Part-Of-Speech Tagging

Rahmad Mahendra

rahmad.mahendra@cs.ui.ac.id

Slide acknowledgement:
Lecture Notes from Edinburgh University, University of Texas at Austin, and University of Illnois

Natural Language Processing & Text Mining
Short Course

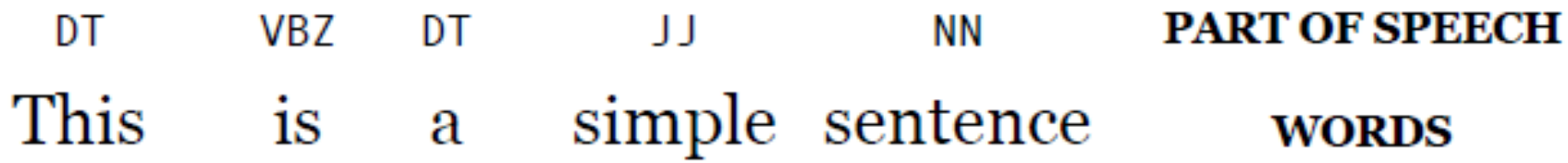
Pusat Ilmu Komputer UI
22 – 26 Agustus 2016

Why POS Tagging?

- Speech synthesis
  - How to pronounce “live”?
- Machine Translation
  - The noun “content” may have a different translation from the adjective.
- Information extraction
  - Finding names, relations, etc.

Part-Of-Speech

- Annotate each word in a sentence with a part-of-speech marker.



- Useful for subsequent syntactic parsing and word sense disambiguation.

English Part of Speech

- Noun (person, place or thing)
  - Singular (NN): dog, fork
  - Plural (NNS): dogs, forks
  - Proper (NNP, NNPS): John, Springfields
  - Personal pronoun (PRP): I, you, he, she, it
  - Wh-pronoun (WP): who, what
- Adjective (modify nouns)
  - Basic (JJ): red, tall
  - Comparative (JJR): redder, taller
  - Superlative (JJS): reddest, tallest
- Determiner
  - Basic (DT): a, an, the
  - WH-determiner (WDT): which, that

English Part of Speech

- Verb (actions and processes)
  - Base, infinitive (VB): eat
  - Past tense (VBD): ate
  - Gerund (VBG): eating
  - Past participle (VBN): eaten
  - Non 3rd person singular present tense (VBP): eat
  - 3rd person singular present tense (VBZ): eats
  - Modal (MD): should, can
  - To (TO): to (to eat)
- Adverb (modify verbs)
  - Basic (RB): quickly
  - Comparative (RBR): quicker
  - Superlative (RBS): quickest

English POS Tagsets

- Brown corpus tagset: 87 tag (Francis and Kucera, 1982)
- Penn Treebank tagset: 45 tag (Marcus et al., 1993)
- C5 CLAWS BNC tagset: 61 tag (Garside et al., 1997)
- C7 tagset: 146 tag (Leech et al., 1994)

Penn Treebank Tagsets

Tag	Description	Example	Tag	Description	Example
CC	coordin. conjunction	*and, but, or*	SYM	symbol	+,%, &
CD	cardinal number	*one, two, three*	TO	"to"	*to*
DT	determiner	*a, the*	UH	interjection	*ah, oops*
EX	existential 'there'	*there*	VB	verb, base form	*eat*
FW	foreign word	*mea culpa*	VBD	verb, past tense	*ate*
IN	preposition/sub-conj	*of, in, by*	VBG	verb, gerund	*eating*
JJ	adjective	*yellow*	VBN	verb, past participle	*eaten*
JJR	adj., comparative	*bigger*	VBP	verb, non-3sg pres	*eat*
JJS	adj., superlative	*wildest*	VBZ	verb, 3sg pres	*eats*
LS	list item marker	*1, 2, One*	WDT	wh-determiner	*which, that*
MD	modal	*can, should*	WP	wh-pronoun	*what, who*
NN	noun, sing. or mass	*llama*	WP$	possessive wh-	*whose*
NNS	noun, plural	*llamas*	WRB	wh-adverb	*how, where*
NNP	proper noun, singular	*IBM*	$	dollar sign	$
NNPS	proper noun, plural	*Carolinas*	#	pound sign	#
PDT	predeterminer	*all, both*	"	left quote	` or "
POS	possessive ending	*'s*	"	right quote	' or "
PRP	personal pronoun	*I, you, he*	(	left parenthesis	[, (, {, <
PRP$	possessive pronoun	*your, one's*	)	right parenthesis	],), }, >
RB	adverb	*quickly, never*	,	comma	,
RBR	adverb, comparative	*faster*	.	sentence-final punc	. ! ?
RBS	adverb, superlative	*fastest*	:	mid-sentence punc	: ; ... – -
RP	particle	*up, off*			

Universal POS Tags

- NOUN (nouns)
- VERB (verbs)
- ADJ (adjectives)
- ADV (adverbs)
- PRON (pronouns)
- DET (determiners and articles)

Universal POS Tags

- ADP (prepositions and postpositions)
- NUM (numerals)
- CONJ (conjunctions)
- PRT (particles)
- '.' (punctuation marks)
- X (anything else, such as abbreviations or foreign words)

Closed vs Open Class

- ***Closed class*** categories are composed of a small, fixed set of grammatical function words for a given language.
  - Pronouns, Prepositions, Modals, Determiners, Particles, Conjunctions
- ***Open class*** categories have large number of words and new ones are easily invented.
  - Nouns, Verbs, Adjectives, Abverb

Why POS Tagging is Hard?

- Ambiguity
  - glass of water/NOUN vs. water/VERB the plants
  - lie/VERB down vs. tell a lie/NOUN
  - wind/VERB down vs. a mighty wind/NOUN
- Sparse data
  - Words we haven't seen before
  - Word-Tag pairs we haven't seen before

POS Tag Ambiguity

How many tags does each word type have?

(Original Brown corpus: 40% of tokens are ambiguous)

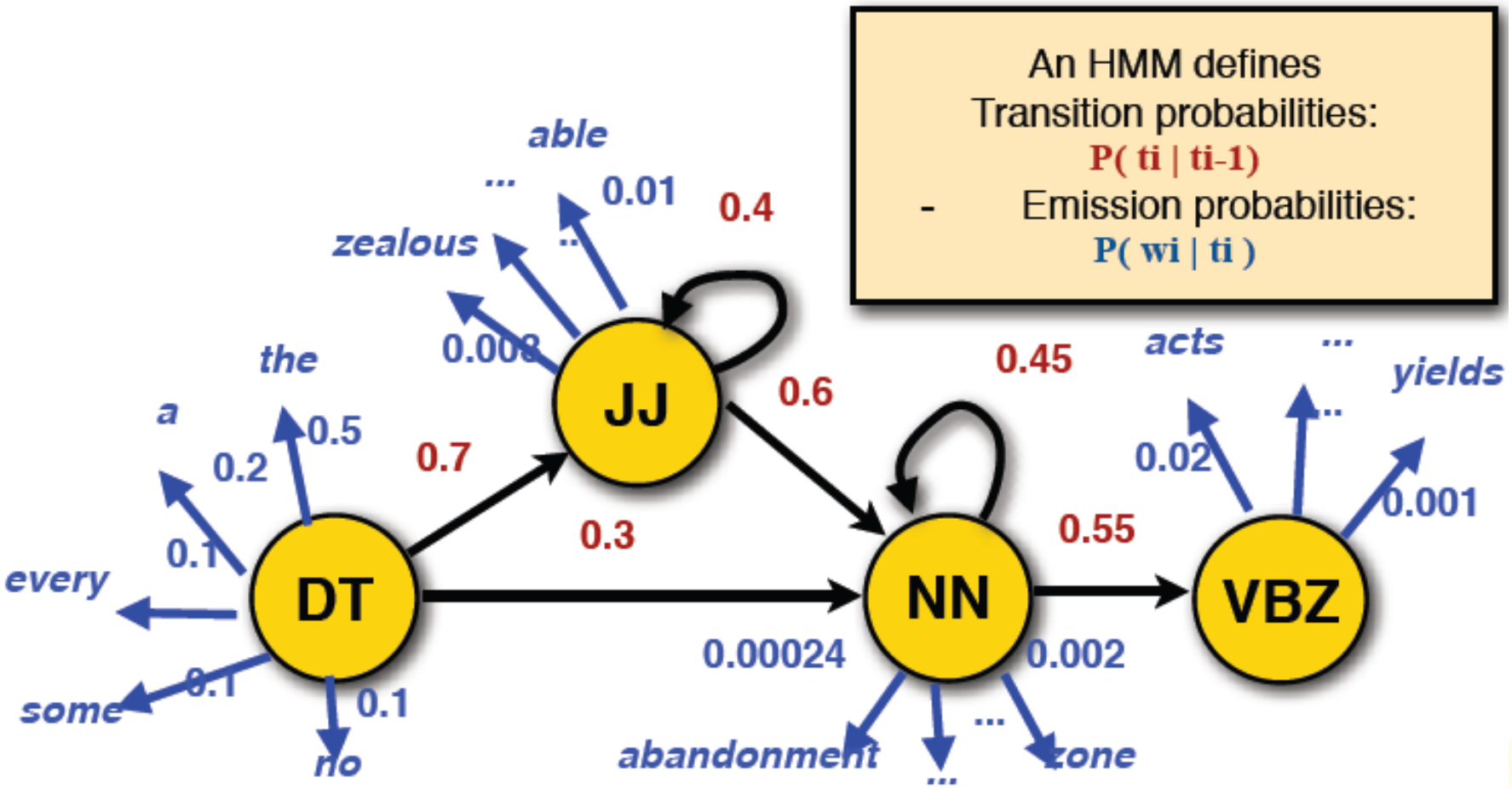
		87-tag Original Brown		45-tag Treebank Brown	
Unambiguous (1 tag)		**44,019**		**38,857**	
Ambiguous (2–7 tags)		**5,490**		**8844**	
Details:	2 tags	4,967		6,731	
	3 tags	411		1621	
	4 tags	91		357	
	5 tags	17		90	
	6 tags	2	(*well*, *beat*)	32	
	7 tags	2	(*still*, *down*)	6	(*well*, *set*, *round*, *open*, *fit*, *down*)
	8 tags			4	(*'s*, *half*, *back*, *a*)
	9 tags			3	(*that*, *more*, *in*)

pusilkom ui
quality results

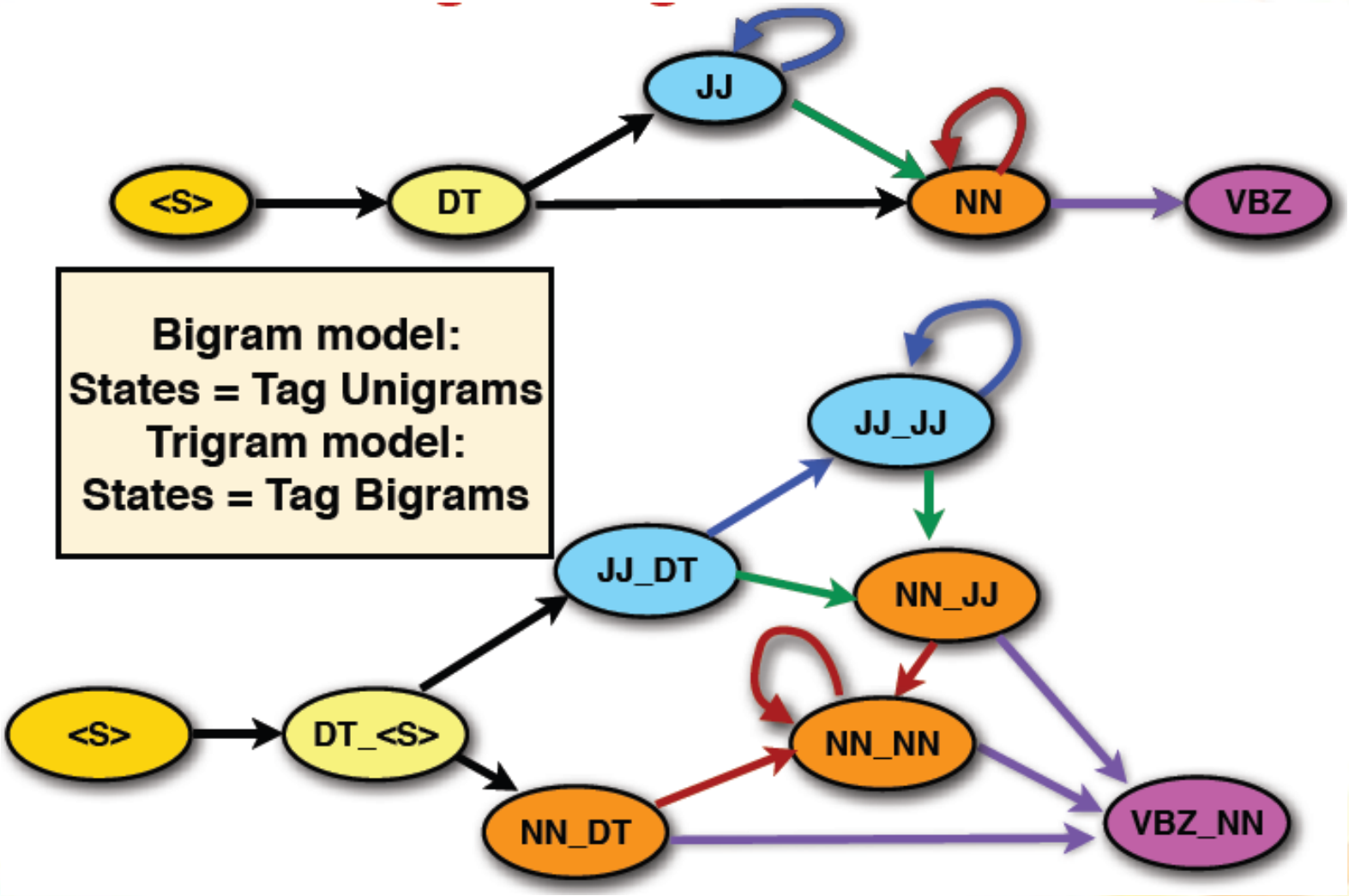
POS Tagging Approaches

- **Rule-Based**: Human crafted rules based on lexical and other linguistic knowledge.
- **Learning-Based**: Trained on human annotated corpora like the Penn Treebank.
  - **Statistical models**: Hidden Markov Model (HMM), Maximum Entropy Markov Model (MEMM), Conditional Random Field (CRF)
  - **Rule learning**: Transformation Based Learning (TBL)

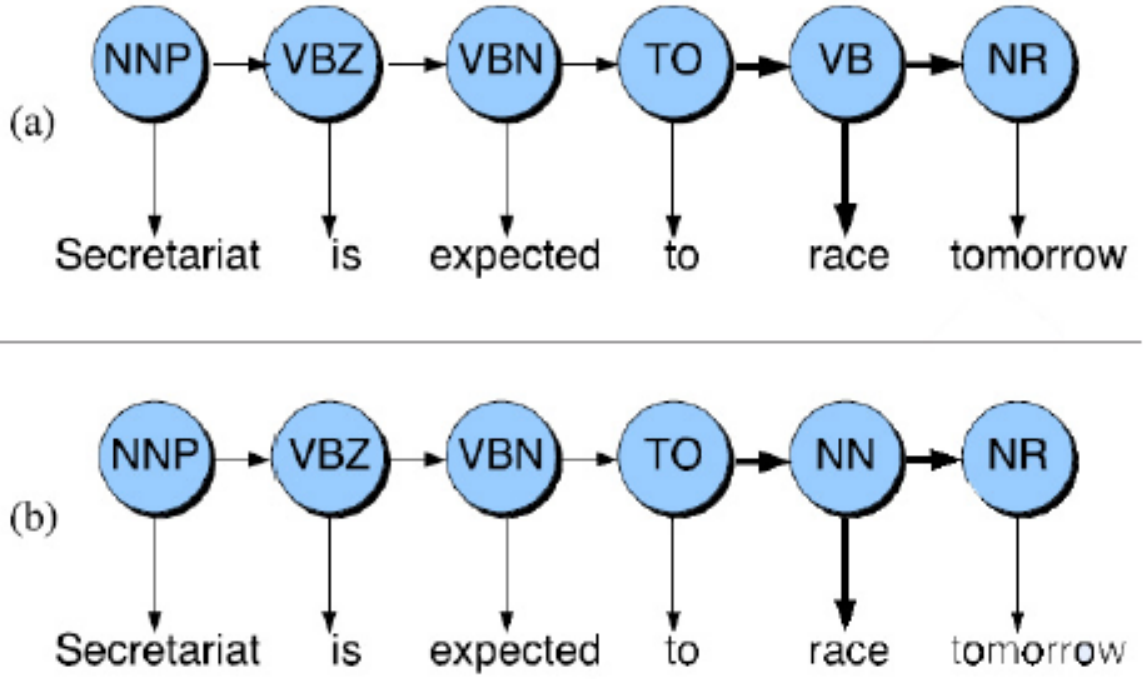
Hidden Markov Model



Hidden Markov Model



Word and Tag Sequence Probability



1. $P(VB|TO) \times P(race|VB) = 0.34 \times 0.00003 = 0.00001$
2. $P(NN|TO) \times P(race|NN) = 0.021 \times 0.00041 = 0.000007$

Thank you