

# Part-of-Speech Tagging & Sequence Tagging

(slides are modified from Claire Cardie / Ray Mooney)

# Part-of-Speech Tagging

**Assign the correct part of speech (word class) to each word/token in a document**

“The/DT planet/NN Jupiter/NNP and/CC its/PPS moons/NNS are/VBP in/IN effect/NN a/DT mini-solar/JJ system/NN ,/, and/CC Jupiter/NNP itself/PRP is/VBZ often/RB called/VBN a/DT star/NN that/IN never/RB caught/VBN fire/NN ./.”

# 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.
  - Tagset used in these slides.
  - 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.

# Penn Tree Tagset

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

# English Parts of Speech

- Noun (person, place or thing)
  - Singular (NN): dog, fork
  - Plural (NNS): dogs, forks
  - **Proper Noun** (NNP, NNPS): John, Springfields
  - Personal **pronoun** (PRP): I, you, he, she, it
  - Wh-pronoun (WP): who, what

# English Parts of Speech

- Verb (actions and processes)
  - Base, infinitive (VB): eat
  - Past tense (VBD): ate
  - Gerund (VBG): eating
  - Past participle (VBN): eaten
  - Non 3<sup>rd</sup> person singular present tense (VBP): eat
  - 3<sup>rd</sup> person singular present tense: (VBZ): eats
  - Modal (MD): should, can

# English Parts of Speech (cont.)

- Adjective (modify nouns)
  - Basic (**JJ**): red, tall
  - Comparative (JJR): redder, taller
  - Superlative (JJS): reddest, tallest
- Adverb (modify verbs)
  - Basic (**RB**): quickly
  - Comparative (RBR): quicker
  - Superlative (RBS): quickest

# English Parts of Speech (cont.)

- I am going **to** go **to** school.
- I will take **over** the world.
- I will place it **over** there.



# English Parts of Speech (cont.)

- Preposition (**IN**): on, in, by, to, with
- To (**TO**): as in “to eat”
- **Determiner (Article)**:
  - Basic (**DT**) a, an, the
  - WH-determiner (WDT): which, that
- Coordinating Conjunction (**CC**): and, but, or
- **Particle (RP)**: off (took off), up (put up)

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

# Function Words / Content Words

- **Function words (closed class words)**
  - words that have little lexical meaning
  - express grammatical relationships with other words
  - Prepositions (in, of, etc), pronouns (she, we, etc), auxiliary verbs (would, could, etc), articles (a, the, an), conjunctions (and, or, etc)
- **Content words (open class words)**
  - Nouns, verbs, adjectives, adverbs etc
  - Easy to invent a new word (e.g. “google” as a noun or a verb)
- **Stop words**
  - Similar to function words, but may include some content words that carry little meaning with respect to a specific NLP application (e.g., “have”, “want”, “get”, etc)

# Part-of-Speech Tagging

- Needed as an initial processing step for a number of NLP applications.
- Among easiest of NLP problems
  - State-of-the-art methods achieve ~97% accuracy.
  - Simple heuristics can go a long way.
  - ~90% accuracy just by choosing the most frequent tag for a word (**MLE**)

# POS Tagging Tools

- Online (and Offline)
  - <http://cogcomp.cs.illinois.edu/demo/pos/?id=4>
  - <http://nlp.stanford.edu:8080/parser/>
- Offline
  - CLAWS
  - LingPipe
  - OpenNLP
  - CRFTagger

# Ambiguity in POS Tagging

- **Particle (RP) vs. preposition (IN)**
  - He talked *over* the deal.
  - He talked *over* the telephone.
- **past tense (VBD) vs. past participle (VBN)**
  - The horse *walked* past the barn.
  - The horse *walked* past the barn fell.
- **noun vs. adjective?**
  - The *executive* decision.
- **noun vs. present participle**
  - *Fishing* can be fun

# Ambiguity in POS Tagging

- “Like” can be a verb or a preposition
  - I **like**/**VB**P candy.
  - Time flies **like**/**IN** an arrow.
- “Around” can be a preposition, particle, or adverb
  - I bought it at the shop **around**/**IN** the corner.
  - I never got **around**/**RP** to getting a car.
  - A new Prius costs **around**/**RB** \$25K.

# 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.
  - Hidden Markov Model (HMM), Maximum Entropy Markov Model (MEMM), Conditional Random Field (CRF), Transformation Based Learning (TBL)
- Generally, learning-based approaches have been found to be more effective overall, taking into account the total amount of human expertise and effort involved.



# Sequence Labeling as Classification

- Today we will see three different ways of labeling a sequence by casting the problem as a simple “classification”.
- Make a separate classification for each word, using *local context* (words in a fixed window).

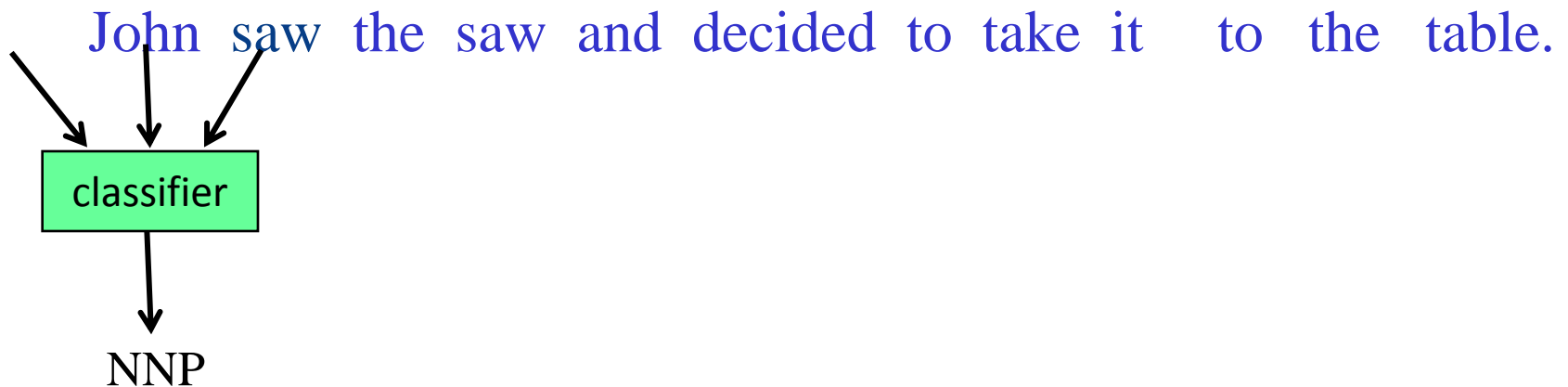
# Sequence Labeling as Classification

## -- First Trial

- Classify each token independently but use as input features, information about the surrounding tokens (sliding window)

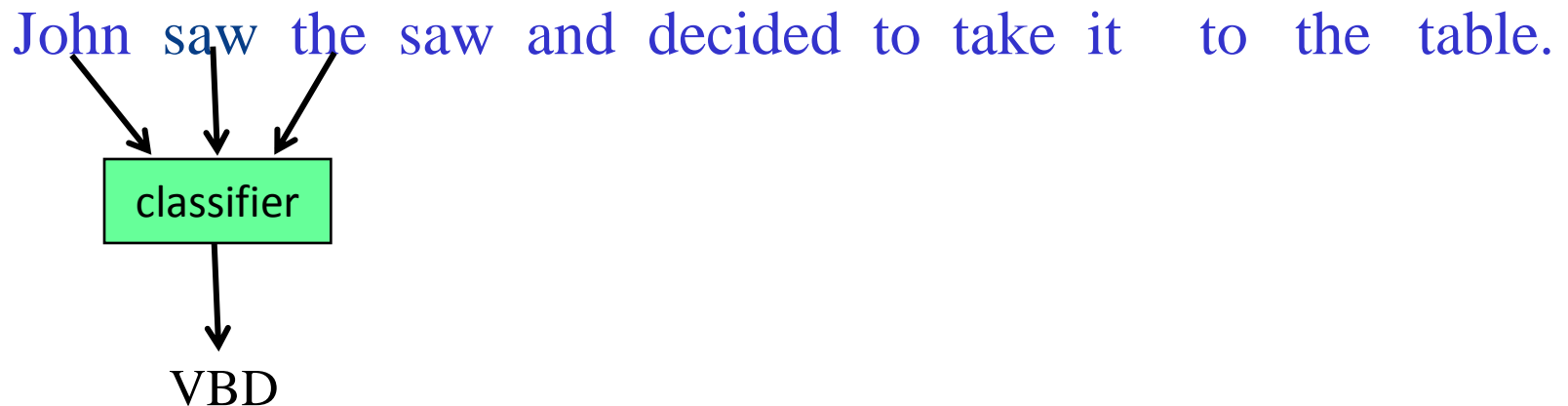
# Sequence Labeling as Classification

## -- First Trial



# Sequence Labeling as Classification

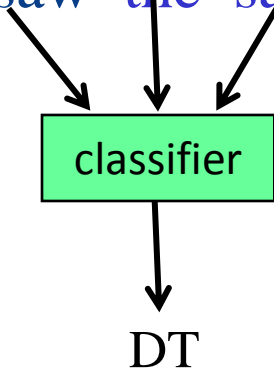
- Classify each token independently but use as input features, information about the surrounding tokens (sliding window).



# Sequence Labeling as Classification

## -- First Trial

John saw the saw and decided to take it to the table.



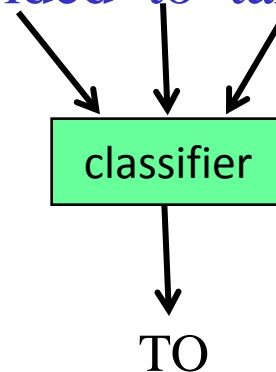
classifier

DT

# Sequence Labeling as Classification

## -- First Trial

John saw the saw and decided to take it to the table.



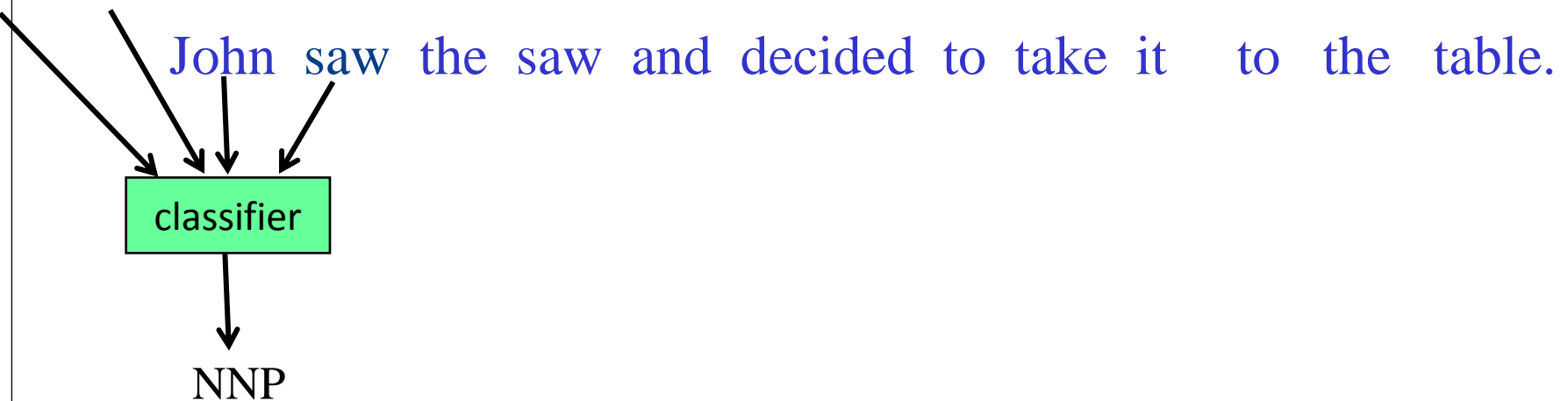
# Sequence Labeling as Classification

-- Second Trial: previous output as features

- Previous approach makes each decision independently from all other decisions.
  - ➔ The output of neighboring words can provide extra information for the current decision.
  - ➔ Use the output of previous decision as extra features for the current decision.

# Sequence Labeling as Classification

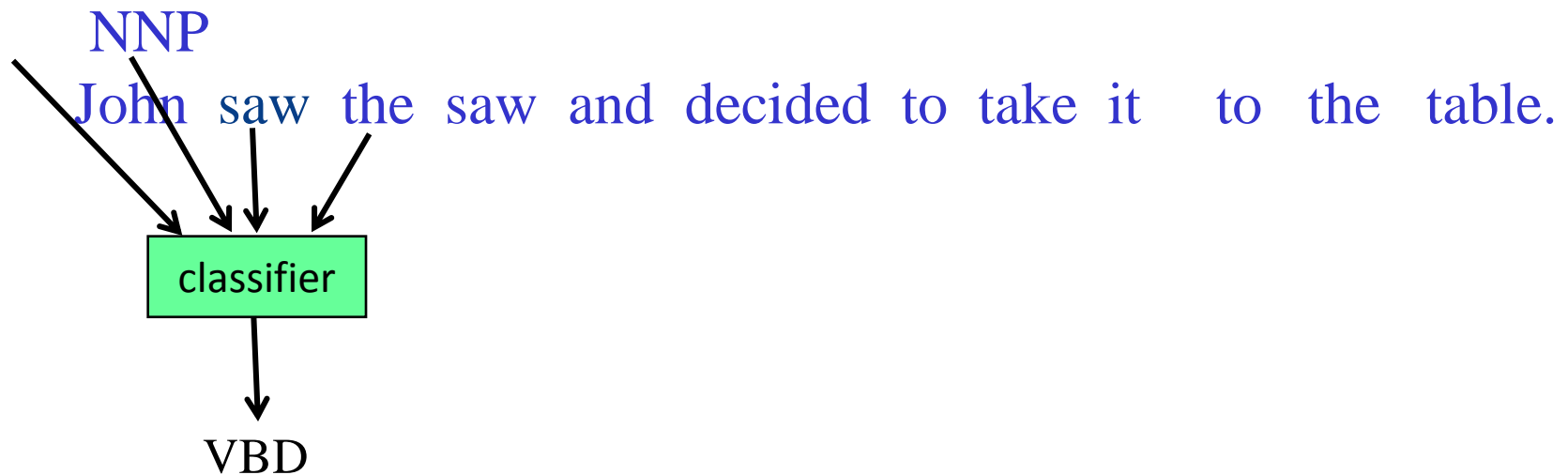
-- Second Trial: previous output as features





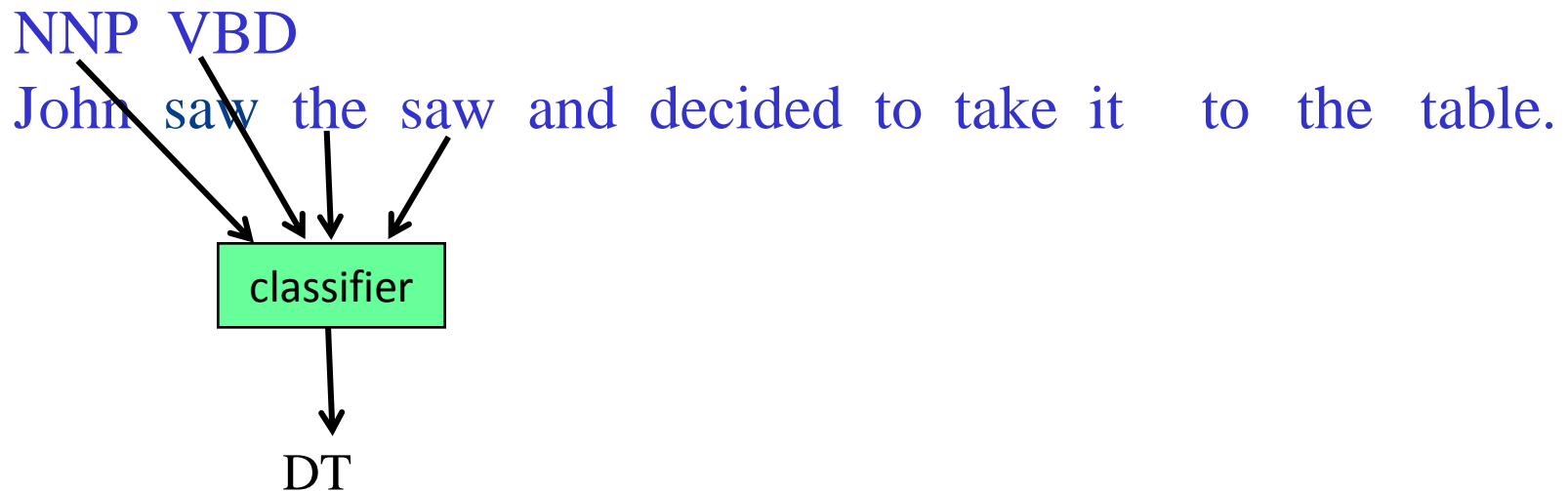
# Sequence Labeling as Classification

-- Second Trial: previous output as features



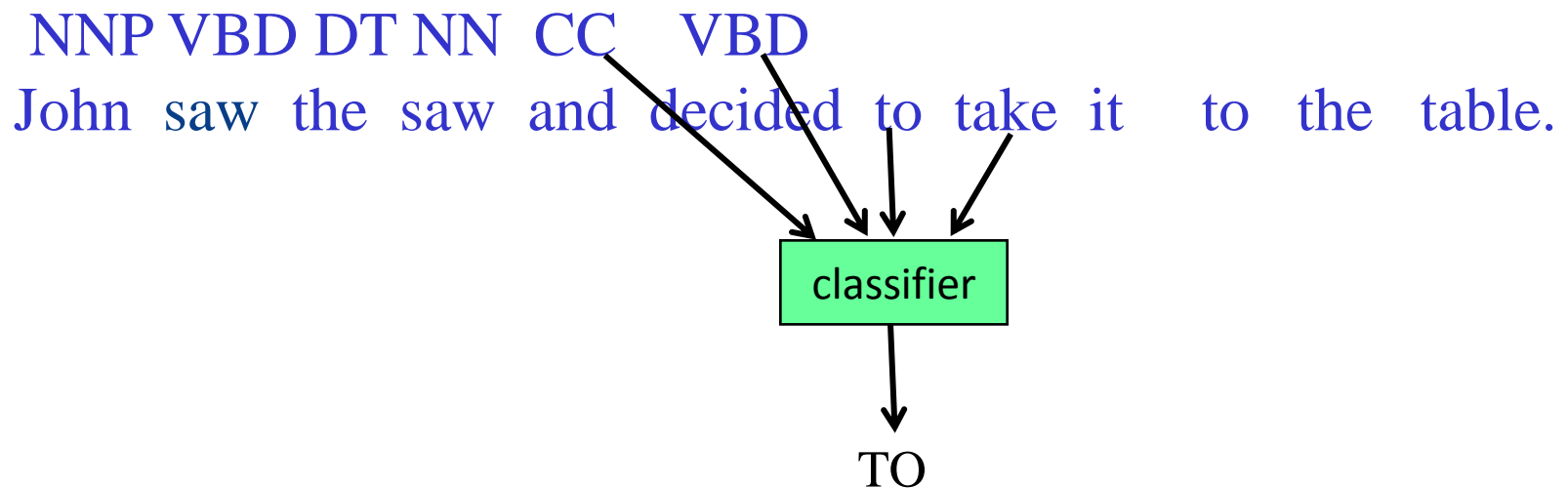
# Sequence Labeling as Classification

-- Second Trial: previous output as features



# Sequence Labeling as Classification

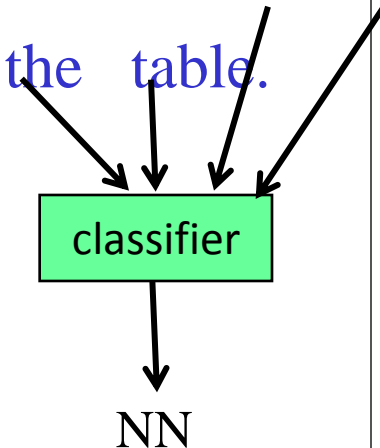
-- Second Trial: previous output as features



# Sequence Labeling as Classification

-- Third Trial: previous output as features

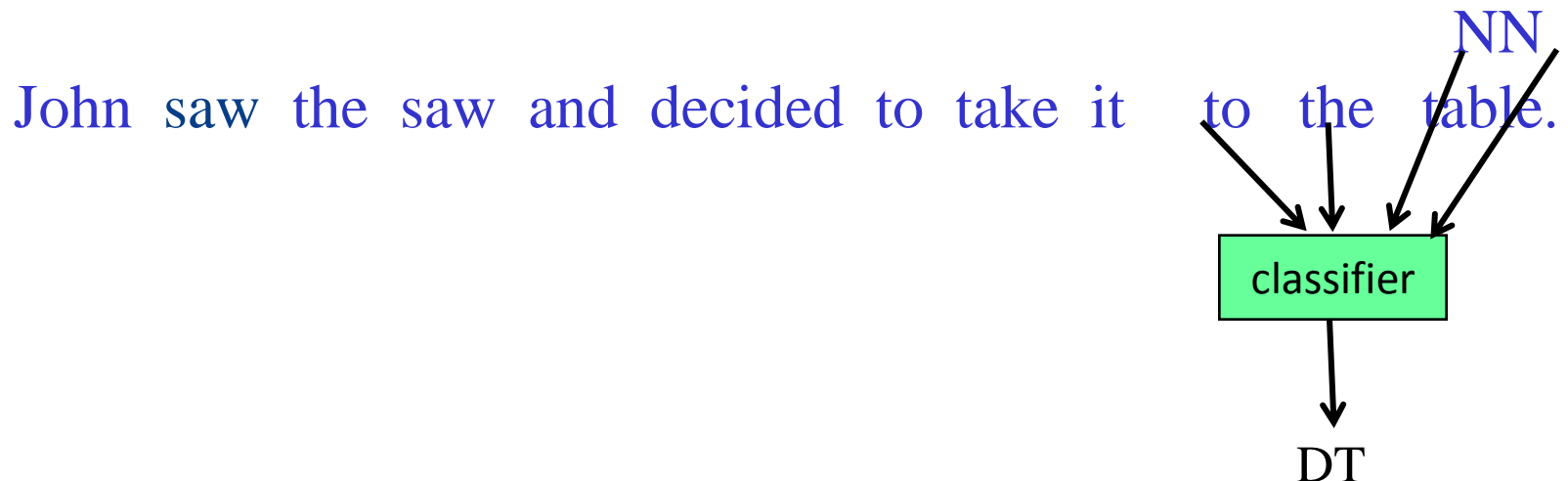
John saw the saw and decided to take it to the table.



- Disambiguating “to” would be even easier if we process the sequence backward.

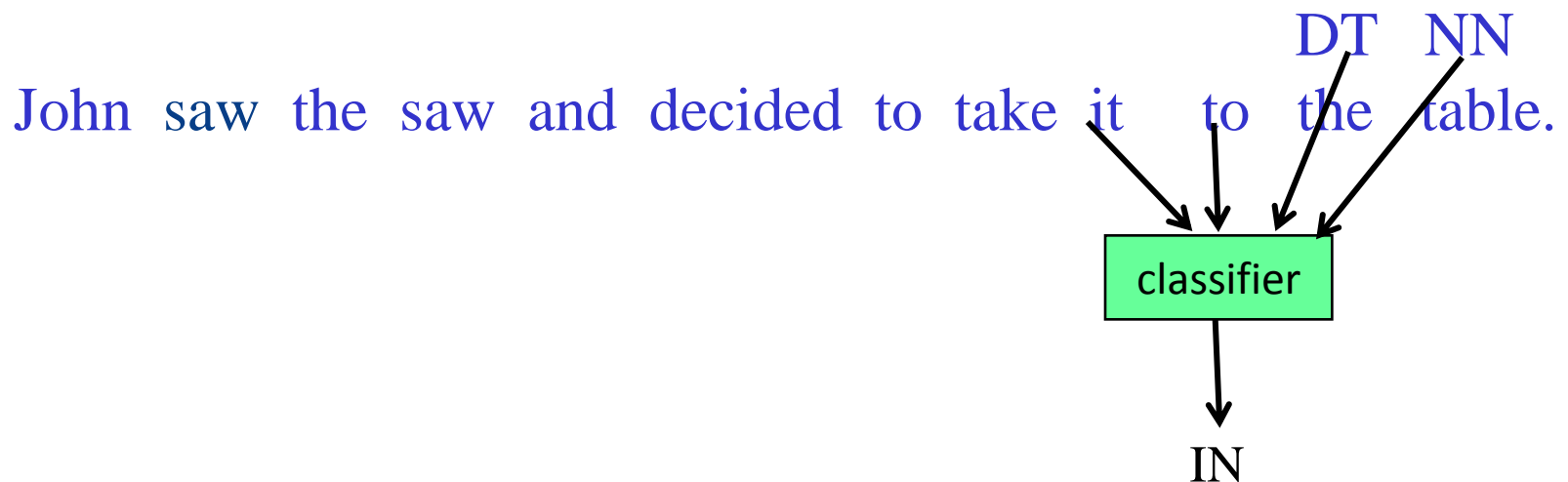
# Sequence Labeling as Classification

-- Third Trial: previous output as features



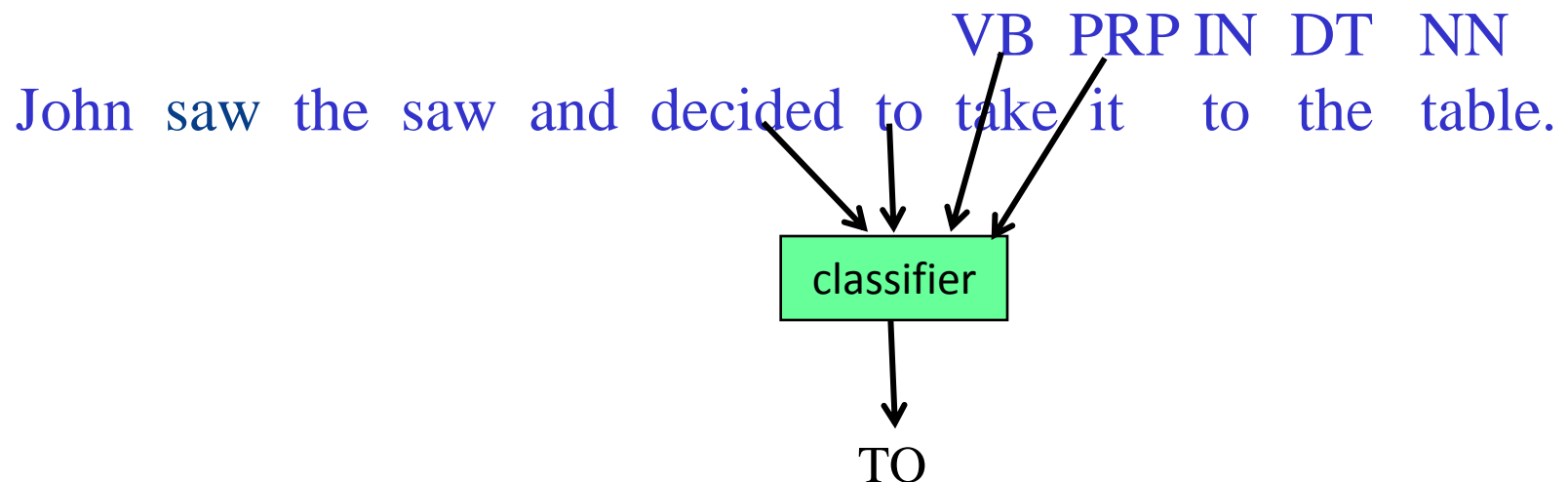
# Sequence Labeling as Classification

-- Third Trial: previous output as features



# Sequence Labeling as Classification

-- Third Trial: previous output as features



# Problems with Sequence Labeling as Classification

- Not easy to integrate information from the output of neighboring words from both directions
- Difficult to propagate uncertainty between decisions and “collectively” determine the most likely joint assignment of categories to all of the tokens in a sequence.
- Once you make a decision (classification) for each word, that’s the final decision for that word – that is, you don’t go back to words for which you have already made decisions in order to fix your previous decisions.
  - ➔ Why would you want to change your mind?



The horse *walked* past ...

The horse *walked* past ...



past tense (VBD)

The horse *walked* past the barn **fell**.



~~past tense (VBD)~~

past particle (VBN)

# Probabilistic Sequence Models

- Probabilistic sequence models allow integrating uncertainty over multiple, interdependent classifications and collectively determine the most likely global assignment.
- Two standard models
  - Hidden Markov Model (HMM)
  - Conditional Random Field (CRF)

# Sequence Tagging problems in NLP

- Part-of-speech tagging
- Information Extraction
  - Named Entity Recognition
    - Person
    - Organization
    - Company
    - Nation
    - Time

# Information Extraction Example-1

- Task: Find location, time, speaker of the talk from the email announcement

The first talk of this year's Distinguish Lecture Series will be given tomorrow (Friday, 9/17) at 2:30pm in CEWIT 200 by Ed Felten of Princeton University. A catered reception will follow. All Computer Science students and faculty are invited to attend.

# Information Extraction Example-1

- Task: Find location, time, speaker of the talk from the email announcement

The first talk of this year's Distinguish Lecture Series will be given tomorrow (Friday, 9/17) at 2:30pm [time] in CEWIT 200 [location] by Ed Felten [speaker] of Princeton University. A catered reception will follow. All Computer Science students and faculty are invited to attend.

# Information Extraction Example-2

- Task: Find opinion expressions (phrases that contain opinions), sources of opinions, targets of opinions.

In a statement headed 'The Tyrant Visits Tirana' carried by the Cuban news agency, Castro slammed Bush for voicing support for Kosovo's independence "without the least respect for the interests of Serbia and Russia.

...



# Information Extraction Example-2

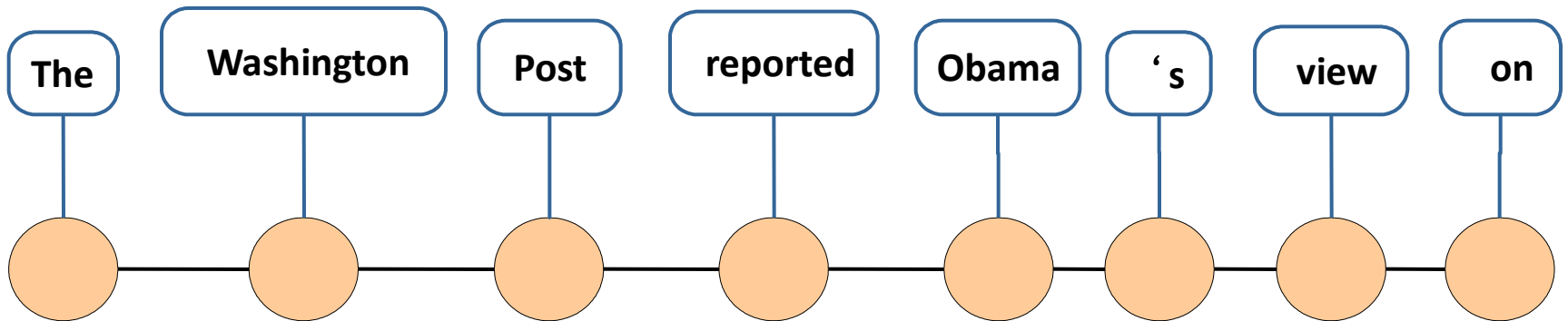
- Task: Find opinion expressions (phrases that contain opinions), sources of opinions, targets of opinions.

In a statement headed 'The Tyrant Visits Tirana' carried by the [Cuban news agency], [Castro] slammed [Bush] for voicing support for [Kosovo]'s independence "without the least respect for the interests of [Serbia] and [Russia].

...

# Sequence Tagging

**<The Washington Post>** reported **<Obama>**'s view on the oil crisis.



# Sequence Tagging

- Prediction using **BIO** tagging

<The Washington Post> reported <Obama>'s view on the oil crisis.

