ allows a) syntactic
b) semantic understanding

→ Process of Understanding words : involves.
a) morphology - Var FEF (PCS).
b) word structure - Read - phonology, write o the
c) linguistic expression :- Semantics, lexicology, etymology

a) * Study of variable forms & functions
* Syntax — arrangement of words into
a) phrases.
b) clauses
c) sentences.

b) * constraints due to pronunciation :
described by ~~for~~ phonology.
* conventions for writing — Orthography

c) Evolution of words : a) etymology
b) lexicology
c) semantics.

→ Morphological parsing — Technique for discovering of word structure.
used to : a) Identify words
b) model diff internal structure with
gp & LC.

Components of words

Words → smallest linguistic units
Morphemes → minimal parts of words (meaning)
Morphems are
a) graphemes (writing)
b) phonemes (speaking)

Noun	Noun+S(pl)	Noun+S(pos)	Pronounciation
thrush	thrushes	thrush's	iz
toy	toys	toy's	z
block	blocks	block's	s

Words & their Components: Tokens, Lexemes, Morphemes, Typology

Tokens

Tokens — Syntactic words.
Eg:- I don't want to buy this product.

Linguists → 2 syntactic words → do not called tokens.

Token → independent role.
reverted back to normalized form

Other words —treated as→ Independent Single token

In Eng:
limited set of cases → tokenization
↖ can be applied ↗ normalization

Other lang: treated in less literal manner

__/__/____

Lexemes

alternative forms that can be expressed for a given word.

Such sets → lexemes (OR) lexical items.

Constitute → lexicon of lang.

Divided by → lexical categories :
- a) verb
- b) noun
- c) adjective
- d) adverb etc

Lemma → citation form of lexeme.

Inflect → plural mic from singular morse.

Derive → reverse
- ↳ reverir
- ↳ reception

Morphemes

Morphs :→ Structural components that associate the properties of word forms

Eg :- dis-agree-ment-S.

agree → free lexical morpheme.

Other elements → bound grammatical morph

Morph 1 + Morph 2 ∞ → phonological + Orthographic changes.

alternative forms → ALLOMORPHS.
Eg :- a) past tense morphemes.
b) plural morphemes.

Typology

Morphological typology → divides languages in groups.

Outline of Typology → based on Quant Relatn b/w
a) words.
b) morphems
c) features

a) Isolating or Analytical lang →
include no or few words.
that would
comprise more than 1 morphene.

b) Synthetic lang → combine → more morphems.
→ in one word.

agglutinative → single func at time → Korean Japanese

fusional. → feature-per-morpheme ratio higher than 1. → Arabic, Latin, German

word formation process

concatenative
↓
linking morphs or morphemes.

Nonlinear
↓
merge non sequentially
↓
apply tonal morphs.

✡ Issues & challenges-

a) Irregularity

Morphological parsing

numerous → Irregularities
↓
is all about
↓
forms & structures
↓
not described by
↓
prototypical linguistic model.

gives rise to → Ambiguity
↓ is
Indeterminacy of Interpretation
↓ of
expressions of lang.
↓
other Issues

Syncretism
Systematic ambiguity

Morphological Modelling

provides ← faces prob.

↓

productivity & creativity.

↓

Generalisation
abstraction

English

→ past forms
↓
-ed a -t

→ Irregular verbs
↓
diff forms
↓
dep on lang of words
↓
Eg :- blow, blew, blown

Ambiguity

Word forms → look same
→ diff fun or meanings
→ called homonyms
→ Eg (kind, rug, right, rose)

Ambiguity presence → morphological process
lang processing at large

Complete Disambiguation :- MP is not considered

Morphological Disambiguation

↓

Analii

↓

encompass not only

↓

resolution of structural components of word and their actual morpho - syntactic prop but also

↙ ↘

lokinization normalization

Eg :- Inverting Sandi during lokinization can provide multiple sols to the prob of ambiguity

Na Asatah Vidyate Bhavah —

no

The unreal has no existance

Syncretism :- Morphologi Phenomenon

↙ ↘

some words mod classes.

↘ ↙

show instances of

↓

Systematic homonymy

Homonymy

occurs due to

neutralization → is all alit → Syntactic irrelevance

uninflectedness → is all alit → being unresponsive to a feature → that is → Syntactic relevant

English ÷ gender category → syntactically neutralized → pronouns

The diff b/w he & she, him and her are only Semantic

Productivity

Members of corpus → word types

Long utterances → word form → word token

Distribution of words → 80/20 rule → law of vital few

Negation → productive Morphological operation

English: dis, non, un

Ex: Acc to wiki — Googol.
— google.
(google is misspelled here.

Now both of these words entered English lexicon.

☆ MORPHOLOGICAL MODELS.

INTRO:

Morphological parsing is a process by which word forms of a lang are lass with corresponding linguistic descriptions.

Many approaches → designing ee imp MM.

Domain specific lang → created → useful in implementing theoretical prob with minimal prog effort.

Other approaches → ✗ domain specific

Dictionary Lookup

dictionary → data structure
enables precomputed word analysis.

Implemented → a) Lists
b) BST
c) Tries.
d) # tables

set of ass → b/w → word forms & their descriptions

generative power — × exploited.

Problem → how the associated annotations are constructed?
→ how informative & accurate they are.

Finite State Morphology

FSM models → directly compiled → FS transducers

eg → 1) Xerox finite State Tool.
2) LEX tool.

Set of possible seq → accepted by transducer — defines → I/p lang

Set of possible seq → emitted by transducer defines → o/p lang

In FS Computational Morphology →
i/p word forms → surface strings
o/p " → lexical

Relation R.
denote it with Σ
set of all sequences over some set of symbols Σ
so that DOMAIN & RANGE of R.
are subsets of Σ
Now R is a func mapping i/p string to set of o/p strings

$R: \Sigma \rightarrow \{\Sigma\}$
$R: String \rightarrow \{string\}$

Morphological Operations & processes
In human lang
expressed in
finite state terms

Problem — reduplication of words found in many natural lang

Uses — applied to MM of isolating & agglutinating lang
— FST — limited extent — morphological analyser or generators

Unification based Morphology.

UBM approaches.

Inspired by
a) The formal linguistic framework like HPSG.
b) Lang for lexical knowledge representation like DATR

UBM concepts & methodologies → closely connected → logical programming → PROLOG.

High level approaches → expressed by → more appropriate data structure → which can include complex values → unlike FSM.

Morphological parsing 'P' associates
lexical forms ϕ
alternative structured content Ψ.

$$P : \phi \rightarrow \{\Psi\}$$
P : forms → content.

For MM, WF are best captured by RE

Linguistic content is best described through feature structure.

Unification → key operation → by which feature structure can be merged into more informative feature structure

MM of this kind → typically formulated → logic programs.

Unification is used to solve the system of constraints imposed by model

Functional Morphology

FM → defines it modes → using → POFM & & Type Theory.

treats → morphological operation & process → as pure Math func.

Organises → linguistic, abstract } into } diff types of values & type classes.

Func morphology → useful for → functional Morphology

Implementation a) MP.
b) NL generation.
c) lexicon Browsing etc.

I = inflection
D = derivation
L = lookup.

I : lexeme → {parameters} → {forms}.
D : lexeme → {parameters} → {lexeme}
L : content → {lexeme}

Morphology Induction

Problem of — discovering & inducing word structure without human insight.

Automatic acquisition of M & L Info, even if not perfect → bootstrapping → classical MM too.

Several challenging issues → deducing words → structure just from the forms & their context.

__/__/____

Finding Structure of Docs

Intro:

* words form sentences.
* Sentences form paragraphs.
* Automatic extraction of structure of docs help in:
  a) parsing
  b) machine Translation
  c) Semantic Role labelling.
* Sentence boundary annotation is important for human readability of output of ASR system.
* Chunking: the i/p text provides better org & indexing of data.

Sentence Boundary Detection:-

* automatically segmenting a seq of word tokens into sentence units.
* Sentence — start — capital
  — end — .
* In addition:
  capital letters — distinguish proper nouns.
  periods — used in abbr & no's
  Other punctuation marks: used inside proper names

Examples

Dr. → doctor / drive.
I spoke with Dr. Tina.
My house in on Mountain Dr.

Problem — SS — written text in SMS —
poorly used punctuation

Input → OCR/ASR → translate
errors of handwritten text then
finding of the System boundaries
must handle those errors as well.

OCR ↔ confuse → ,
ASR → ~~p~~ marks → punctuation marks.
Prob1 — Dialogue Act Segmentation
Dialogue acts are better defined for
conversational speech using one of
Markup standards

Dialogue Act Markup in Several Layer (DAMSL),
Meeting Recorder Dialog act (MRDA).
are 2 ex of such markup standards.

As to these std, "Okay no problem"
consists of 2 sentential units
Okay & no problem.

Prob 2 :- Code Switch.
Sentences from multiple lang by
multi lingual speakers.

affect - terminalities

NLP

Components of NLP

(I) Natural Language Understanding:-
- → Lexical Ambiguity (word level).
- → Syntactical " (sentence level).
- → Referential Ambiguity (Referring issue using pronouns)

(II) Natural Language Generation:
- → Text planning (Retrieve Relevant content from KB).
- → Sentence planning (words, meaningful phases, setting tone)
- → Text Realization (Mapping sentence planning into sentence structure)

Steps in NLP

① Lexical Analysis :- we have to analyse the structure of words. The collection of words & phrases in lang is lexicon of lang

② Syntatic Analysis :- We use parsing for analysis of word. Although have to arrange words in a particular manner. That shows the relationship b/w words

③ Semantic Analysis :- It describes a dictionary meaning which is meaningful. In the task domain, mapping syntatic structure & object

④ Discourse Integration

The meaning of any sentence depends upon the meaning of previous sentence. In addition, it brings the meaning to immediately

preceding sentence.

(5) Pragmatic Analysis

In this step, data is interpreted on what is actually meant. Although we have to discuss aspects of lang which require real-world knowledge

Steps in NLP

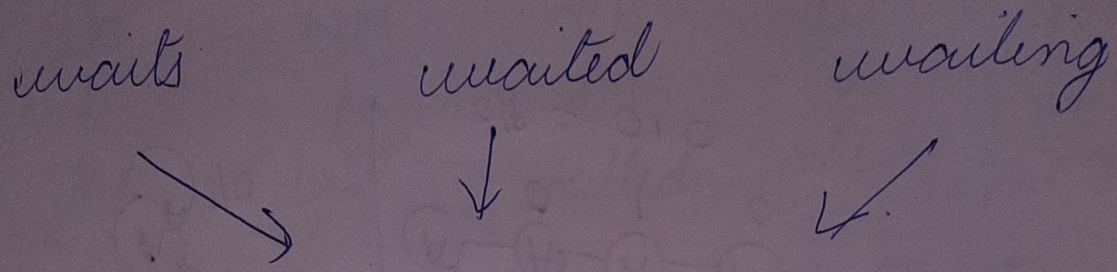
(1) Tokenization :

- Cutting big sentences into small tokens.
- Example : Welcome to my channel.

[Welcome] [to] [my] [channel]

(2) Stemming :

- Normalize words into base or root forms



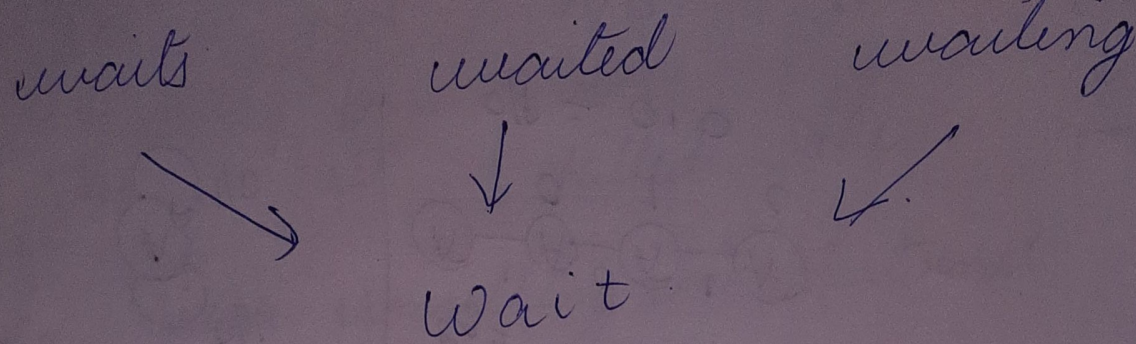
(3) Lemmatization

* Group together different inflected forms of words called Lemma.
* Somehow similar to Stemming, as it maps several words into one common root.
* Output of lemmatization is a proper word.

Eg:- gone, going and went → Go

(4) POS tags

- POS stands for Parts of Speech tags.
- It indicates how a word function in meaning as well as grammatically in a sentence



(3) Lemmatization

* Group together different inflected forms of words called Lemma.
* Somehow similar to stemming, as it maps several words into one common root.
* Output of lemmatization is a proper word.

Eg:- gone, going and went → Go

(4) POS tags

- POS stands for Parts of Speech tags.
- It indicates how a word function in meaning as well as grammatically in a sentence

eg

DT	NN	VB	DT	NN.
↓	↓	↓	↓	↓
The	cat	killed	the	rat

eg: 'Google' something on Internet.
(NN, VB)

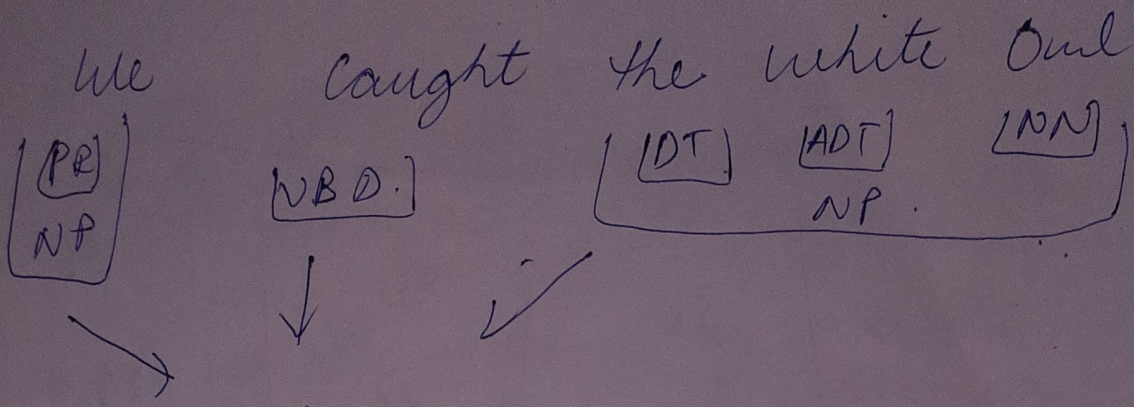
(5) Name Entity Recognization

- It is a process eg detecting name entity such as: person, Organisation, location, Quantities etc.

eg: Google's CEO Sundar (person) Pichai introduced the new Pixel 3. at New York Central Mall.
(Org) (location)

(6) Chunking

Picking individual pieces of info & grouping them into bigger pieces.



- This helps getting insight & meaningful info from the text.

Applications of NLP

- Chat Box
- Speech Recognization
- Machine translation
- Spell checking
- Keyword Searching
- Advertisement matching

Words & their Components

* Words is the most lang are the smallest unit that can form a complete utterance by themselves.

* Three important terms which are integral parts of words are.

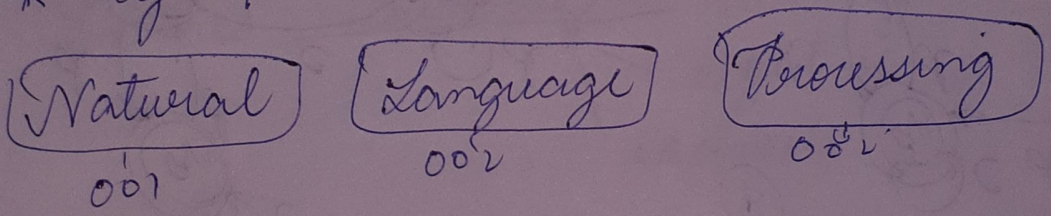
(a) Phonemes: Units of sound in spoken lang.
Graphemes: Smallest unit of written lang.
Morphemes: minimal part.

* In words & components we have Tokens, lexemes, Morphemes, Typology

- Tokens:
* Cutting big sentences into smaller words & encoding each word.
* This

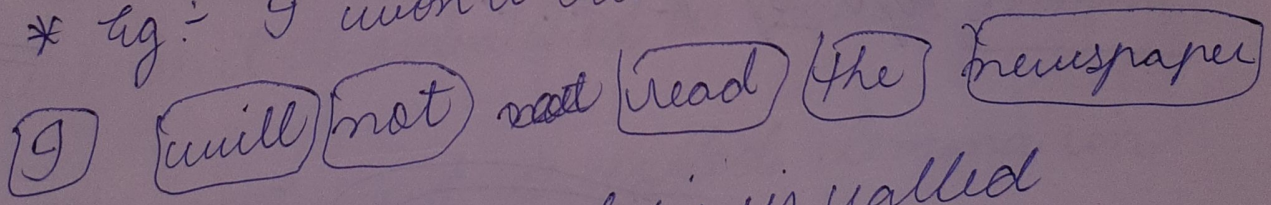
Tokens

* Tokenization is essentially splitting a phase, sentence, paragraph or an entire text document into smaller units & encoding them.
* Each of this smaller units are called tokens
* Eg Natural Language processing.



3 tokens

* Eg:- I won't read the newspaper



* This type of analysis is called tokenization & normalization.

Lexemes

* There are a of alternative forms that can be expressed for a word. Such sets are called as lexemes or lexical items.
* ~~They~~ Lexemes constitute to lexicon of a lang.
* Lexemes — lexical categories → noun, verb, adjective, adverb.

* The citation form of lexeme by which it is identified is called lemma.
* The root word is called lemma in lemmatization process.
* Eg In case of regular & irregular verbs we derive forms from verb to regular

Morphemes

* The structural components that associate the properties of words forms are called morphs.
* The morphs that by themselves represent some aspect of the meaning of a word is called morphemes of some function.
* Eg:- disagree-ments where agree is a free lexical morpheme & other elements &, are bounded grammatical morphemes.
* Morphs when they interact with each other undergo additional phonological & orthographical changes.
* These alternant forms are called allomorphs
* Ex:- the past tense morphemes, plural morphemes

Typology

* Morphological typology divides lang in. groups.
* Isolating lang include no words that would compromise more than 1 morpheme.
* Synthetic lang can combine more morphemes in one word
* Concatenative lang linking morphemes one after the other
* Non-linear lang allowing simultaneous components to merge non-sequentially

Issues & Challenges

- Irregularity
- Ambiguity
- Productivity :-

a) Irregularity :-

* Morphological parsing procedures generalisation & abstraction in the world of words.
* In English the general past form forms by adding -ed or -t (accepted & built)
* The irregular verbs in English tend to take different forms in the past or in present participle depending on the origin of word

Verb	Past tense	Past Participle
blow	blew	blown
break	broke	broken
buy	bought	bought

Ambiguity

* Words forms that look the same but have distinct functions or meanings are called homonyms (example: kind, ring, right, rose).
* Ambiguity is present in all aspects of morphological processing
* Inverting of Sandhi during tokenization can provide multiple solutions to the problem of Indian lang (na asatah vidyate bhavah) which means the unreal has no existence)

Productivity

* An important question to be answered - Is the inventory of words in a language finite or infinite?
* In one view lang can be seen as simply a collection of utterances actually pronounced or written.

* This data set can be linguistic corpora - a finite collection of linguistic data.
* If we consider language as a system, we discover structural clemies like recursion, iteration that allow to produce an infinite set of concrete linguistic utterances
* This process is called morphological productivity.
* Eg Au to maki "googol" is a made-up word denoting a no "one followed by hundred zeros.
* The name of the company google is actually a misspelled word.
* Now both of these words entered the English lexicon.

Date :

Page No.

Q) Parsing Natural Languages.

eg(1) He wanted to go for drive in movie
He wanted to go for drive in the country

* There is a natural pause between drive & in in the 2nd sentence.

* The gap reflects an underlying hidden structure.

* Parsing provides a structural description that identifies such a break in the intonation.

eg(2) The cat who lives dangerously had nine lives.

* A text to speech system needs to know that the first instance of a word is verb & the second is noun.

* This is an instance of POS tagging problem.

* Another important application where parsing is important is text summarizatio

Date :

Page No.

(4.3) Open borders imply increasingly racial fragmentation in EUROPEAN COUNTRIES.

* In the above sentence, the words capitalized phase can be replaced with other phrases without changing the meaning of sentence.

* Open borders imply increasingly racial fragmentation in the countries of Europe

* " " " " ,
" " "

European countries

* " " " " "
" " "

Europe

* In NLP, Syntatic parsing is used in many app's like

* Statistical Machine Translation
* Information extraction from text collection
* Language summarization
* Grammar correction in text
* knowledge acquisition from language.

Date :

Page No.

Q2)) Tree banks: A Data Driven Approach to Syntax.

* Parsing recovers info that is not explicit to the input sentence.
* Parsers requires some additional knowledge beyond the input sentence that should be produced as output.
* We can write down rules of the syntax of a sentence as CFG.
* Here we have CGF which represents a simple grammar of transitive verbs in English

(Ans)

S → NP VP.
NP → 'John' | 'pockets' | DN | NP PP.
VP → VNP | VP PP
V → wrought.
D → a
N → 'shirt'
PP → P NP.
P → 'with'

* The above CFG can produce the syntax analysis of sentence like: John wrought a shirt with pockets.

* Parsing the previous sentences gives.

(S (NP John) (VP (VP (V brought) (NP (D a) (N Shirt))) (PP (P with) (NP pockets))))	(S (NP John) (VP (V bought) (NP (NP (D a) (N shirt)) (PP (P with) (NP pockets)))))

* Writing a CFG for syntactic analysis of NLP is problematic.
* A simple list of rules does not consider interactions between different components in the grammar.
* Listing all posible syntactic constructions in a language is difficult task.
* It is difficult to exhaustively list the lexical properties of words.
* This is a typical knowledge acquisition problem
* Other problem is that rules interact with each other in combinatorially explosive ways

Ex2 :- Noun phrases as binary branching tree

N → NN (Recursive Rule).
N → 'natural' | language | 'processing' | 'book'.

* For input 'natural language processing' the recursive rules produce two ambiguous parsers.

(N (N (N natural) (N language)). (N processing))	(N (N natural) (N (N language). (N processing)))

* For CFG's it can be proved that the no. of parsers obtained by using recursive rule n times are Catalan number (1, 1, 2, 5, 14, 42, 132, 429) of n.

$$cat(n) = \frac{1}{n+1}\binom{2n}{n}$$

* For the input "natural language processing book" only one of the 2 parse obtained by using CFG is correct.

(N (N (N (N natural) (N language)). (N processing)). (N book)).

* This is the second knowledge acquisition problem.
* We not only need to know the rules but also analysis.
* The construction of a tree bank is a data driven approach to syntax analysis
* It allows us to address both the knowledge acquisition bottlenecks in one stroke
* A Treebank is a collection of sentences.
* Where each sentence is provided by complex syntax analysis
* The syntax analysis for each sentence has been judged by a human expert.
* A set of annotation guidelines is written to before the annotation process to ensure
* Treebank contains annotations of syntactic structure of large set of sentences.
* Supervised ML techniques are used to train the parser from the training data extracted from tree banks.
* The first knowledge acquisition problem is addressed by providing syntactic analysis directly instead of grammar.

* The second knowledge acquisition problem is solved as for each sentence in a treebank the most feasible syntactic analysis is provided.
* Using supervised ML techniques are used to learn scoring functions for all possible syntax analysis

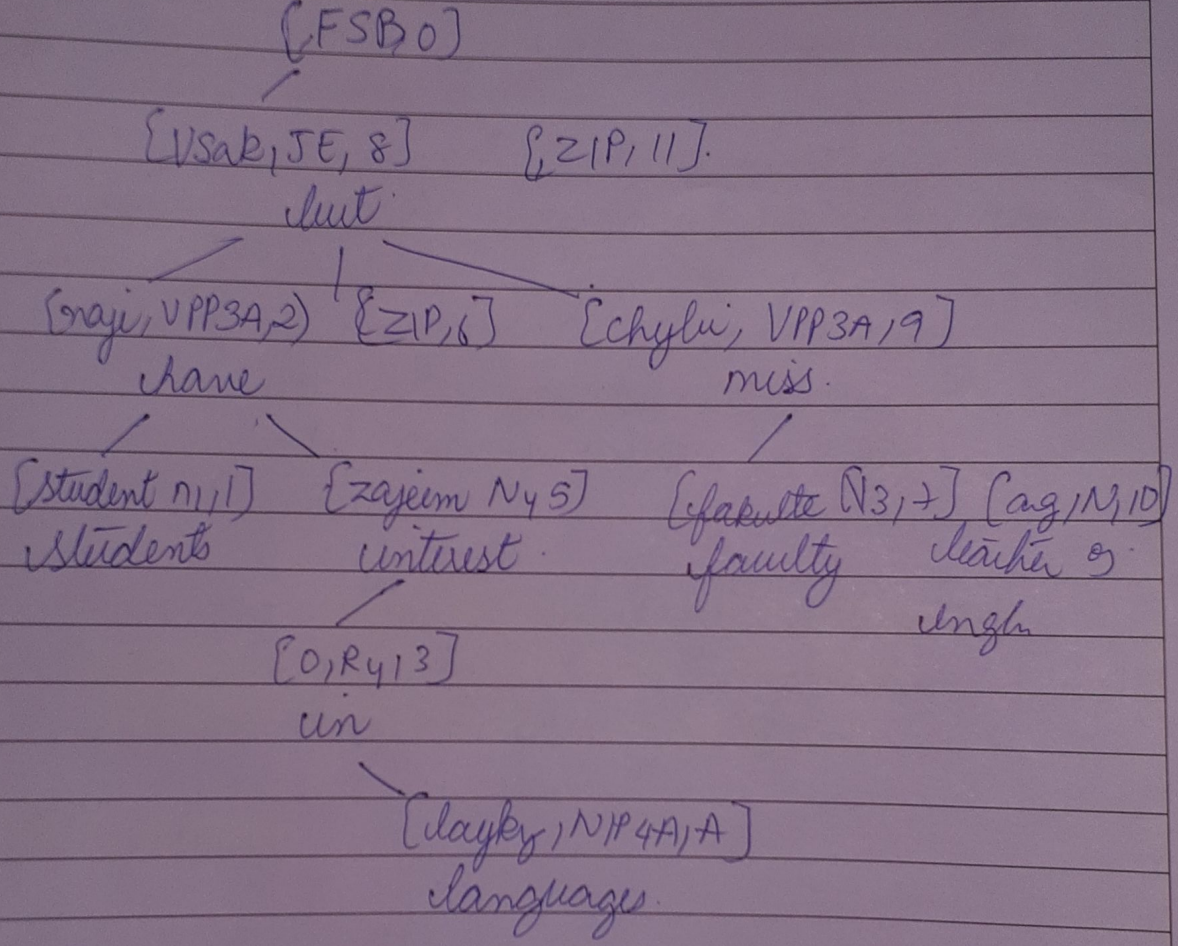
Q3 Representation of Syntactic Structure
(OR)
Syntax Analysis using dependency graphs.

* Dependency graphs connects head of a phrase to its dependents
* Definition: In dependency graph
  a) nodes are words of input sentences.
  b) arcs are binary relations from head to dependent.
* It is often assumed that all words except one, have a syntactic head.
* It means graph will have a tree with single independent node as root
* In labelled dependency parsing the parser assigns a specific type to each dependency relation holding b/w head word & dependent word

Date :

Page No.

* In dependency tree, 0 index is used to indicate the root symbol



* This is an example of a dependency graph syntax analysis for a Czech sentence taken from Prague Dependency Treebank.
* Each node in the graph is a word, its parts of speech & not the word in sentence
* Eg :- [fakulte, N3, 7) is the 7th word in the sentence with F, tag N3 which

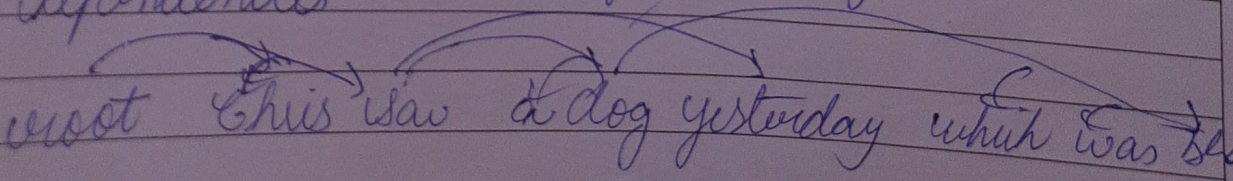
also tells us that it has dative case.

* The node [ZSB, 0] is the root node in dependency tree.
* The english equivalent is provided to each node.
* It is observed that dependency analysis make minimal assumption about syntatic structure for avoiding annotation of hidden structure like empty elements to represent missing arguments of predicate.

The dia: The students are interested in languages, but the faculty is missing teachers of English.

Projectivity

* Projectivity is the constraint on semantic analysis due to effect of linear order of words on dependencies b/w words.
* The ex shows the english sentence with requirements of crossing dependencies.



Parsing Algo

4) Shift reduce parser

2. Data structure stores
   * stack - symbols of grammer.
   * Input buffer - Input string

* $ specifies bottom of the stack & end of input buffer

* Actions of shift reduce parser
  a) shift
  b) reduce
  c) accept
  d) error

* There are 2 types of conflicts.
  a) Shift reduce conflict.
  b) reduce reduce conflict

Ex: E → E+T | T.
T → T*F | F.
F → (E) / id.
id * id.

Stack	Input buffer	Action
$	id * id $	Shift
$id	* id $	Reduce by F → id
$F	* id $	Reduce T → F
$T	* id $	Shift
$T*	id $	Shift
$T*id	$	Reduce F → id

$T*F	$	Reduce T→T*F
$T	$	Reduce T→E
$E	$	Accepted

Shift reducing parsing algorithm

* Start with an empty stack.
* The buffer must contain the input string.
* Exit with success if,
  a) the top of the stack contains start symbol of the grammer.
  b) and if the buffer is empty.
* Choose one the following 2 steps.
* If choice is ambiguous, choose one based on an oracle.
* Shift a symbol from the buffer onto the stack.
* If the top r symbols of the stack are $\alpha_1, \ldots \alpha_r$ then replace the top r symbols with A, left hand side of non terminal 'A'
* Exit the failure if no action can be taken in previous steps.
* Else go to

Explain

Hyper graphs & chart parsing (OR) CYK Algo

* It is a membership algo.
* It checks whether a string 'abbb' is a valid member of full CGF.
* CYK is applicable for CNF only.

CNF: $A \to BC$ (OR) $A \to a$

* It is universal (applicable for all grammer)
* Time complexity = $O(n^3)$.
* Space complexity = $O(n^2)$

Eg abbb.
1234

$S \to AB$.
$A \to BB \mid a$.
$B \to AB \mid b$.

4

Use make 4×4 matrix
Omit half

	4	3	2	1
1	S,B	S,B,A	S,B	A
2	S,B,B	A	B	
3	A	B		
4	B			

2,4.
22 44
S,B. B

11 44.
A B

ROUGH

S,4
(33) (44)
B B.
A

(13)
11 33
A B.
S,B.

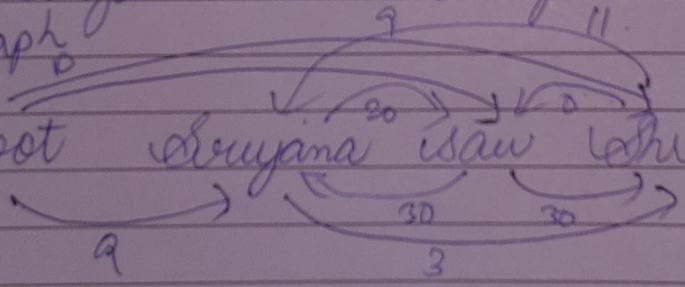
1234

1 2.
(11) (22)
A. B.

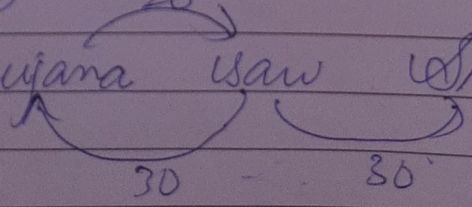
2 3
(22) (33)
B B.
A.

6). Minimum Spanning Tree.

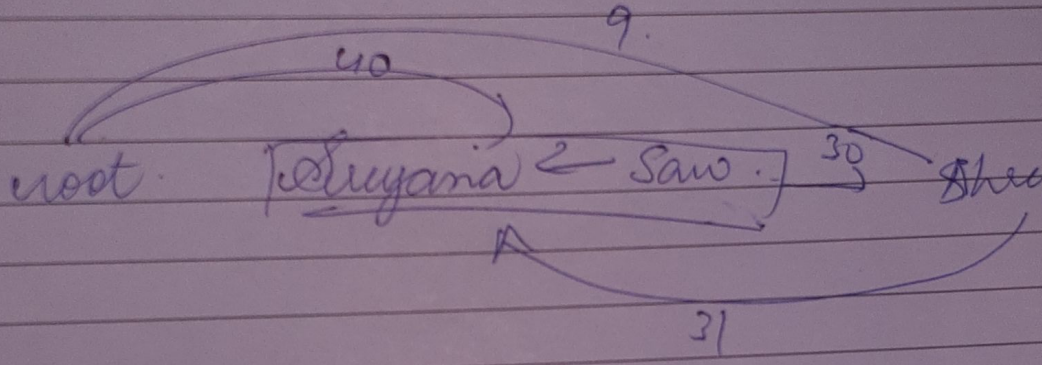
* The minimum Spanning Tree corresponds to the optimum branching problem in directed graphs which are roots and does not have cycles.
* The basic prequisite is all the dependency links b/w the words must have score.
* To Consider input string as Krupa saw Shreya.
* graph



* The scoring function is used to assign weights to edges.
* The algo starts finding the incoming edge with highest score.



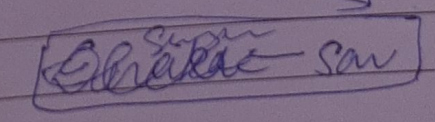
root → Saw → Sreyana = 10 + 30 = 40.
root → Sreyana → Saw = 9 + 20 = 29
Shreya → saw - Sreyana 0 + 30 = 30.
Shreya → Oruka - Saw. 11 + 20 - 51.



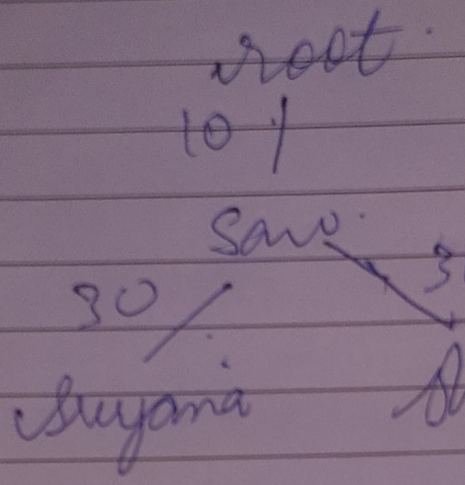
* MST is applied recursively.
* Least incoming edges to each vertex is found.
* In next iteration

root → Shreya → [Sreyana → Saw] 9 + 31 = 40.
root → [Sreyana - Saw] → Shreya will have.
40 + 30 = 70.

the least nodes chosen are



And we get highest scoring dependency parse as



(Q7) Models for Ambiguity Resolution in Parsing

a) Probabilistic Context Free Grammer.
b) Generative Models for parsing
c) Discriminative Models for Parsing

Date :

Page No.

28) Semantic parsing

a) Intro

* It's an imp phase in NLP.
* Any document is comprised of sets of sentences.
* These sentences are formed by the words arranged according to the rule based grammar as per constructs in a particular language.
* To extract meaning out of this syntactically arranged sentences is the object of semantic parsing.
* Parsing is analysing a text into logical syntactic components
* And Semantics is study of meaning.
* The information pieces are identified and related in Semantic parsing.
* The meaningful parts are identified in the text
* Then they are transformed into suitable data structure for higher level task accomplishment

b) Types of Semantic parsing :-

* Semantic parsing is about identifying the meaning.
* There are 2 approaches

Page No.

(1)

* Consider domain specific applications like travel reservation, gaming simulation etc.
* Such meaning extraction can be done based on restricted limits of particular domain
* In this approach based on query, output is generated from the meaning representation.
* This approach is known as deep semantic parsing
* little or no scope of reusability

(2)

* The second approach is not domain specific
* It is more generic.
* In this approach, the task of meaning representation is divided into small parts.
* It is known as shallow semantic parsing.
* These parts are responsible for capturing small manageable components that represent meaning
* High reusability

c) Semantic Interpretation

* Semantic Interpretation is joining of different components
* The components define the meaning representation of text.
* This representation when fed to computer can be further processed for computational manipulations by search for any applications.
* Semantic parsing is the part of Semantic interpretation

Components of Semantic interpretation

Following are the major components in this process.

① Structural Ambiguity.

* Any sentence is represented by its Syntactic structure.
* Syntax and Semantics are closely related to each other.
* We consider that the Semantic interpretation is based on underlying Syntactic structure
* So Syntactic processing is the first step in Semantic interpretation

② Word Sense

* Many times in any language a same word is used in different meaning depending upon world knowledge
* Eg:- word bank represents a money bank or it can also means river bank.
* Due to inherent intelligence of language vocabulary by humans, it is not a difficult for them to understand the meaning of word uttered by speaker or author.
* Consider the foll examples
  - ⓐ Go to bank to withdraw money.
  - ⓑ Fetch some water from river bank.
* It's easy task for humans to disambiguate the meaning of word bank in above sentences.
* But for machine this word sense disambiguation is a challenging task & it plays an imp role in semantic interpretation.
* ~~But for machine,~~

③ Entity and Entity resolution.

* Any text consists of various entities.

falling in different categories like person name, locations, quantities etc.
* Identification of these entities is the major task in semantic interpretation system.
* Named Entity System (NER): is a subtask of information extraction for locating & classifying named entities.
* Ex- [Divyana]$_{person}$ bought 100 shares of [Infosis]$_{org}$ in [2022]$_{time}$.
* Another important task is of co-reference resolution.
* Co-reference occurs when 2 or more expressions in text refer to same person.
* Eg:- Divya said she will sing. In this sentence, proper noun Divya & she
* This task is also under information extraction
* And it is a major component in semantic interpretation

(4) Predicate Argument Structure
* After finishing all the above mentioned tasks of word sense disambiguation, NER co-reference sol.
* Next task is to identify what is a

role of entity in particular event
* This is known as resolving argument structure of predicate in a sentence
* It is typically an identification of who did what to whom, how where, when etc

(5) Meaning Representation :-

* This is the last step in semantic interpretation
* Also called deep representation
* It involves building meaning representation
* Which can be used by algorithms for various applications
* Research is still going on in the area of general purpose
* Much study has been done in domain specific apps.
* Eg :- Consider a Geoquery domain the example query can be represented as which a river is longest
Ans (x1, longest (x, river (x1)))

Date :

Page No.

Q10)) System Paradigms. (OR)
Approaches for practical simple mentation of semantic interpretation.

* Based on diversity of languages & levels of granularity ...
* The main constraint is posed by lack of data availability of lang.
* Due to these constraints, some successful approaches for practical implementation fall in these 3 categories

(1) System Architecture.

a) Knowledge Based :
The systems are designed using predefined set of rules for finding the solution to the particular problem.

b) Unsupervised :
Sol to a particular application is found by minimum sum of ~~constraint~~ incoherent

c) ~~Semi~~ Supervised :
* The technique involves model training by application of diff ML algo.
* Feature functions are created to create functions which are used to predict labels to handle unseen data

(2) Scope

(a) Domain dependent - System designed for specific domain such as travel reservation.

(b) Domain independent - It includes general techniques.

(3) Coverage

a) Shallow : In this approach, intermediate representation is generated to be used by machine.

b) Deep : Through this approach representation, which is created, directly used by machine

(Q11) Word Sense 221.