stop at any point of time (depends how many communities we want)

5/04/19 (1 lecture missing)

→ SNA can also be based on Content Analysis. (till now → structural analysis.)
Ex. can we infer something from a colln 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, colln of words → sentences, colln 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%

Q. → How to input the content & 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. Conjuction
7. Participle
8. Article.

→ Basis of most of European language

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

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

Jagging Algos

* 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 Ru~~ How you treat a word in that sentence is decided by set of rules (disambiguate)

- No learning ∵ learning is done by human being

✓ 2. statistic~~al~~s based :

Have stats of how this word is used previously.

Ex. $\boxed{A}$ student

↳ most probably an article. – determiner ∵ (1) telling no.

Stats: 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.

→ There languages in which we don't have any adjective

challenges behind building Rule based approach

→ Need to know English structure

English word Classes

- Open Class
  - → new words added regularly in this set
  - → Nouns, Verbs, Adjectives, etc.
- Closed Class
  - → no new words are being added
  - → Prepositions, articles, etc

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

- Nouns
  - Proper Noun (name of person, city) etc.
  - Common Nouns
    - Count Nouns (having singular & plural eg goat, goats) (can say 1 goat, 2 goat)
    - Mass Nouns (group of similar entity eg. snow, water) (can't say 1 water, 2 water, ..)
- Verbs (action or process)
  - ↳ all kinds of forms eg. eat, ate, eats, ... eating
  - ↳ Auxillaries
    - ↳ closed form
- Adjectives (describe noun)
  - ↳ Ex. white, black, ~~etc~~, beautiful, good, bad
- Adverbs (add values to verb)
  - ↳ Ex. Unfortunately, John walked home extremely slowly yesterday.
  - all 4 words are describing his "walk" so, all these are adverbs
  - home → directional adverb
  - extremely → degree "
  - slowly → manner "
  - yesterday → temporal

② **Closed classes**

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.) Auxillary 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^n with verbs

Ex. turned down $\xrightarrow{\text{meaning}}$ reject
↓ verb ↓ particle

but when given to parser → verb + particle → should be considered as 1 meaning word

Ex. rule out → eliminate
find out → discover
} → algo. has to look at both words (If prev. word is verb → find meaning of comb^n of words)

Ex. Rule based. → Whenever we see down, look prev. words & find meaning accordingly.
(In state based → it'll look at previous usage of word & classify accordingly)

→ Conjunctions
↓ Co-ordinating
→ and, or
→ joining 2 ~~verbs~~ sentences.

↓ Sub-ordinating
→ that

Ex. → I thought that this course will be easy.
verb
embedded in verb 'thought'
'that' is telling about the verb.

→ Pronouns

Personal	Possessive	wh- pronouns
→ He, she, I, me	→ mine, your, his, her	→ who, what, whom, whoever

→ Auxillary Verbs : be, do, can, have, had, are

some other set of words :

* Interjection : oh, ah, un, alas

* Negation : no, not

* Politenes marker : please, thankyou

→ There're various algo available to label. we'll see → Rule 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 C 5 Tagset (1997)

→ found from 87 - tag Brown Corpus (1960-70) → done in Brown University

Q → why reduced 87 → 45 ?

Algo is more effecient 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 → 2nd Edition
(15th chapter : 1st 4 sub-sections)

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

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

Rule based approach :

EngCG → One of the algos
↓
constrained
grammar

lexicon → dictionary.

classifying

Step 1 → Finding all 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-sem : 33%
After ~ : 66% Today's → very less

15 minutes video Report → 28th
Video → 10th May