CSE528

Natural Language Processing

Venue:ADB-405 Topic: PartsOfSpeach Tagging

Prof. Tulasi Prasad Sariki,

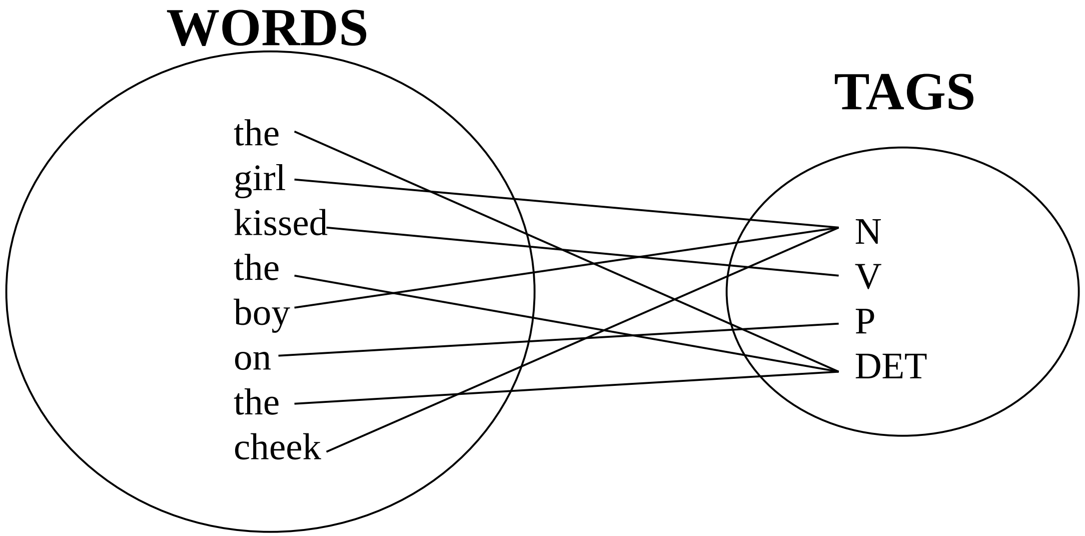
SCSE, VIT Chennai Campus

www.learnersdesk.weebly.com

VIT®
UNIVERSITY
(Estd. u/s 3 of UGC Act 1956)
www.vit.ac.in
Vellore • Chennai
CHENNAI CAMPUS
Vandalur - Kelambakkam Road, Chennai - 600127

Definition

The process of assigning a part-of-speech or other lexical class marker to each word in a corpus.



Definition

- ❑ Annotate each word in a sentence with a part-of-speech marker.
- ❑ Lowest level of syntactic analysis.
- ❑ Useful for subsequent syntactic parsing and word sense disambiguation.
- ❑ Example

John saw the saw and decided to take it to the table.
NNP VBD DT NN CC VBD TO VB PRP IN DT NN

An Example

WORD	**LEMMA**	**TAG**
the	the	+DET
girl	girl	+NOUN
kissed	kiss	+VPAST
the	the	+DET
boy	boy	+NOUN
on	on	+PREP
the	the	+DET
cheek	cheek	+NOUN

English POS Tagsets

- ❑ Original Brown corpus used a large set of 87 POS tags.
- ❑ Most common in NLP today is the Penn Treebank set of 45 tags.
  - ❑Reduced from the Brown set for use in the context of a parsed corpus (i.e. treebank).
- ❑ The C5 tagset used for the British National Corpus (BNC) has 61 tags.

Word Classes

Basic word classes: Noun, Verb, Adjective, Adverb, Preposition, …

Open vs. Closed classes

- Open:
  - Nouns, Verbs, Adjectives, Adverbs.
  - Why “open”?
- Closed:
  - determiners: a, an, the
  - pronouns: she, he, I
  - prepositions: on, under, over, near, by, …

Closed vs. Open Class

Closed class categories are composed of a small, fixed set of grammatical function words for a given language.

- prepositions: on, under, over, …
- particles: up, down, on, off, …
- determiners: a, an, the, …
- pronouns: she, who, I, ..
- conjunctions: and, but, or, …
- auxiliary verbs: can, may should, …

Closed vs. Open Class

Open class categories have large number of words and new ones are easily invented.

- ❑ Nouns new nouns: Internet, website, URL, CD-ROM, email, newsgroup, bitmap, modem, multimedia
- ❑New verbs have also : download, upload, reboot, right-click, double-click,
- ❑ Verbs (Google),
- ❑ Adjectives (geeky)
- ❑ Abverb (chompingly)

English Parts of Speech (Nouns)

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

English Parts of Speech (Nouns)

Proper nouns (Penn, Philadelphia, Davidson)

- ❑ English capitalizes these.

Common nouns (the rest).

Count nouns and mass nouns

- ❑ Count: have plurals, get counted: goat/goats,
- ❑ Mass: don’t get counted (snow, salt, water,)

English Parts of Speech (Verbs)

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)

English Parts of Speech (Adjectives)

Adjective (modify nouns, identify properties or qualities of nouns)

- ❑ Basic (JJ): red, tall
- ❑ Comparative (JJR): redder, taller
- ❑ Superlative (JJS): reddest, tallest

Adjective ordering restrictions in English:

- ❑ Old blue book, ***not*** Blue old book
- ❑ the **44th** president
- ❑ a **green** product
- ❑ a **responsible** investment
- ❑ the **dumbest**, **worst** leader

English Parts of Speech (Adverbs)

Adverb (modify verbs)

- Basic (RB): quickly
- Comparative (RBR): quicker
- Superlative (RBS): quickest

Unfortunately, John walked home extremely slowly yesterday

- Directional/locative adverbs (here, downhill)
- Degree adverbs (extremely, very, somewhat)
- Manner adverbs (slowly, slinkily, delicately)
- Temporal adverbs (yesterday, tomorrow)

English Parts of Speech (Determiner)

Is a word that occurs together with a noun or noun phrase and serves to express the reference of that noun or noun phrase in the context.

That is, a determiner may indicate whether the noun is referring to a definite or indefinite element of a class, to a closer or more distant element, to an element belonging to a specified person or thing, to a particular number or quantity, etc.

English Parts of Speech(Determiner)

Common kinds of determiners include

- definite and indefinite articles (*the,* a, *an)*
- demonstratives (*this*, t*hat, these*)
- possessive determiners (*my*, *their*)
- quantifiers (*many*, *few* , *several*).

English Parts of Speech (preposition)

Preposition (IN): a word governing, and usually preceding, a noun or pronoun and expressing a relation to another word or element in the clause, as in ‘the man ***on*** the platform’, ‘she arrived ***after*** dinner’.

Ex: on, in, by, to, with

English Parts of Speech

Coordinating Conjunction (CC): that connects words, sentences, phrases or clauses.

the truth of nature, ***and*** the power of giving interest

Ex: and, but, or.

Particle (RP): a particle is a function word that must be associated with another word or phrase to impart meaning, i.e., does not have its own lexical definition.

Ex: off (took off), up (put up)

POS tagging

- ❑ POS Tagging is a process that attaches each word in a sentence with a suitable tag from a given set of tags.
- ❑ Tagging is the assignment of a single part-of-speech tag to each word (and punctuation marker) in a corpus.
- ❑The set of tags is called the Tag-set.
- ❑ Standard Tag-set : Penn Treebank (for English).

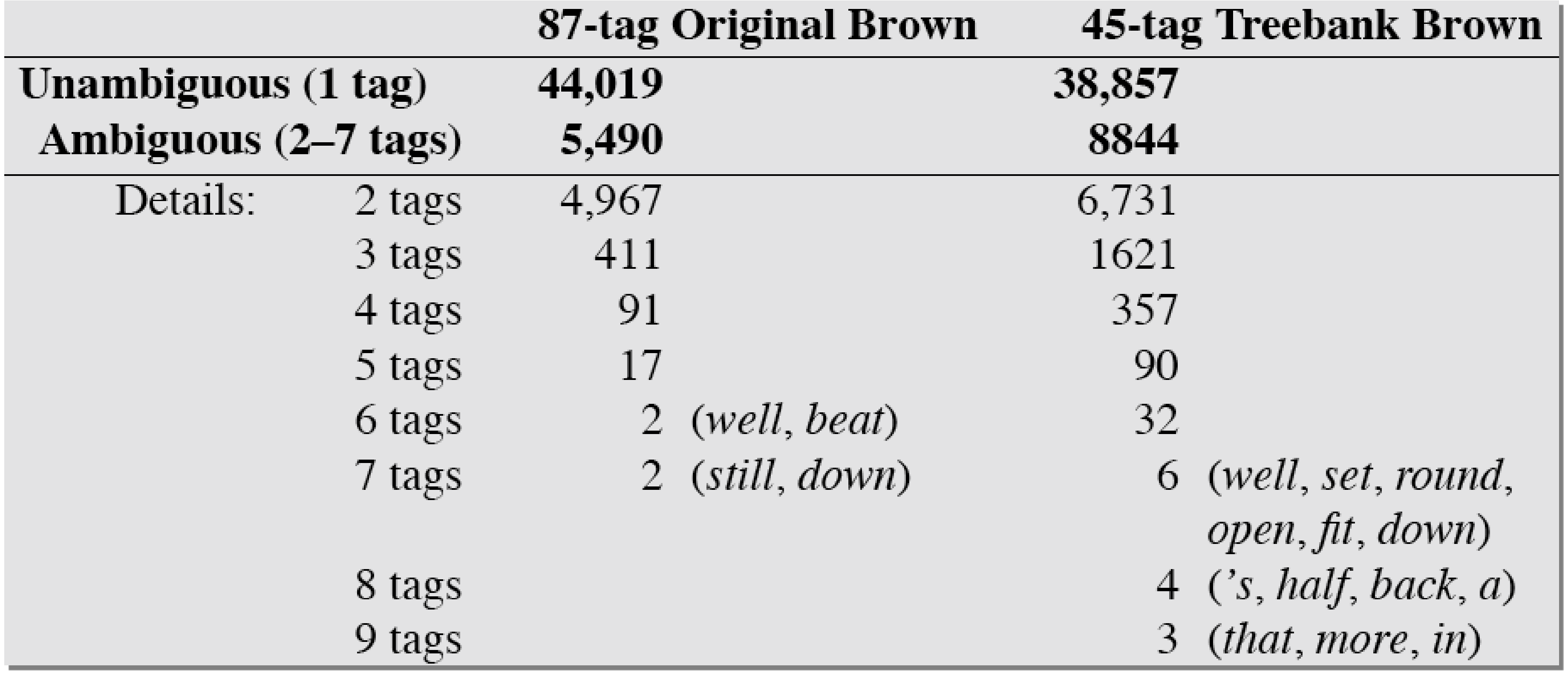
POS tagging

- ❑ There are so many parts of speech, potential distinctions we can draw.
- ❑ To do POS tagging, we need to choose a standard set of tags to work with.
- ❑ Could pick very coarse tag sets.
  - ❑N, V, Adj, Adv.
- ❑ More commonly used set is finer grained (Penn TreeBank, 45 tags)
  - ❑PRP$, WRB, WP$, VBG

POS Tag Ambiguity

- ❑ Deciding on the correct part of speech can be difficult even for people.
- ❑ In English : I bank1 on the bank2 on the river bank3 for my transactions.
  - ❑ Bank1 is verb, the other two banks are nouns
- ❑ In Hindi :
  - ❑ “Khaanaa” : can be noun (food) or verb (to eat)

Measuring Ambiguity



		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*)

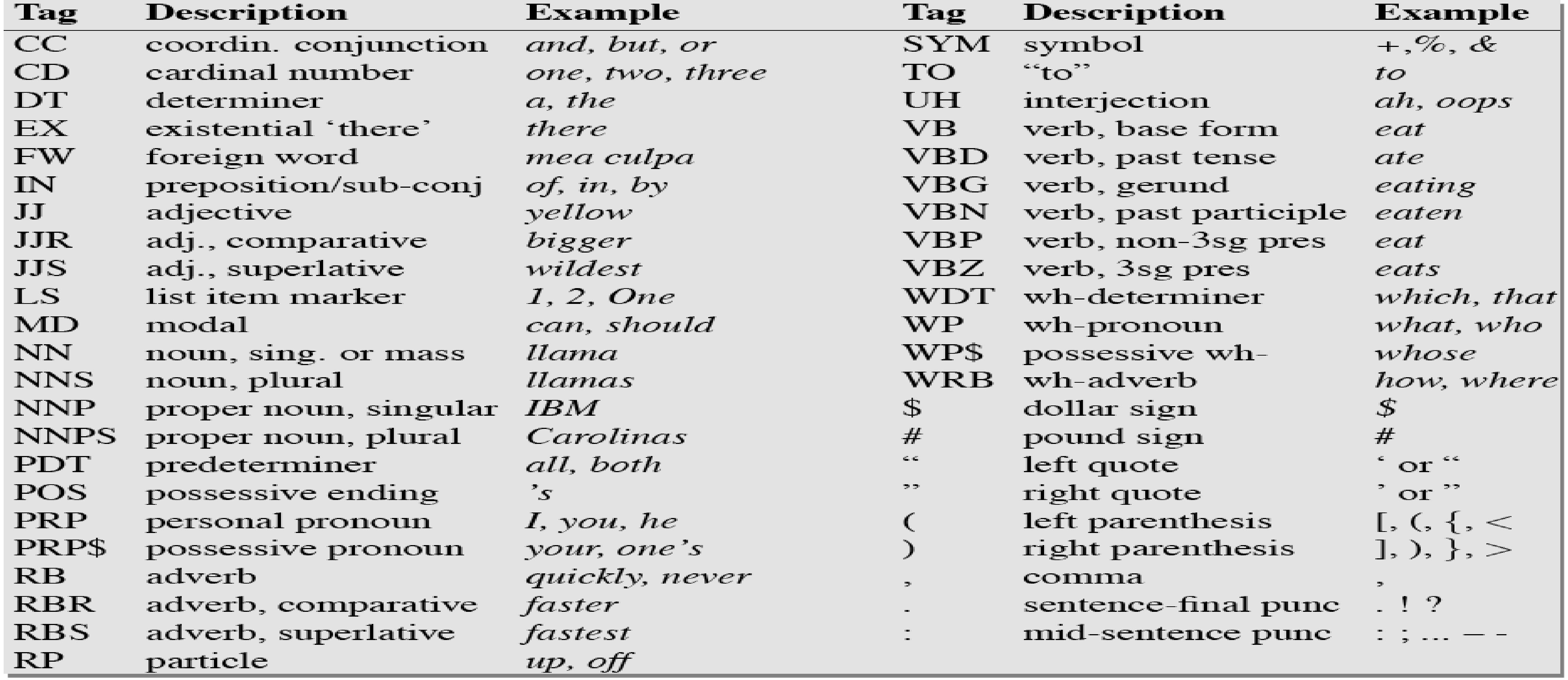
How Hard is POS Tagging?

- ❑ About 11% of the word types in the Brown corpus are ambiguous with regard to part of speech

- ❑ But they tend to be very common words

- ❑ 40% of the word tokens are ambiguous

Penn TreeBank POS Tagset



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*			

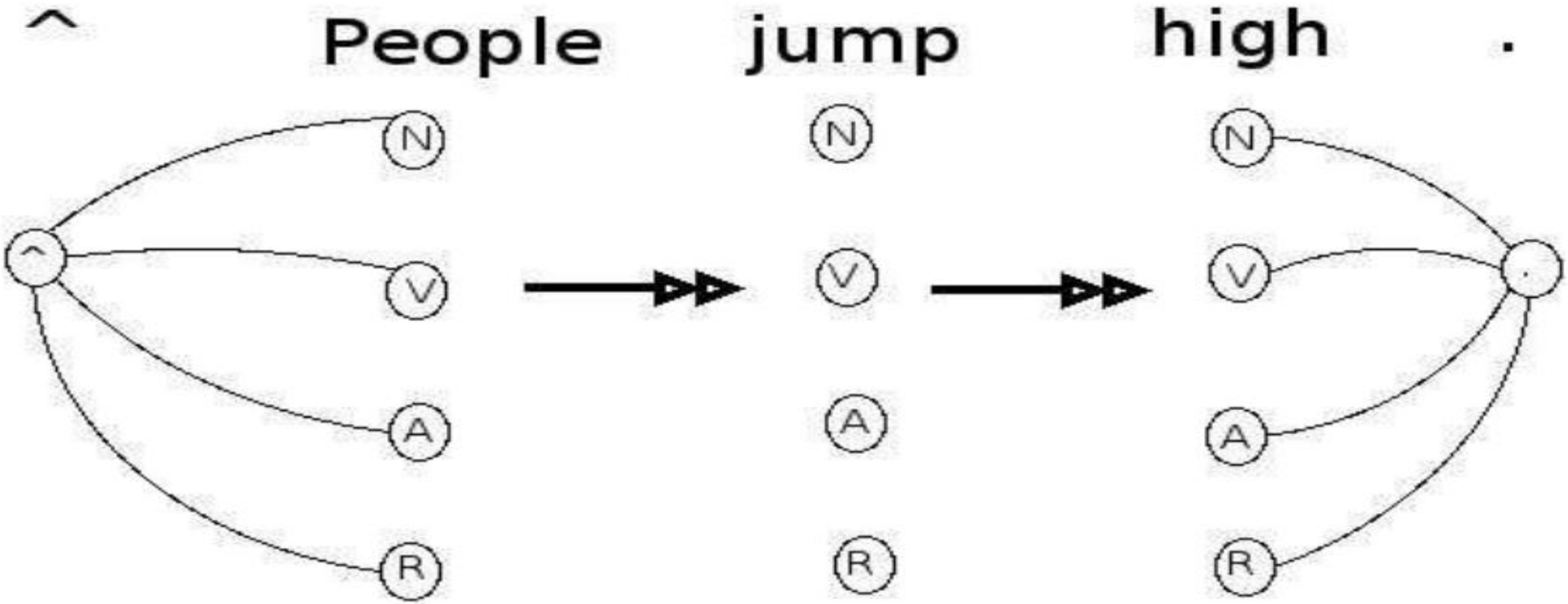
Using the Penn Tagset

- □ The/DT grand/JJ jury/NN commmented/VBD on/IN a/DT number/NN of/IN other/JJ topics/NNS ./.
- □ Prepositions and subordinating conjunctions marked IN ("although/IN I/PRP..")
- □ Except the preposition/complementizer "to" is just marked "TO".

Process

- List all possible tag for each word in sentence.
- Choose best suitable tag sequence.
- Example
  - ”People jump high”.
  - People : Noun/Verb
  - jump : Noun/Verb
  - high : Noun/Verb/Adjective
  - We can start with probabilities.

Example



Why POS

- POS tell us a lot about a word (and the words near it).
  - E.g, adjectives often followed by nouns
  - personal pronouns often followed by verbs
  - possessive pronouns by nouns
- Pronunciations depends on POS, e.g.
  - object (first syllable NN, second syllable VM), content, discount
- First step in many NLP applications

Rule-Based Tagging

- ❑ Start with a dictionary.
- ❑ Assign all possible tags to words from the dictionary.
- ❑ Write rules by hand to selectively remove tags.
- ❑ Leaving the correct tag for each word.

Step1: Start with a Dictionary

she:	PRP
promised:	VBN,VBD
to:	TO
back:	VB, JJ, RB, NN
the:	DT
bill:	NN, VB

Etc… for the ~100,000 words of English with more than 1 tag

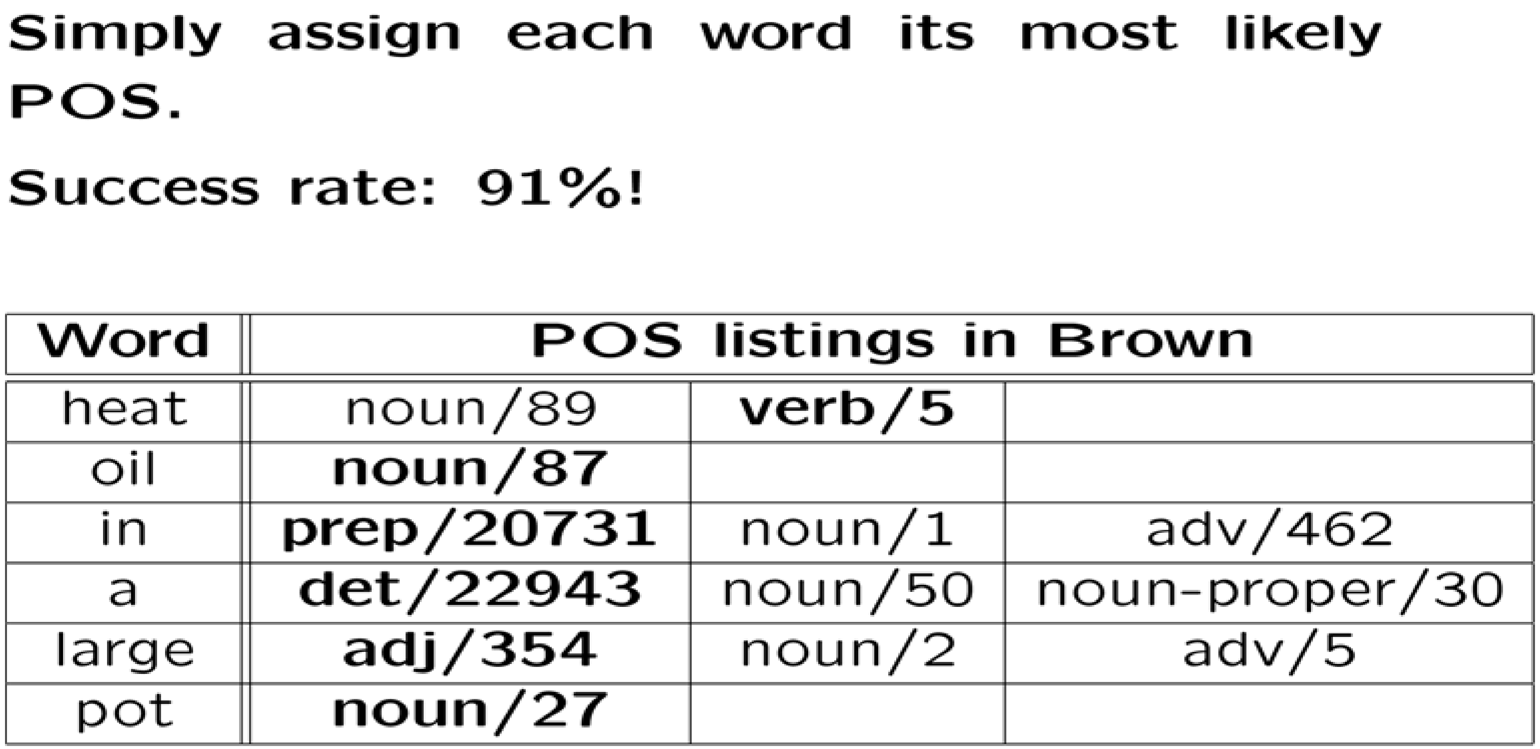
Step2: Assign Every Possible Tag

			NN		
			RB		
	VBN		JJ		VB
PRP	VBD	TO	VB	DT	NN
She	**promised**	**to**	**back**	**the**	**bill**

Step3: Write Rules to Eliminate Tags

Eliminate VBN if VBD is an option when VBN|VBD follows “<start> PRP”

			NN		
			RB		
	~~VBN~~		JJ		VB
PRP	VBD	TO	VB	DT	NN
She	**promised**	**to**	**back**	**the**	**bill**



Simply assign each word its most likely POS.

Success rate: 91%!

Word	**POS listings in Brown**		
heat	noun/89	**verb/5**	
oil	**noun/87**		
in	**prep/20731**	noun/1	adv/462
a	**det/22943**	noun/50	noun-proper/30
large	**adj/354**	noun/2	adv/5
pot	**noun/27**		

END