

119 (1 lecture missing)

5/10/19

of info in any point of time (depends how many communities we want)

→ SNA can also be based on Content Analysis. (till now → structural analysis)  
Ex. can we infer something from a coll<sup>n</sup> of tweets

→ Here, we're looking at English level. But it can be extended to other languages

→ NLP :-

POS Tagging [Part of Speech] → small part of NLP

→ To find meaning of content : we've to look at words, coll<sup>n</sup> of words → sentences, coll<sup>n</sup> of sentences and so on.  
→ Looking at word → Noun / Pronoun / Adjective / ...

↳ Have to classify in automated version

→ Most algo are statistical in nature with accuracy ~ 97%

→ Have to input the context & issues faced while doing POS

Tagging & Content Analysis.

→ In 100 B.C., they defined 8 classes :

1. Noun

2. Verb

3. Pronoun

4. Prepositions

5. Adverb

6. Conjunction

7. Participle

8. Article.

→ Basis of most of European language

Objective of this : to understand word, we need to know its class.

→ Now, we have almost 45 tags (these 8 classes are further classified)

### Judging Ages

- \* 1. Rule based : People have read articles & labelled the words manually & classify accordingly. They've made their own rules.

Dictionary → 1. Tells class of word.

- 2. ~~Set of rules~~ How you treat a word in that sentence is decided by set of rules (disambiguated)
  - No learning : learning is done by human being

### 2. Statistical<sup>s</sup> based :

Have state of how this word is used previously.

Ex: [A] student

→ most probably an article - determiner ; (1) telling no. states: This classifier will tell it to be a determiner

◦ Learn from past history in a statistical way & disambiguate words accordingly.

1) → Not very popular (we've very good ML models in 2)

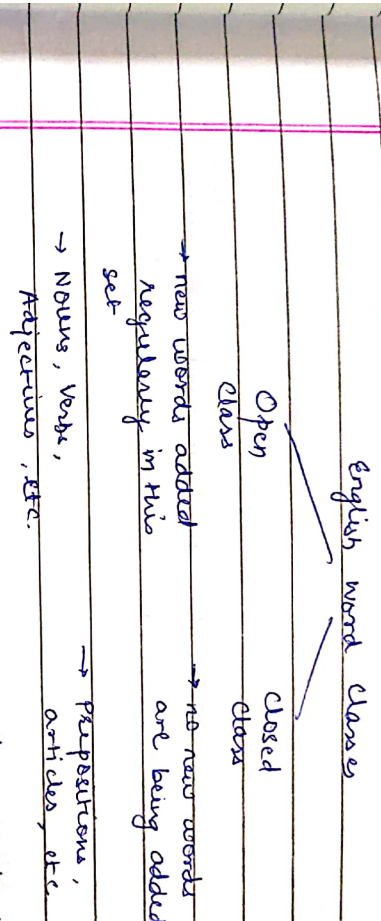
2) → we've HMM.

We'll be looking at Rule-based approach.

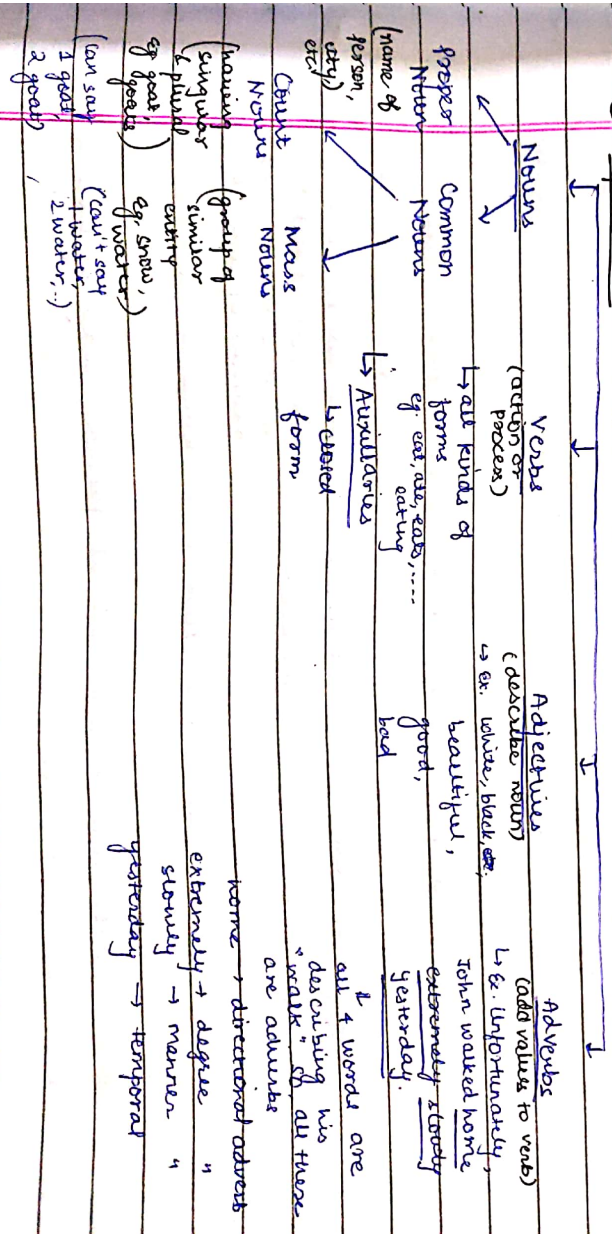
→ we're comparing in number we don't have any adjectives

### challenge behind building Rule based approach

→ Need to know English structure



① Open class → some of class's are made to understand contents effectively (diff. from English language)





closed class

- 1) Prepositions — on, in, at, by
- 2) Determiners — Articles determining something  
a, an, the, this, that
- 3) Pronouns — he, she, who, I
- 4) Conjunctive — and, but, or (join 2 sentences)
- 5) Auxiliary Verbs — can, may, should
- \* 6) Particles — up, down, on, off, in, out (give a little bit)  
↓  
of ambiguity
- 7) Numerals — one, two, first, second

Note

- Preposition always occur before nouns
  - Particles are used in "combin" with verbs
- Ex: turned down  $\xrightarrow{\text{meaning}}$  reject
- verb  $\downarrow$  particle

but when given to power → verb + particle → should be considered as  
↓  
word meaning

ex. rule out  $\rightarrow$  eliminate  
 ruled out  $\rightarrow$  disown  
 both words (if poss.)  
 word is verb  $\rightarrow$  find

meaning of comb<sup>n</sup> of words)

By rote board. → whenever we see down, look over words & find meaning accordingly.

(In strait based  $\rightarrow$  it'll look at previous usage of word & classify accordingly)

→ Conjunctions

Co-ordinating

↓  
sub-ordinating  
embedded  
in verb  
'thought'

→ and, or  
→ joining 2 ~~words~~ sentences.

Ex. → I thought that this course will be easy.

Teacher's Signature: .....

that is telling about the verb

→ Personal

Personal

→ He, she, I, me

Positive

→ mine, yours, his, hers

wh - Pronouns

→ who, what, whom, wherever

→ Auxiliary Verbs

be, do, can, have, had, are

Some other set of words:

\* Interjection : oh, ah, um, alas

\* Negation : no, not

\* Politeness marker : please, thank you

→ There's no various algo available to label. we'll see → <sup>Rule</sup> based algo

→ People have classified set of tags for POS tags (to maintain std)

A popular tag is:

Tag sets:

1) 45 - tag Penn TreeBank tag set (1993) → more popular

2) 61 - tag C5 Tagset (1997)

→ found from 87 - tag Brown Corpus (1960-70) → <sup>done in Brown</sup> University

Q → why reduced 87 → 45?

Atgo is more efficient when it works with less no. of tags.

→ It's observed only 14% of words are ambiguous but they come most frequently in sentences.

So, they give 14% ambiguity in sentences.

Used → Speech & language Processing  
by Jurafsky & Martin → 2<sup>nd</sup> Edition  
15th chapter - 1st & 2nd sections

PAGE NO.:

→ In some analysis, we can't simply remove '\$', '(', ')', etc.  
" " " " so, we use tags for these also.

→ Have to check whether these add values or lessen values  
of contents.

### Rule based approach:

EngCG → One of the algorithms  
↓  
constraint grammar

lexicon → dictionary

step 1 → finding possibilities of all the words

step 2 → applies large set of constraints to input sentence to rule out  
incorrect POS (from list we obtained earlier)

End term  
Before mid-term : 33%

After : 66% Today's → very less

15 minutes video Report → 28th

video → 10th May