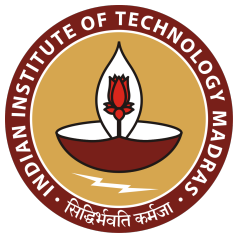
BSCS5002: Introduction to Natural Language Processing

Part-of-Speech Tagging

Parameswari Krishnamurthy



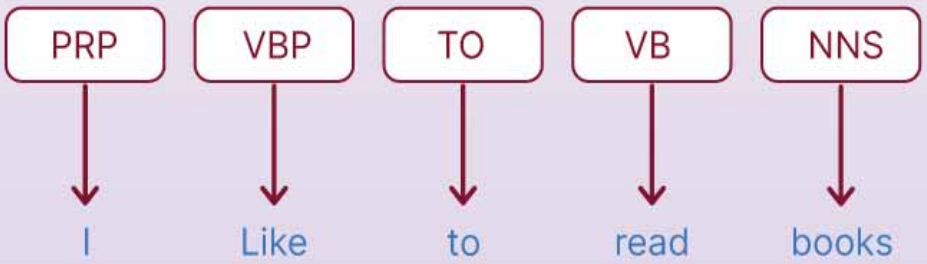
Language Technologies Research Centre
IIIT-Hyderabad

param.krishna@iiit.ac.in

Part-of-Speech (POS) Tagging

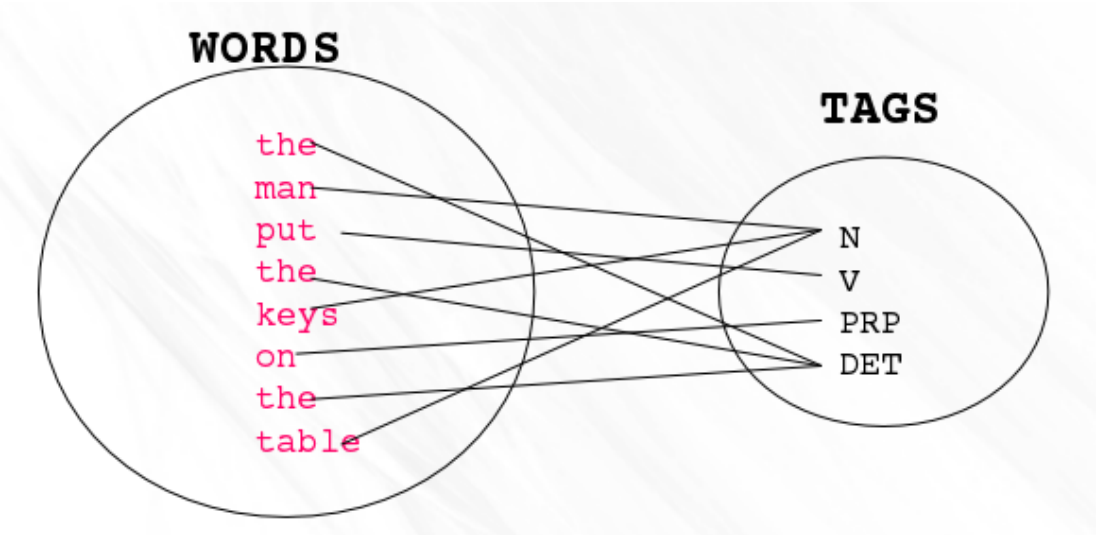
- Assigning grammatical categories (e.g., noun, verb, adjective) to words in a sentence.
- Helps in understanding the syntactic structure and meaning of sentences.
- Done using POS taggers, which assign tags based on word context and linguistic rules.
- Example POS tags:
  - Noun (NN)
  - Verb (VB)
  - Adjective (JJ)
  - Adverb (RB)
  - Pronoun (PRP)
  - Determiner (DT)
- POS tagging is a fundamental task in natural language processing (NLP).
- Used in various applications such as information retrieval, sentiment analysis, and machine translation.

POS Tagging
In NLP



What is POS Tagging?

- The process of assigning a part-of-speech to each word in a sentence.
- It is normally a *sentence-based approach*.
- Given a sentence formed of a sequence of words, POS tagging tries to label (tag) each word with its correct part of speech (also named *word category*, *word class*, or *lexical category*).



Some Examples

NN	noun	chair, bandwidth, pacing
VB	verb	study, debate, munch
ADJ	adjective	purple, tall, ridiculous
ADV	adverb	unfortunately, slowly
PRP	preposition	of, by, to
PRO	pronoun	I, me, mine
DET	determiner	the, a, that, those
CC	conjuncts	and, but

Difference b/w Morphological Analyzer and POS

MA	POS
>>Finds internal structure of a word (root form, affixes, etc.)	assigning a POS tag to the given surface form word
>>word-based approach	sentence-based approach
>>needs smaller number of categories	needs a lots of categories
>>one/more analysis for a word	one analysis for a word
e.g.	
back <back,adj,norm> | <back,n,3,sg,0> | <back,adv> | <back,v>	The *back* door = JJ (ADJ) On my *back* = NN Win the voters *back* = RB (ADV) Promised to *back* the bill = VB

Why POS tagging?

- First step in parsing
- More tractable than full parsing, intermediate representation
- Resolves lexical ambiguity
  e.g. this
  This is a nice day = PRP
  This day is nice = DET
  You can go *this* far = ADV
  The POS tagging problem is to determine the POS tag for a particular instance of a word.

Why POS tagging?

- Useful as a step for several other, more complex NLP tasks, e.g.
  - Speech synthesis pronunciation

Lead	Lead
INsult	inSULT
OBject	obJECT
OVERflow	overFLOW
DIScount	disCOUNT
CONtent	conTENT

  - Parsing: e.g. Time flies like an arrow
    Is flies a N or V?
  - Word prediction in speech recognition
    - Possessive pronouns (my, your, her) are likely to be followed by nouns
    - Personal pronouns (I, you, he) are likely to be followed by verbs

Challenges in POS Tagging

Why POS tagging is hard?

Challenges:

- Ambiguity: multiple category
- In a sentence, obviously there exist some words for which more than one POS tag is possible.
- e.g. *Can can can the can*

While disambiguating a particular word, humans exploit several mechanisms and information sources:

- the roles of other words in the sentence,
- the syntactic structure of the sentence,
- the domain of the text, and
- the world knowledge.

Challenges

- Garden-Path Sentences
- Examples:
  - *The horse raced past the barn fell.*
  - *The government plans to raise taxes were defeated.*
    first interpretation : The government is planning to raise taxes...
    final interpretation : The plans of the government to raise taxes were defeated.
  - *The old man the boat*
    first interpretation : The man, who is old...
    final interpretation : The boat is manned by the old.

Example-1

- Ambiguous POS contexts
  Girls have broken hearts
  Time flies like an arrow.
  - Possible POS assignments:
  - Time/[V,N] flies/[V,N] like/[V,Prep] an/Det arrow/N
  - Time/N flies/V like/Prep an/Det arrow/N
  - Time/N flies/N like/Prep an/Det arrow/N
  - Time/N flies/N like/V an/Det arrow/N
- Unknown Words
  - I like that *app*

How to Procede?

A natural question that may arise is:
what are these parts of speech?
or how do we specify a set of suitable parts of speech?
Famous traditional POS tags :

- noun
- verb
- adjective
- adverb
- pronoun
- preposition
- conjunction
- interjection

What about *five*, *the*, *$*?

Word Classes

The significance of the POS for language processing is that it gives a significant amount of information about the word and its neighbors.
Two types of Word Classes:

- Open word classes
  - new members are added
  - Four major open classes: *nouns*, *verbs*, *adjectives*, and *adverbs*.
- Closed word classes
  - Having relatively fixed membership
  - closed classes: *prepositions*, *determiners*, *conjunctions* etc

Open Word Classes

(i) Nouns

Nouns are traditionally grouped into **proper nouns** and **common nouns**. **Proper nouns:** (NNP)

Rani, Kasargod, and IBM

Not preceded by articles, e.g., the book is upstairs, but (*the) Rani is upstairs.

Common nouns: (NN)

1. Count nouns:
Allow grammatical enumeration, i.e., both singular and plural (goat/goats), and can be counted (one goat/ two goats)

2. Mass nouns:
Something is conceptualized as a homogeneous group, snow, salt, and communism.
Appear without articles where singular nouns cannot (Snow is white but not *Goal is white)

Nouns of Space and Time (NST)

Nouns of space and time (NST) or adverbial nouns (Krishnamurti & Gwynn, 1985:98; Whitman, 2002:561) form a special type of nouns without number marking.
These forms also function as postpositions when they take nouns as their complements.
When they occur without noun complements, they are categorized as nouns with the following features.
(i) NST have the ability to form an oblique stem, which can be used adjectively;
(ii) NST have the ability to add case markers and postpositions to the oblique stem;
(iii) NST have the ability to add third-person pronominal suffixes to the oblique stem.

2. Verbs

- Most of the words referring to actions and processes including main verbs like *draw*, *provide*, *differ*, and *go*.
- A number of morphological forms: non-3rd-person-sg (eat), 3rd-person-sg(eats), progressive (eating), past participle (eaten)

i. Finite Verbs (VF)
A finite form is one that can stand as the main verb of a sentence and occur before a final pause (full stop).
e.g. Tam: *paṭittāṉ* 'he read'

ii. Non-finite Verbs (VNF)
A non-finite form cannot stand as a main verb and rarely occurs before a final pause
e.g. avaṉ *paṭittā* nallatu 'It is good if he reads'

iii. Infinitives (VINF)
e.g. nāṉ cinimā *pārkka* vantēṉ 'I came to see the cinema'

iv. Gerunds (VNG)
e.g. *cirittal* uṭmpukku nallatu. 'laughing is good for health'

v. Auxiliary (VAUX)
e.g. nī vara *vēṉṭām*. ‘you don’t come’

Adjectives (JJ)

- Terms describing properties or qualities
- Most languages have adjectives for the concepts of color (white, black), age (old, young), and value (good, bad), but

Adverbs (RB)

- Words viewed as modifying something (often verbs)
- Directional (or locative) adverbs: specify the direction or location of some action, hoe, here, downhill
- Manner adverb: describe the manner of some action or process, slowly, slinkily, delicately
- Temporal adverbs: describe the time that some action or event took place, yesterday, Monday
- Degree adverbs: specify the extent of some action, process, or property, extremely, very, somewhat

Adverbs (JJ)

- Words viewed as modifying something (often verbs)
- Directional (or locative) adverbs: specify the direction or location of some action, hoe, here, downhill
- Manner adverb: describe the manner of some action or process, slowly, slinkily, delicately
- Temporal adverbs: describe the time that some action or event took place, yesterday, Monday
- Degree adverbs: specify the extent of some action, process, or property, extremely, very, somewhat

Closed Classes:

- Pronouns: *I*, *you*, *he*
- Prepositions : *on*, *under*, *over*, *near*, *by*, *at*, *from*, *to*, *with*
- Determiners: *a*, *an*, *the*
- Pronouns: *she*, *who*, *I*, *others*
- Conjunctions: *and*, *but*, *or*, *as*, *if*, *when*
- Particles: *up*, *down*, *on*, *off*, *in*, *out*, *at*, *by*
- Numerals: *one*, *two*, *three*, *first*, *second*, *third*

Tagsets

- Part-of-Speech (POS) tagsets are collections of labels used to classify words based on their grammatical roles in sentences
- Granularity: Need to Decide the level of granularity needed for the tagset.
- For example, more detailed tagsets might distinguish between different types of nouns or verbs.
- Types:
  - **Universal tagsets:** Designed for cross-linguistic compatibility (e.g., Universal Dependencies).
  - **Standardized tagsets:** Widely accepted and used in specific languages (e.g., Penn Treebank for English).
  - **Language-specific tagsets:** Tailored for specific languages (e.g., BIS for Indian languages).

Penn Treebank Tagset:

Tag	Description	Example	Tag	Description	Example	Tag	Description	Example
CC	coordinating conjunction	*and, but, or*	PDT	predeterminer	*all, both*	VBP	verb non-3sg present	*eat*
CD	cardinal number	*one, two*	POS	possessive ending	*'s*	VBZ	verb 3sg pres	*eats*
DT	determiner	*a, the*	PRP	personal pronoun	*I, you, he*	WDT	wh-determ.	*which, that*
EX	existential 'there'	*there*	PRP$	possess. pronoun	*your, one's*	WP	wh-pronoun	*what, who*
FW	foreign word	*mea culpa*	RB	adverb	*quickly*	WP$	wh-possess.	*whose*
IN	preposition/ subordin-conj	*of, in, by*	RBR	comparative adverb	*faster*	WRB	wh-adverb	*how, where*
JJ	adjective	*yellow*	RBS	superlatv. adverb	*fastest*	$	dollar sign	*$*
JJR	comparative adj	*bigger*	RP	particle	*up, off*	#	pound sign	*#*
JJS	superlative adj	*wildest*	SYM	symbol	+,%, &	"	left quote	' or "
LS	list item marker	*1, 2, One*	TO	"to"	*to*	"	right quote	' or "
MD	modal	*can, should*	UH	interjection	*ah, oops*	(	left paren	[, (, {, <
NN	sing or mass noun	*llama*	VB	verb base form	*eat*	)	right paren	],), }, >
NNS	noun, plural	*llamas*	VBD	verb past tense	*ate*	,	comma	,
NNP	proper noun, sing.	*IBM*	VBG	verb gerund	*eating*	.	sent-end punc	. ! ?
NNPS	proper noun, plu.	*Carolinas*	VBN	verb past part.	*eaten*	:	sent-mid punc	: ; ... – -

Universal Dependencies-POS

- Universal Dependencies(UD) POS tagset is a standardized set of tags used for part-of-speech tagging across multiple languages.
- Developed as part of the Universal Dependencies project, it aims to provide a consistent POS tagging framework for cross-linguistic research and applications.
- It supports the development of cross-linguistic NLP tools and resources.
- Universal Coverage: Designed to capture syntactic and grammatical categories across different languages in a consistent manner.
- Minimalist Approach: Focuses on core grammatical categories to ensure broad applicability while avoiding language-specific complexities.

UD POS Tags

Open class words	Closed class words	Other
ADJ	ADP	PUNCT
ADV	AUX	SYM
INTJ	CCONJ	X
NOUN	DET	
PROPN	NUM	
VERB	PART	
	PRON	
	SCONJ	

Indian language Tagset: BIS

- The BIS POS tagset is specifically designed for Indian languages under the Bureau of Indian Standards (BIS) (IS 17627 : 2021)
- It was developed to standardize POS tagging across various Indian languages.
- The tagset is multilingual and covers a wide range of Indian languages, including Hindi, Tamil, Telugu, Bengali, Marathi, and others.
- The BIS tagset is based on a hierarchical structure that captures both coarse-grained and fine-grained linguistic categories.
- It includes both universal categories (e.g., Nouns, Verbs) and language-specific tags to address the unique features of Indian languages.

BIS-POS tags

Common Noun (NN); Proper Noun (NNP); Noun of Space and Time (NST); Pronoun (PR); Personal (PRP); Reflexive (PRF); Relative (PRL); Reciprocal (PRC); Wh-word (PRQ); Demonstrative (DM); Main Verb(VM); Finite Verb (VF); Non-finite Verb (VNF); Infinitive (VINF); Gerund (VNG); Auxiliary (VAUX); Adjective (JJ); Adverb (RB); Postposition (PSP); Conjunction (CC); Coordinator (CCD); Subordinator (CCS); Quotative (UT); Particles (RP); Classifier (CL); Interjection (INJ); Intensifier (INTF); Negation (NEG); Quantifiers (QT); Residuals (RD); Foreign word (RDF); Symbol (SYM); Punctuation (PUNC); Unknown (UNK); Echowords (ECH)

Methods of POS tagging

- Rule-based POS tagging
- Transformation based POS tagging
- Statistical based POS tagging