

# Part of Speech Tagging Lecture #3a

Natural Language Processing  
School of Computing

# Part of Speech

- a category to which a word is assigned in accordance with its syntactic functions.
- Also known as : **POS, Word Classes, or Syntactic Categories**
- **Ex :**
  - **Nouns:** people, animals, concepts, things
  - **Verbs:** expresses action in the sentence
  - **Adjectives:** describe properties of nouns
- **Give information about a word and its neighbors**
- **Ex :** Nouns are likely to be preceded by determiners and adjectives, while verbs are by nouns

# Part of Speech

- Dionysius Thrax dari Alexandria (100 B.C.) define 8 PoS :
  - Noun
  - Verb
  - Pronoun
  - Preposition
  - Adverb
  - Conjunction
  - Participle
  - Article

# Part of Speech : Now

- 45 from Penn Treebank dataset
- 87 from Brown Corpus
- 146 from C7 tagset

# PoS Tag Examples

Number	Tag	Description	Number	Tag	Description
1.	CC	Coordinating conjunction	21.	RBR	Adverb, comparative
2.	CD	Cardinal number	22.	RBS	Adverb, superlative
3.	DT	Determiner	23.	RP	Particle
4.	EX	Existential <i>there</i>	24.	SYM	Symbol
5.	FW	Foreign word	25.	TO	<i>to</i>
6.	IN	Preposition or subordinating conjunction	26.	UH	Interjection
7.	JJ	Adjective	27.	VB	Verb, base form
8.	JJR	Adjective, comparative	28.	VBD	Verb, past tense
9.	JJS	Adjective, superlative	29.	VBG	Verb, gerund or present participle
10.	LS	List item marker	30.	VBN	Verb, past participle
11.	MD	Modal	31.	VBP	Verb, non-3rd person singular present
12.	NN	Noun, singular or mass	32.	VBZ	Verb, 3rd person singular present
13.	NNS	Noun, plural	33.	WDT	Wh-determiner
14.	NNP	Proper noun, singular	34.	WP	Wh-pronoun
15.	NNPS	Proper noun, plural	35.	WP\$	Possessive wh-pronoun
16.	PDT	Predeterminer	36.	WRB	Wh-adverb
17.	POS	Possessive ending			
18.	PRP	Personal pronoun			
19.	PRP\$	Possessive pronoun			
20.	RB	Adverb			

# POS Examples

- Noun : book/books, nature, Germany, Sony
- Verb : eat, wrote
- Auxiliary : can, should, have
- Adjective : new, newer, newest
- Adverb : well, urgently
- Numbers : 872, two, first
- Article/Determiner : the, some
- Conjunction : and, or
- Pronoun : he, my
- Preposition : to, in
- Particle : off, up
- Interjection : Ow, Eh

# Closed vs. Open Class Words

- Closed class: relatively fixed set
  - Prepositions: of, in, by, ...
  - Auxiliaries: may, can, will, had, been, ...
  - Pronouns: I, you, she, mine, his, them, ...
  - Usually function words (short common words which play a role in grammar)
- Open class: productive
  - Nouns, Verbs, Adjectives, Adverbs
    - Nouns (Proper Nouns) -> Telkom University, Mukidi
    - Verbs -> Please **email** me the invitation

# Applications : Grammatical Checking

- He went to a restaurant yesterday
- She goes to a bookstore today
  
- What is the difference between using Language Modeling and POSTag information?



# Applications : Information Extraction

- help to find named entities (people, organization)

# Applications : Machine Translation

- Give me a **round** figure. [adjective]
- Shall we play another **round** of cards? [noun]
- He had a look **round** before he kept going. [adverb]
- They walked **round** the tree. [preposition]
- The floor function **rounds** down. [verb]

# Applications : Paraphrasing and Summarization

- The **well** was drilled fifty meters deep

- All is **well** with us

so [conjunction]

spring [noun]

- The **grow** [verb] was drilled fifty meters deep

okay [adjective]

successfully [adverb]

anyway [interjection]

so [conjunction]

spring [noun]

- All is **grow** [verb] with us

okay [adjective]

successfully [adverb]

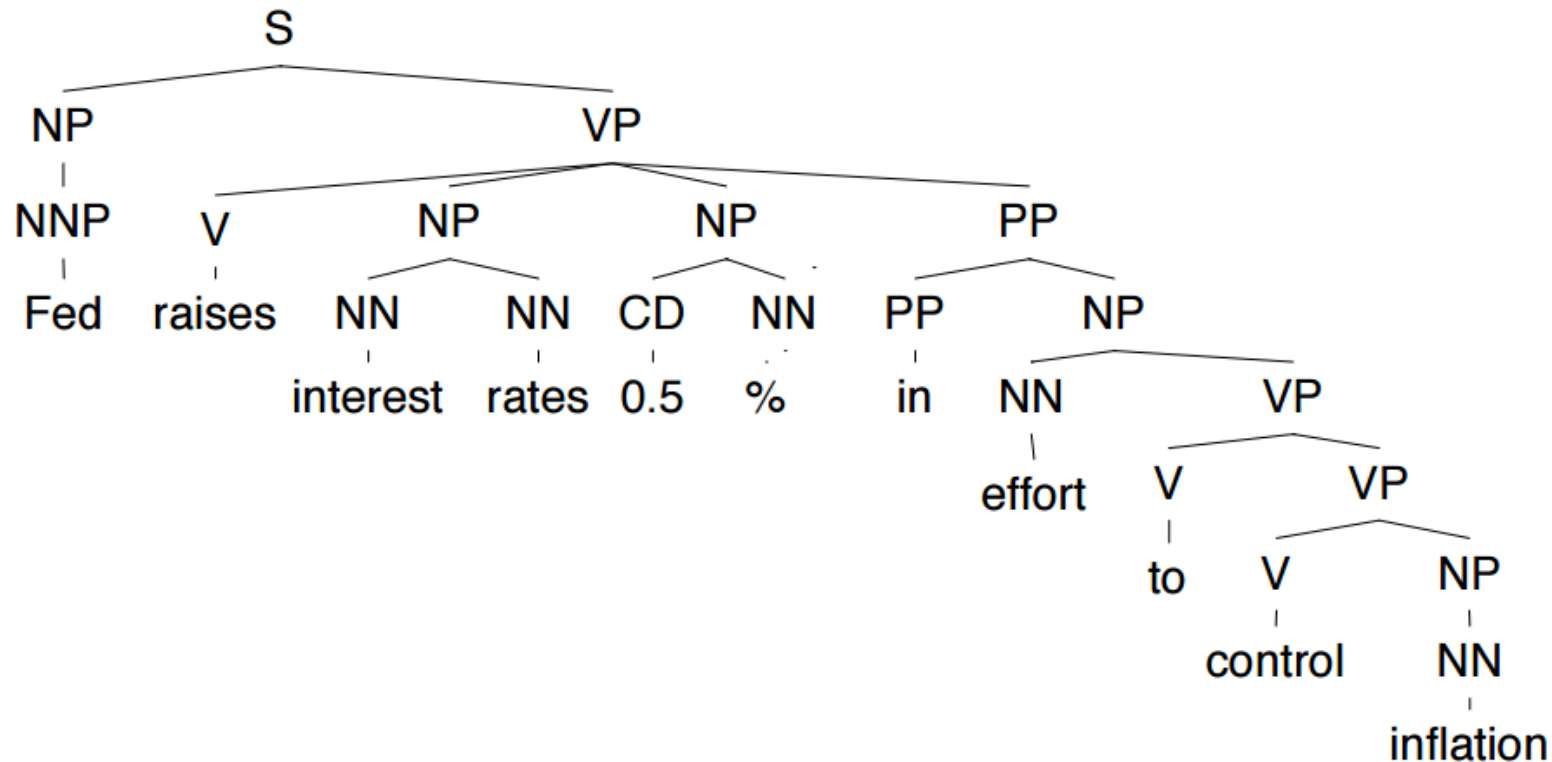
anyway [interjection]

# Applications : Speech Synthesis

- They **live/verb** in Bandung
- Opening PON is **live/adj** on TV
- Eggs have a high protein **content/Noun**
- She was **content/Verb** to step down after four years as CEO

# Applications : Parsing

- Fed raises **interest** rates 0.5% in effort to control inflation



# PoS Tagging

The process of assigning a part of speech to each word in a text

Challenge: words often have more than one POS **AMBIGUITIES DETECTED**

- earnings growth took a **back/JJ** seat
- a small building in the **back/NN**
- a clear majority of senators **back/VBP** the bill
- Dave began to **back/VB** toward the door
- enable the country to buy **back/RP** about debt
- I was twenty-one **back/RB** then

# Tagging Task

- Time flies like an arrow → Time/N flies/V like/Prep an/Det arrow/N
- Fruit flies like a banana → Fruit/N flies/N like/V a/DET banana/N
- Pemerintah/NNP kota/NNP Delhi/NNP mengerahkan/VB monyet/NN untuk/SC mengusir/VB monyet-monyet/NN lain/JJ yang/SC berbadan/VB lebih/RB kecil/JJ dari/IN arena/NN Pesta Olahraga/NNP Persemajukan/NNP ./Z

# Distribution of Ambiguities

- 45-tags Brown corpus (word types)
- Unambiguous (1 tag): 38,857
- Ambiguous: 8,844
  - 2 tags: 6,731
  - 3 tags: 1,621
  - 4 tags: 357
  - 5 tags: 90
  - 6 tags: 32
  - 7 tags: 6 (well, set, round, open, fit, down)
  - 8 tags: 4 ('s, half, back, a)
  - 9 tags: 3 (that, more, in)



# Approaches

Approach :

- Rule Based
- Statistical/Stochastic

Degree of Supervision :

- Supervised : Training corpus is tagged by humans
- Unsupervised : Training corpus isn't tagged
- Partly Supervised : e.g. Training corpus isn't tagged, but we have dictionary giving possible tags for each word

# Rule-Based Tagging

- Typically...start with a dictionary of words and possible tags
- Assign all possible tags to words using the dictionary
- Write rules by hand to *selectively remove* tags
- Stop when each word has exactly one (presumably correct) tag

# Rule Based Tagging : Example

- Assign All Possible POS to Each Word

			NN		
			RB		
	VBN		JJ		VB
PRP	VBD	TO	VB	DT	NN
<b>She</b>	<b>promised</b>	<b>to</b>	<b>back</b>	<b>the</b>	<b>bill</b>

# Rule Based Tagging : Example

- Apply Rules Eliminating Some POS

*E.g., Eliminate VBN if VBD is an option when VBN/VBD follows “<start> PRP”*

			NN		
			RB		
	<del>VBN</del>		JJ		VB
PRP	VBD	TO	VB	DT	NN
<b>She</b>	<b>promised</b>	<b>to</b>	<b>back</b>	<b>the</b>	<b>bill</b>

# How Hard is PoS Tagging ?

- Many words have only one POS tag (e.g. *is, Mary, very, smallest*)
- Others have a single ***most likely*** tag (e.g. *dog -> photographers seemed to dog her every step ???* )
- Using state-of-the-art automated method, how many tags are correct?
  - About 97% currently
- But baseline is already 90%
- Baseline is performance of simplest possible method:
  - Tag every word with its most frequent tag
  - Tag unknown words as nouns

# Improving Language Model

- Tags also tend to *co-occur* regularly with other tags (e.g. Det, N)
- In addition to conditional probabilities of words  $P(w_1 | w_{n-1})$ , we can look at POS likelihoods ( $P(t_1 | t_{n-1})$ ) to disambiguate sentences and to assess sentence likelihoods