Word class -
Part of Speech Tag (POS)

It is a syntactic category or class of any word in a natural language sentence eg. Noun, Verb, Adjective, Adverb etc.

Ex:- Our dog chased a brown cat away from the home.

Word	Pos category
Our	pronoun
a, the	determiners
dog, cat, home	Nouns
brown	adjective
chased	verb (past tense)
away	adverb
from	preposition.

POS Tagging Problem:-

Problem is to identify what is the actual category for each of the word.

Given a text of English, identify the parts of speech of each word.

Text: The boy put the toys in the bag

POS category	Words
Noun (N)	boy, toys, bag
Verb (V)	put
Preposition (P)	in
Determiner (Det)	the

→ POS tags tell us more information about the word and its overall role in the sentence.

→ POS tags of a word also tells some information about its neighbour words.

eg. Nouns are generally preceded by adjectives and/or determiners
The girl, an ant, handsome boy.

→ Identifying POS tags is an important initial step for more complex downstream NLP tasks.

↳ parsing.
↳ Information extraction
↳ Sentiment Analysis.
↳ Machine Translation.

→ POS tags can be broadly categorized into two categories.

1. closed class
2. open class.

Closed class:-

↳ Relatively fixed set of words (limited in numbers)
↳ Addition of new closed words is very rare.
↳ Mostly functional: to tie the concept of a sentence together.

eg. preposition,
Determiners,
Pronoun,
Connectives.

Open class :-

* Cannot associated a fixed set of words, new words can be frequently encountered with such POS tags.

* Mostly content bearing: they refer to objects, action and features in the world.

* Addition of new open words is very frequence

eg. Nouns,
Verbs,
Adjectives,
Adverbs

POS : Level of Details

Decision of to take a very coarse grained level or to go to the more fine grained level.

In coarse grained level, we can identify a word as a noun but we can't tell whether it is a singular noun or a plural noun.

In fine grained level, we can give grammatical details of the word.

eg :- for verb it belong to past tense, future or present tense etc.

In very fine grained level, too many part of speech tags which leads to confusion.

POS Tag Set Examples:-

- Brown corpus tagset (87 tags):
- Penn Treebank tagset (45 tags):
- C7 tagset (146 tags)

Penn Treebank [PTB]

→ Very popular part of speech tagset is by University Pennsylvania.
→ It is called Penn Treebank (PTB)
→ It contains 45 part of speech tags.
→ Fairly standarized for English.
→ Popular NLP tools such as standford CoreNLP, spaCY used this tag set.
→ POS Tags are small and compact with 2-4 ~~captical~~ capital characters

UPenn Treebank Tagset:-

Noun Types.

POS Type	Explanation	Ex.
NN	Singular common noun	Women, Orange, Table
NNS	plural Common Noun	Women, Oranges, Tables
NNP	Singular proper Noun	Priya, Zenith, Jack
NNPS	Plural Proper Noun	Indians

Verb Types

POS Type	Explanation	Ex.
VB	Base form of a verb	Walk, play, eat, read.
VBD	Past tense of a verb	Walked, played, ate, read
VBN	Past Participle of a verb	walked, played, eaten, read.
VBG	Gerund form of a verb	fishing, walking, reading.
VBZ	3rd person verb on present tense	walks, plays, eats, read, is
VBP	Non 3rd person verb on present tense	Walk, play, eat, read, am, are
MD	Modal Verb	can, may, should.

Adjective & Adverb Types

POS type	Explanation	Ex.
JJ	Adjective	Intelligent, Small, fast
JJR	comparative Adjec	Better, Smaller
JJS	superlative Adjec	Best, Smallest
RB	Adverb	Back, Behind, fast, slow
RBR	comparative adverb	slower, faster
RBS	Superlative adverb	slowest, fastest

Pronoun, Determiner, Preposition Types

POS Type	Explanation	Ex.
PRP	Pronoun	He, She, they, I, we.
PRP$	Possessive Pronoun	His, Her, Your, our
POS	possessive Marker	India's, Asian's
DT	Determiner	The, a
CC	Coordinating conjunction	And, or, also, but.
IN	Preposition	In, Under, Of, from, with
CD	Cardinal Number	20, two

Ex:-

statement :- I am a girl

Word	POS Type	Explanation
I	PRP	Pronoun
am	VBP	Non 3rd person verb on present tense
a	DT	Determiner
girl	NN	Singular Common Noun

Statement :- Kavya is a intelligent girl

word	POS Type	Explanation
kavya	NNP	singular proper noun
is	VBZ(Aux. verb)	3rd person verb on present tense
a	DT	Determiner
Intelligent	JJ	Adjective
girl	NN	Singular Common Noun

Statement : She plays tennis

Word	PosType	Explanation
she	PRP	Pronoun
plays	VBZ	3rd person verb on present tense
tennis	NN	Singular common noun

Statement : They play football

Word	PosType	Explanation
They	PRP	Pronoun
play	VBP	Non 3rd person verb on present tense.
football	NN	Singular Common Noun.

Statement : And now for something completely different.

Word	Pos Type	Explanation
And	CC	Coordinating Conjunction
now	RB	Adverb
for	IN	Preposition
something	NN	Singular Common Noun
completely	RB	Adverb
different	JJ	Adjective

Statement : They refuse to permit us.
PRP VBP Det VB PRP.

Statement :

~~They refuse to permit us to obtain~~

They refuse to permit us to obtain refuse permit.
PRP VBP Det VB PRP Det VB NN NN.

Depends on context, the word takes multiple parts of speech tag (POS)

refuse = NN & VBP

permit = VB & NN

Difficulties / challenges :-

*. A word may have different POS tags. for. eg to sort the number.
sort is vert.
Bubble sort : sort is a Noun

*. closed class POS tags are not ambiguous but their frequency is also limited in a statement.

*. Generally words with open class are more prone to multiple pos tags.

*. open class pos tags occur a lot more frequently & its tagging depends on its neighbouring context that is previous & future words.

Exs of ambiguities in POS tagging :-

(1) - The attack (NN) was brutal
- King was planning to attack (VB) neighbouring states
- Tigers usually attack (VBP) their prey in a group

(2) - I read that (DT) book several times
- We thought that (IN (preposition)) you are in class
- We were talking about a village group that (WDT) had won the first prize in cricket

Some more ex. which are ambiguous for human also!!!

Look and Flies words are Noun as well as verb.

Look Word:

Look with Noun tag	Looks as Verb tag.
* Her looks ars keen	* He looks (VBZ) angry
* Take a quick look of a room	* She looked (VBD) straight

Fly Word:

Fly with Noun tag	Fly as Verb tag
* Flies are irritating	* kite flies swyftly in air
* candidates were dropping like flies during the technical interview	* Insects flying over the water.

→ Each word does not have a unique part of speech tags and it might depend on the context in this the word is being used.

→ In English language, ambiguity with no. of pos tags is uncommon (more than four) but 2 tags are common. We can't ignore.

→ success of pos tagging is crucial for NLP application in which machine have to understand language at syntactic level.

Application such as: Machine translation.

* Some word are unambigious in nature. carry only one POS.
* Some words are ambiguous.
* Probability technique can be applied.
* Also when we use particular language (English language), we follow certain rules such as.

1) Two determiners rarely follow each other (eg: the a girl)

2) Similar 2 base forms of the word they donot follow each other.
(eg want eat — is wrong
want to eat - is right)

3) Determiner is always followed by an adjective or noun.
(eg the table,
handsome boy,
the beautify girl)

4) Distinguishing past tense (VBD) with past participles (VBN) for irregular verbs is easy because they have different forms

VBN : has or have + the past participle

be + the past participle (be sometime is omitted)

Modal Verbs as auxillary Verbs :

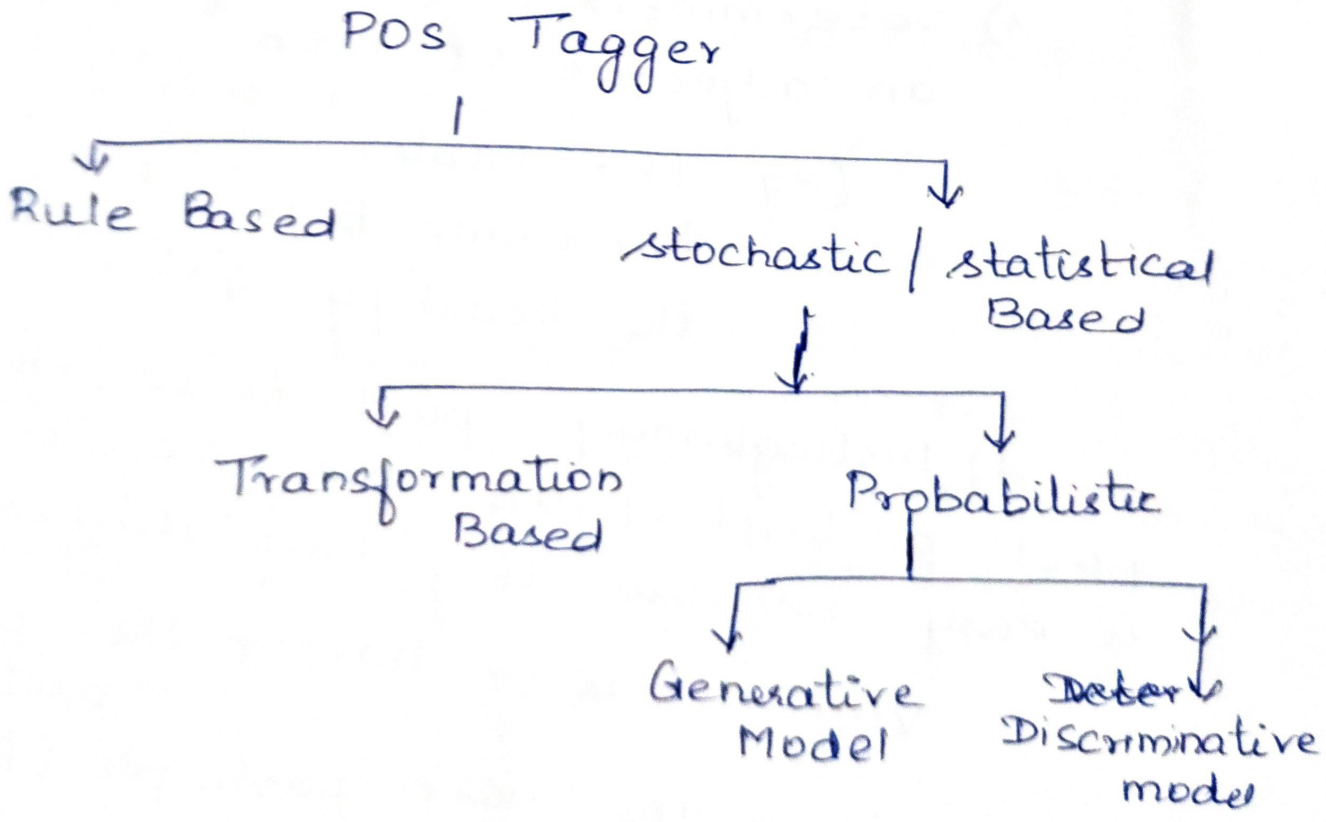
could / may / should / might / would / must + have + VBN

VBNs often used as participial adjectives whereas VBD often follow a subjective.

eg. The girl talked (VBD) to me

I have forgotten (VBN) my lines.

Classification of Taggers



Rule-based Approach

→ Rule-based approach : Rule-based pos tagging is the oldest approach that uses hand-written rules for tagging.

→ Rule based taggers depends on dictionary or lexicon to get possible tags for each word to be tagged.

→ Some language expert will sit together find out what are the symbol patterns.

→ Hand-written rules are used to identify the correct tag when a word has more than one possible tags.

→ Disambiguation is done by analyzing the linguistic features of the word, its preceding words, its following word & other aspects.

For eg, if the preceding word is article the word in question must be noun. This information is coded in the form of rules.

The rules may be context-pattern rules or as regular expressions compiled into finite-state automata that are intersected with lexically ambiguous sentence representation.

- TAGGIT, the first large rule based tagger, used context-pattern rules.

→ TAGGIT used a set of 71 tags & 3300 am disambiguation rules.

These rules disambiguated 77% of words in the million-word Brown University corpus.

Ex: I want to read a book.

Book: Noun or verb

Rule: before book, determiner (a) is there, therefore it is a noun

Limitation of Rule-based Approach:-

*. Hard coded rule are required.

*. Rules has to be updated based on language change.

*. It needs to check lexicon every time which is not efficient approach

*. strong language experts team is required.

Stochastic / Statistical based Approach :-

→ It is based on the concept of probability and machine learning.

→ Training & Test corpus are used.

→ Two approaches are there:

a) Transformation based tagger

b) Probabilistic based tagger

a) Transformation based tagger:

→ This tagger is based on concept of Transformation - Based Learning (TBL) approach.

→ TBL uses supervised learning.

→ There is assumption of pre-tagged training corpus.

→ It combines idea of the rule-based and stochastic taggers.

→ Like the rule based tagger, TBL is based on rules that specify what tags should be assigned to what words

→ But like the stochastic taggers, TBL is a machine learning technique, in which rules are automatically induced from the data.

Label the training set with most frequent tags.

eg: The can was rusted

The/DT can/MD was/VBD rusted/VBD

Add transformation rules to reduce training mistakes.

The/DT can/NN was/VBD rusted/VBN

Explanation: 'can' can be noun as well as modal verb also.

It uses machine learning as well as grammar rules.

1) Modal verb is never preceded by determiner

2) Also in corpus (training data) the possibility of 'determiner modal verb' pair is zero.

Therefore can is replaced as Noun tag.

Explanation:- 'rusted' can be VBD as well as VBN also

1) In grammar rule, VBN is preceded by VBD.

2) Also in corpus, possibility of VBD-VBD pair is negligible

Statement :-

Race is incorrectly tagged in the following statement.

IS	expected	to	race	tomorrow
↓	↓	↓	↓	↓
VBZ	VBN	TO/pre	NN	NN.

In second statement, this race is correctly tagged as an NN.

The	race	for	outer	space
↓	↓	↓	↓	↓
Det.	NN	IN/pre	JJ	NN

~~For eg,~~ In Brown corpus, the race is mostly likely to be a noun:

$P(NN|race) = .98$ $P(VB/race) = .02$

Therefore race is labelled as NN because its occurance (probability) is more.

Brill's tagger learned a rule that applies exactly to this mistagging of race:

Change NN to VB when the previous tag is To.

This rule would change race/NN to race/VB in exactly the following situation since it is preceded with to/TO:

is	expected	to	race	tomorrow
|	|	|	|	|
VBZ	VBN	TO	VB	NN

b) Probabilistic:

Approach is to "pick the most-likely tag for this word".

Two approaches are there whether you generate the data from the class or the class from the data.

problem statement: Data available of the form [d, c]

where d is observation & c is hidden classes.

Two types of Model are there.

i) Generative Model
ii) Discriminative Model.

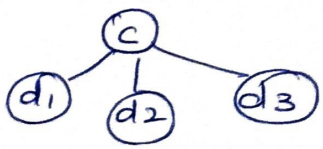
i) Generative Model:

→ In Generative model, we assume that class is there & it generates data.

→ eg. class is given (subjects) & all words are generated from subject (eg. maths, english, science)

→ POS tags are there & words are generated from that. [eg Det → the / a

→ Here flow is downward [classes generates data]



eg:- Hidden Markov Model, Naive Bayes classifier