

# CS 2731 Introduction to Natural Language Processing

## Session 18: POS tagging, NER, HMMs

---

Michael Miller Yoder

October 30, 2023

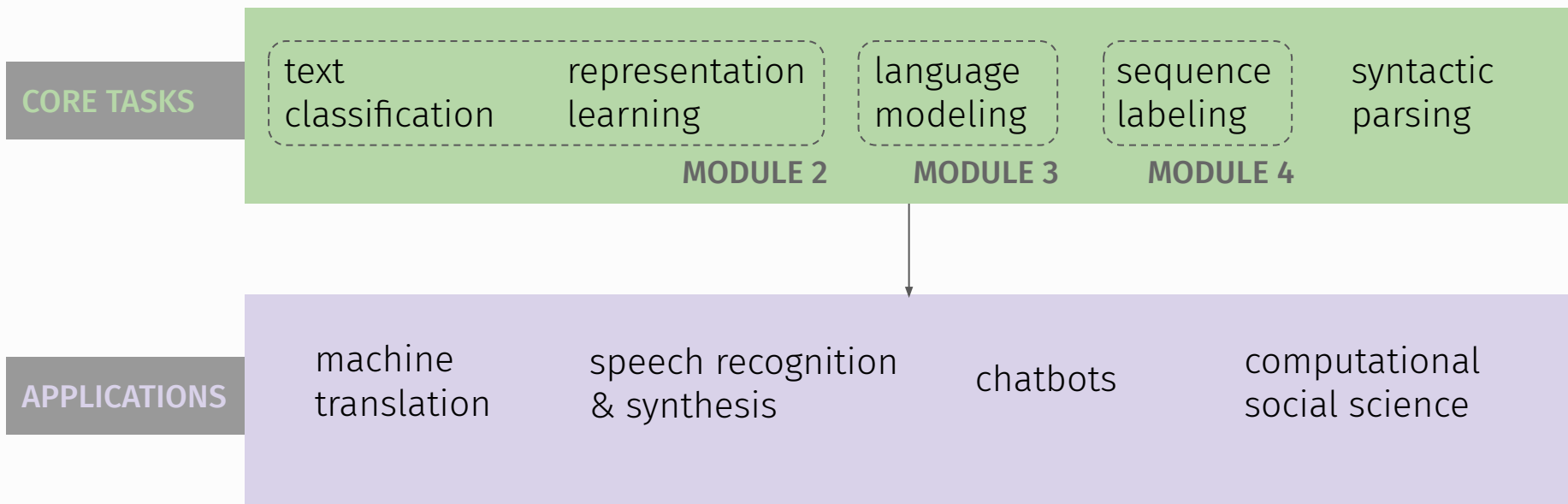
# Course logistics: homeworks

- Homework 2 contest winners!
  - LR with features
    - Tom with 73.7 accuracy on test set
    - *Runner-up*: Birju with 72.4
  - FNN with static word embeddings
    - Ben with 70.9
    - *Runner-up*: RJ with 70.8
- [Homework 3](#) is due **this Thu 11-02 at midnight**
  - Updates: add-one smoothing is now optional, for **extra credit**
  - (Most of) an implementation for perplexity is provided
  - Ask questions in the Canvas discussion forum (or can email)
- Homework 4 will be released today. Is **due Thu 11-09**

# Course logistics

- Pantho's office hours next week will be **Tuesday 2:45-3:45pm** instead of Thursday. Is this better every week?
- Next project milestone is a basic working system **due Thu 11-16**

# Core tasks and applications of NLP



# Overview: POS tagging, NER, HMMs

- Parts of speech
- Part-of-speech (POS) tagging
- Named entity recognition (NER)
- Hidden Markov Models (HMMs)

# Parts of speech

---

My cat who lives dangerously no longer  
has nine lives.

My cat who **lives** dangerously no longer  
has nine **lives**.



My cat who **lives** dangerously no longer  
has nine **lives**.

**lives** /lɪvz/ verb

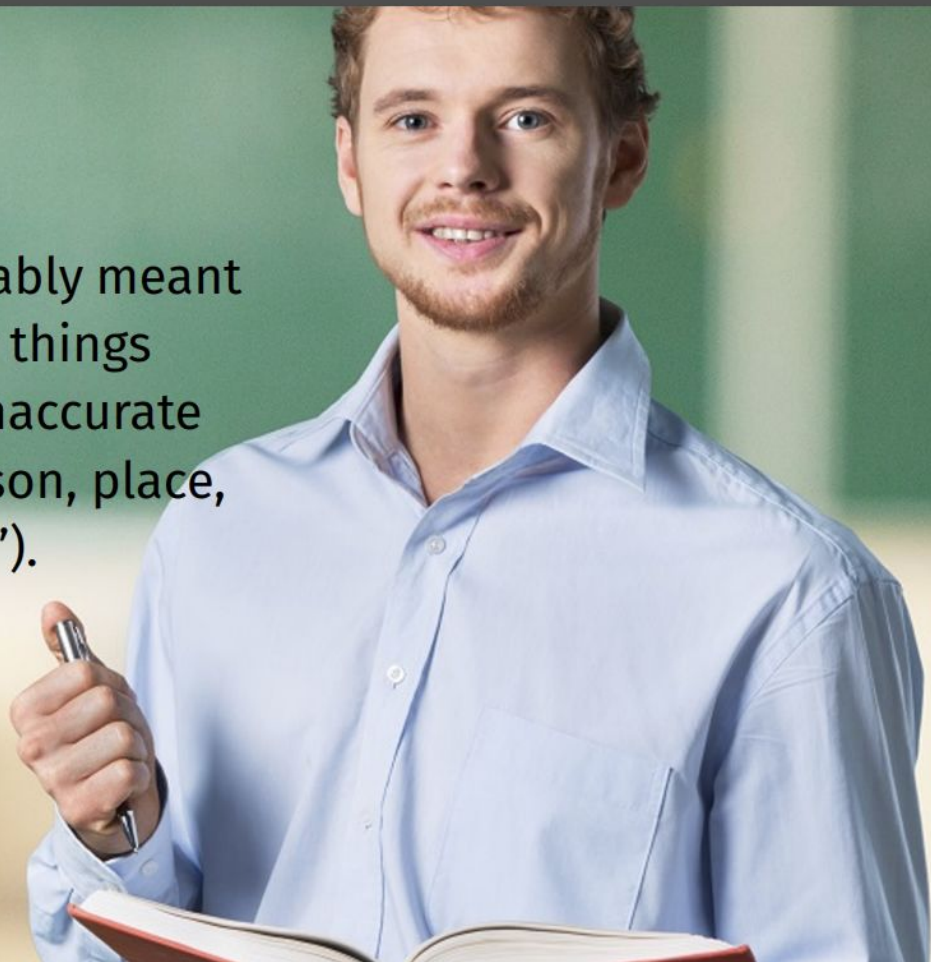
**lives** /ləjvz/ noun

# Examples of Parts of Speech

PART OF SPEECH	EXAMPLES
<b>noun</b>	dog, cat, professor, exam, fear, loathing, oppression, void, text, Bavarian
<b>verb</b>	enjoy, walk, finish, trust, hug, like, understand, be, text, drink
<b>adjective</b>	nice, happy, red, exciting, ludicrous, funny, ancient, Bavarian
<b>adverb</b>	slowly, quickly, shrewdly, foolishly, boisterously, undercover, yesterday
<b>preposition</b>	to, for, from, under, by
<b>auxiliary verbs</b>	be, have, must, might, will, would
<b>determiner</b>	the, a(n), this, that, my, her
<b>pronouns</b>	he, she, it, this, that
<b>conjunctions</b>	and, but, however, nevertheless, so

# Your English Teacher Was a Well-Intentioned Liar

Your English teacher probably meant well, but taught you many things about language that are inaccurate (like that a noun is a “person, place, thing, or abstract concept”).



*Slide credit: David Mortensen*

# Criteria for Parts of Speech

Remember the early 20th century American linguists who wanted to document endangered languages? They wanted to define parts of speech in an objective, language-neutral way, so they defined them **distributionally**. This works better than the semantic criteria that your English teacher taught you.

**morphology** What is the distribution of morphemes within these words?

Same POS  $\Rightarrow$  similar morphemes

**syntax** What is the distribution of words within phrases and sentences?

Same POS  $\Rightarrow$  similar roles/contexts

American Structuralists called these “form classes” but we call them “lexical classes” or “grammatical classes” or “parts of speech”

# Open Class Parts of Speech

Classes to which neologisms are readily added. In English:

<b>nouns</b>	can be both subjects and objects of verbs and objects of prepositions, (usually) be singular or plural, have determiners, be modified by adjectives, and be possessed
<b>verbs</b>	can take noun phrases as arguments and tense morphology and can be modified by adverbs
<b>adjectives</b>	can modify nouns and take comparative and superlative morphology where allowed by prosody
<b>adverbs</b>	can modify verbs, adjectives, or other adverbs



# Closed Class Parts of Speech

Classes to which neologisms are not readily added. In English:

<b>prepositions</b>	occur before noun phrases, connecting them syntactically to larger phrases
---------------------	--

<b>determiners</b>	occur at the beginning of noun phrases
--------------------	--

<b>conjunction</b>	join phrases, clauses, and sentences
--------------------	--------------------------------------

<b>auxiliary verbs</b>	occur before (non-finite) main verbs
------------------------	--------------------------------------

<b>particles</b>	are associated with a verb and are “moveable” (e.g. <i>He tore <b>off</b> his shirt</i> versus <i>He tore his shirt <b>off</b></i> )
------------------	--

<b>numerals</b>	are distributed in some ways like nouns and in others like adjectives
-----------------	---

# What about pronouns?

- Pronouns are generally considered, in English, to be a closed class—it is not easy to add new items to it.
- What are we to make of **neopronouns** like *xe* and *xem* or *ze* and *hir*?
- Their existence suggests that pronouns are not a completely closed class
  - Social movements can change grammar!
  - But it is difficult due to anti-transgender attitudes and to pronouns being a rather closed class in English
- In some languages (e.g., Thai) pronouns clearly are an open class

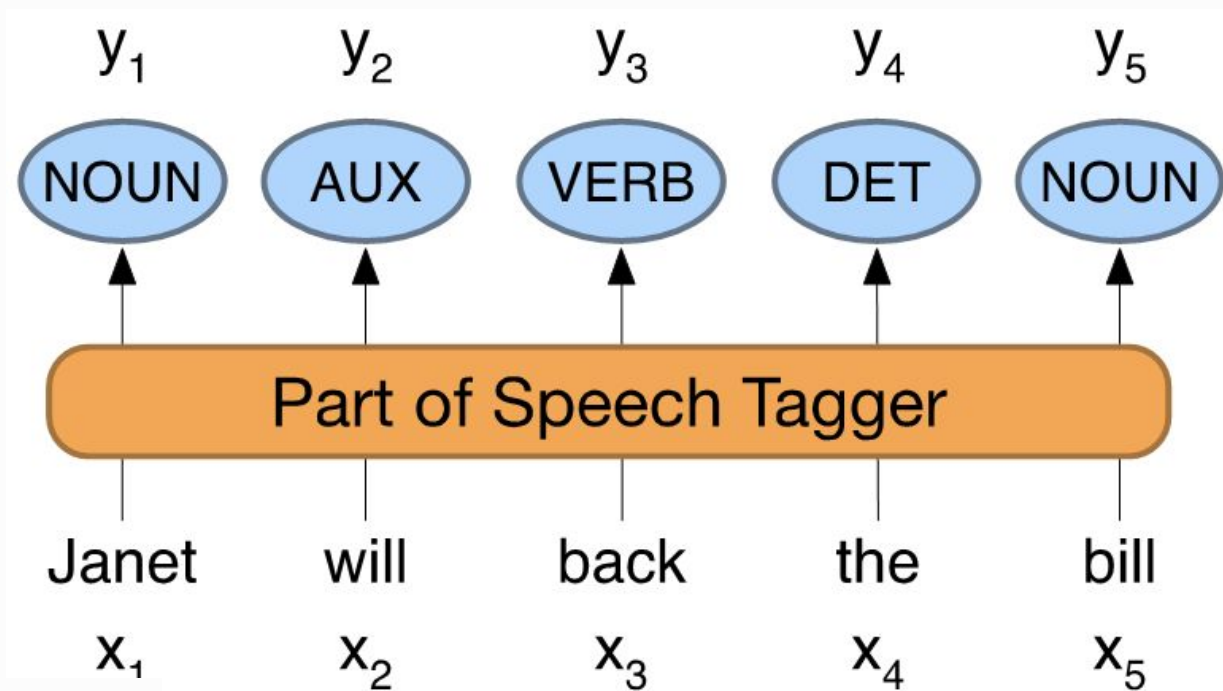
# Part of speech (POS) tagging

---



# Part-of-speech tagging

Map from sequence  $x_1, \dots, x_n$  of words to  $y_1, \dots, y_n$  of POS tags



# “Universal Dependencies” tagset [Nivre et al. 2016]

	Tag	Description	Example
Open Class	<b>ADJ</b>	Adjective: noun modifiers describing properties	<i>red, young, awesome</i>
	<b>ADV</b>	Adverb: verb modifiers of time, place, manner	<i>very, slowly, home, yesterday</i>
	<b>NOUN</b>	words for persons, places, things, etc.	<i>algorithm, cat, mango, beauty</i>
	<b>VERB</b>	words for actions and processes	<i>draw, provide, go</i>
	<b>PROPN</b>	Proper noun: name of a person, organization, place, etc..	<i>Regina, IBM, Colorado</i>
	<b>INTJ</b>	Interjection: exclamation, greeting, yes/no response, etc.	<i>oh, um, yes, hello</i>
Closed Class Words	<b>ADP</b>	Adposition (Preposition/Postposition): marks a noun’s spacial, temporal, or other relation	<i>in, on, by under</i>
	<b>AUX</b>	Auxiliary: helping verb marking tense, aspect, mood, etc.,	<i>can, may, should, are</i>
	<b>CCONJ</b>	Coordinating Conjunction: joins two phrases/clauses	<i>and, or, but</i>
	<b>DET</b>	Determiner: marks noun phrase properties	<i>a, an, the, this</i>
	<b>NUM</b>	Numeral	<i>one, two, first, second</i>
	<b>PART</b>	Particle: a preposition-like form used together with a verb	<i>up, down, on, off, in, out, at, by</i>
	<b>PRON</b>	Pronoun: a shorthand for referring to an entity or event	<i>she, who, I, others</i>
	<b>SCONJ</b>	Subordinating Conjunction: joins a main clause with a subordinate clause such as a sentential complement	<i>that, which</i>
Other	<b>PUNCT</b>	Punctuation	<i>; , ()</i>
	<b>SYM</b>	Symbols like \$ or emoji	<i>\$, %</i>
	<b>X</b>	Other	<i>asdf, qwfg</i>

# Penn TreeBank tagset for English

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>			

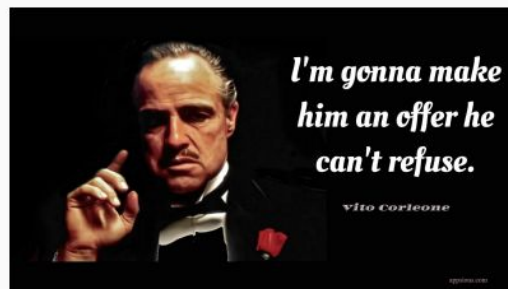
# Why part of speech tagging?

- Can be useful for other NLP tasks
  - Parsing: POS tagging can improve syntactic parsing
  - MT: reordering of adjectives and nouns (say from Spanish to English)
  - Sentiment or affective tasks: may want to distinguish adjectives or other POS
  - Text-to-speech (how do we pronounce “lead” or “object”?)
- Or linguistic or language-analytic computational tasks
  - Need to control for POS when studying linguistic change like creation of new words, or meaning shift
  - Or control for POS in measuring meaning similarity or difference

# POS Tagging is a Disambiguation Task

Consider the following sentences:

I	'm	gonna	make	him	an	offer	he	can	't	refuse
PRO	V	AUX	V	PRO	DET	N	PRO	AUX	ADV	V
			N			V				N



There are eight different ways of tagging this sentence if words are taken out of context. POS Tagging task: **choose the best of these.**

# How difficult is POS tagging in English?

Roughly 15% of word types are ambiguous

- Hence 85% of word types are unambiguous
- *Janet* is always PROP, *hesitantly* is always ADV

But those 15% tend to be very common.

So ~60% of word tokens are ambiguous

E.g., *back*

earnings growth took a *back*/ADJ seat

a small building in the *back*/NOUN

a clear majority of senators *back*/VERB the bill

enable the country to buy *back*/PART debt

I was twenty-one *back*/ADV then

# Sources of information for POS tagging

Janet will back the bill  
AUX/NOUN/VERB? NOUN/VERB?

Prior probabilities of word/tag

- "will" is usually an AUX

Identity of neighboring words

- "the" means the next word is probably not a verb

Morphology and wordshape:

- Prefixes                      unable:                      un- → ADJ
- Suffixes                      importantly:                      -ly → ADJ
- Capitalization                      Janet:                      CAP → PROP

# Standard algorithms for POS tagging

Supervised Machine Learning Algorithms:

- Hidden Markov Models
- Conditional Random Fields (CRFs)
- Neural sequence models (RNNs or Transformers)
- Large Language Models (like BERT), finetuned

All required a hand-labeled training set, all about equal performance (97% on English)

All make use of information sources we discussed

- Via human created features: HMMs and CRFs
- Via representation learning: Neural LMs



# Named entity recognition (NER)

---

# Named entities

- **Named entity**, in its core usage, means anything that can be referred to with a proper name. Most common 4 tags:
  - **PER** (Person): “Marie Curie”
  - **LOC** (Location): “New York City”
  - **ORG** (Organization): “Stanford University”
  - **GPE** (Geo-Political Entity): “Boulder, Colorado”
- Often multi-word phrases
- But the term is also extended to things that aren't entities:
  - dates, times, prices

# Named entity tagging

The task of named entity recognition (NER):

- find spans of text that constitute proper names
- tag the type of the entity.

# NER output

Citing high fuel prices, [ORG United Airlines] said [TIME Friday] it has increased fares by [MONEY \$6] per round trip on flights to some cities also served by lower-cost carriers. [ORG American Airlines], a unit of [ORG AMR Corp.], immediately matched the move, spokesman [PER Tim Wagner] said. [ORG United], a unit of [ORG UAL Corp.], said the increase took effect [TIME Thursday] and applies to most routes where it competes against discount carriers, such as [LOC Chicago] to [LOC Dallas] and [LOC Denver] to [LOC San Francisco].

# Why NER?

- Sentiment analysis: consumer sentiment toward a particular company or person?
- Question Answering: answer questions about an entity?
- Information Extraction: Extracting facts about entities from text.

# Why NER is hard

## 1) Segmentation

- In POS tagging, no segmentation problem since each word gets one tag.
- In NER we have to find and segment the entities!

## 2) Type ambiguity

[PER Washington] was born into slavery on the farm of James Burroughs.  
[ORG Washington] went up 2 games to 1 in the four-game series.  
Blair arrived in [LOC Washington] for what may well be his last state visit.  
In June, [GPE Washington] passed a primary seatbelt law.

# BIO tagging [Ramshaw and Marcus 1995]

How can we turn this structured problem into a sequence problem like POS tagging, with one label per word?

[PER Jane Villanueva] of [ORG United Airlines Holding] , a unit of [ORG United Airlines Holding] , said the fare applies to the [LOC Chicago ] route.

Words	BIO Label
Jane	B-PER
Villanueva	I-PER
of	O
United	B-ORG
Airlines	I-ORG
Holding	I-ORG
discussed	O
the	O
Chicago	B-LOC
route	O
.	O

# BIO tagging

B: token that *begins* a span

I: tokens *inside* a span

O: tokens outside of any span

# of tags (where  $n$  is #entity types):

1 O tag,

$n$  B tags,

$n$  I tags

total of  $2n+1$

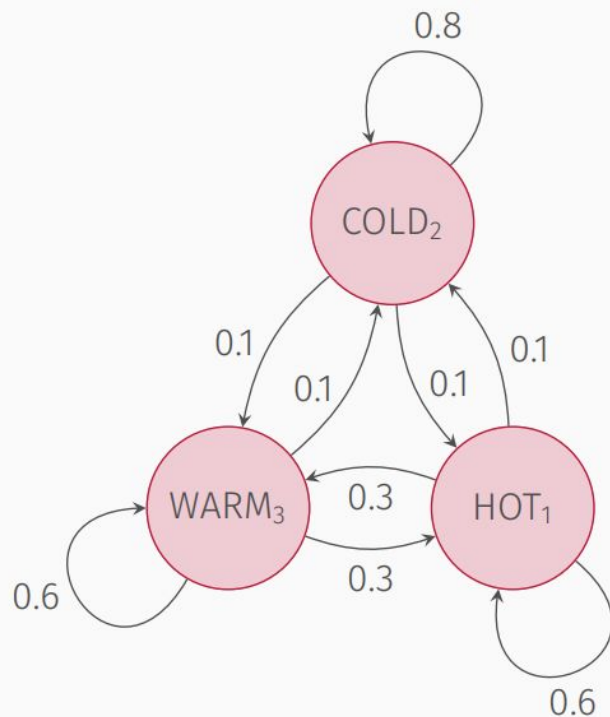
Words	BIO Label
Jane	B-PER
Villanueva	I-PER
of	O
United	B-ORG
Airlines	I-ORG
Holding	I-ORG
discussed	O
the	O
Chicago	B-LOC
route	O
.	O



# Hidden Markov Models (HMMs)

---

# Markov Chains Tell Us about the Probabilities of Sequences of Random Variables



The figure to the left represents a Markov Chain.

- States
- Transitions
- Weights (probabilities)

The probability of COLD<sub>2</sub> at the timestep after COLD<sub>2</sub> is 0.8.

The probability of HOT<sub>1</sub> after COLD<sub>2</sub> is 0.1.

The probability of HOT<sub>1</sub> → WARM<sub>3</sub> → COLD<sub>2</sub> is  $0.3 \times 0.1 = 0.03$

**The Markov Assumption applies.**

# The Markov Assumption

“When predicting the future, the past doesn’t matter—only the present.”

in other words

$$p(q_i = a | q_1 \dots q_{i-1}) = P(q_i = a | q_{i-1})$$

This is the same assumption we made for ngram language modeling.

# A formal definition of the Hidden Markov Model (HMM)

- $Q = q_1, \dots, q_N$  a set of  $N$  **states**
- $A = a_{1,1}, a_{1,2}, \dots$  a **transitional probability matrix** of cells  $a_{ij}$ , where each cell is a probability of moving from state  $i$  to state  $j$ .  
 $\sum_{j=1}^N a_{ij} = 1 \forall i$
- $O = o_1, \dots, o_T$  a **sequence of  $T$  observations**, each drawn from a vocabulary  $V$ .
- $B = b_1, \dots, b_n$  a sequence of observation likelihoods (or **emission probabilities**). The probability that observation  $o_t$  is generated by state  $q_i$ .
- $\pi = \pi_1, \dots, \pi_N$  an **initial probability distribution** over states (the probability that the Markov chain will start in state  $q_i$ . Some states  $q_j$  may have  $p_j = 0$  (meaning they cannot be initial states).  $\sum_{i=1}^N \pi_i = 1 \forall i$

# HMMs Assume the Markov Assumption and Output Independence

Like Markov Chains, HMMs require the Markov Assumption:

$$p(q_i|q_1...q_{i-1}) = P(q_i|q_{i-1})$$

The further assume that the observed outputs depend only upon the state  
(**Output Independence**)

$$P(o_i|q_1, \dots, q_i, \dots, q_T, o_1 \dots, o_i, \dots o_T) = P(o_i|q_i)$$

Where  $q_1, \dots, q_T$  are the states at each time step and  $o_1, \dots, o_T$  are the outputs at each time step. In other words:

- The preceding or following states do not matter (we assume)
- The preceding or following outputs do not matter (we assume)

# We can use Bayes' Rule to pick the right hidden POS tags

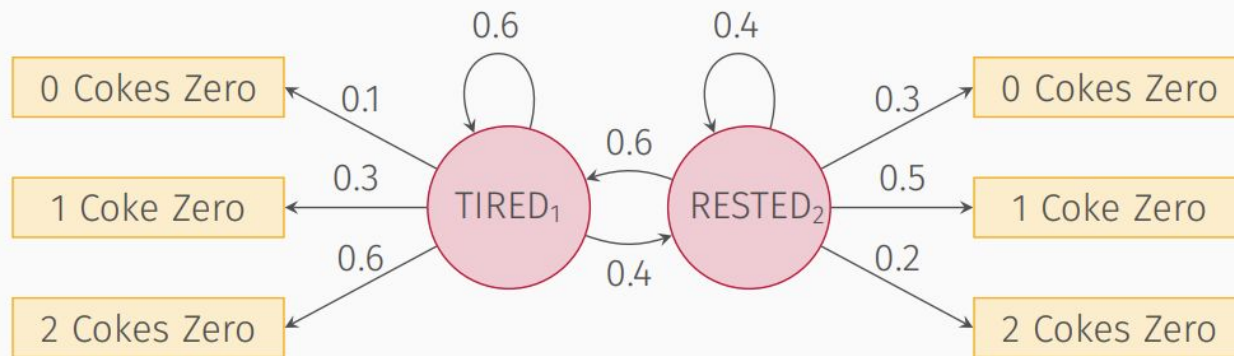
$$\hat{t}_1^n = \operatorname{argmax}_{t_1^n} P(t_1^n | w_1^n) \approx \operatorname{argmax}_{t_1^n} \prod_{i=1}^n P(w_i | t_i) P(t_i | t_{i-1})$$

For timestep 1 through  $n$ :

- $t_1$ : the hidden state at timestep 1
- $w_1$ : the observed word at timestep 1

# The Coke Zero Example

Since I do not drink coffee, I must drink Coke Zero to remain caffeinated. My consumption is related to my exhaustion. Could you build a model to infer my exhaustion from the number of Coke Zero bottles added to my wastebasket each day?



$$\pi = [0.7, 0.3]$$

# Training an HMM

How do we learn the transition and emission probabilities?

- If we have (enough) data labeled with hidden and observed events, can just **use MLE/relative frequencies** with or without smoothing
- If we don't have (enough) labeled data, can use the **Forward-Backward Algorithm**, a special case of the Expectation Maximization (EM) algorithm
  - We won't go into the details of this algorithm, but the overview is that you start with an initial estimate and use that estimate to compute a better one iteratively



# Training HMMs with labeled data

Suppose we knew both the sequence of days in which a grad student is tired or rested and the number of Cokes Zero that she consumes each day:

0	3	1
rested	tired	rested
1	2	2
tired	tired	tired
0	0	2
rested	rested	rested

How would you train an HMM?

# Using MLE to train HMMs

First, compute  $\pi$  from the initial states:

$$\pi_t = 1/3 \quad \pi_r = 2/3$$

Then we can compute the matrix  $A$ :

$$\begin{aligned} p(\text{tired}|\text{tired}) &= 1/2 & p(\text{tired}|\text{rested}) &= 1/6 \\ p(\text{rested}|\text{tired}) &= 1/3 & p(\text{rested}|\text{rested}) &= 2/3 \end{aligned}$$

and then the matrix  $B$ :

$$\begin{aligned} p(0|\text{tired}) &= 0 & p(0|\text{rested}) &= 2/5 \\ p(1|\text{tired}) &= 1/4 & p(1|\text{rested}) &= 1/5 \\ p(2|\text{tired}) &= 1/2 & p(2|\text{rested}) &= 1/5 \end{aligned}$$

# Parameters of an HMM for POS

$A =$

	N	V	O
N	0.1	0.6	0.3
V	0.3	0.3	0.4
O	0.3	0.4	0.3

transition probabilities ←

emission probabilities ↓

$B =$

	I	m	gonna	make	him	an	offer	he	can	t	refuse
N	0.1	0.00001	0.00001	0.2	0.1	0.00001	0.2	0.1	0.1	0.00001	0.19996
V	0.00001	0.1	0.2	0.2	0.00001	0.00001	0.05	0.00001	0.19995	0.00001	0.25
O	0.00001	0.00001	0.00001	0.00001	0.00001	0.5	0.00001	0.00001	0.00001	0.49991	0.00001

# Wrapping up

- Parts of speech are grammatical classes of words like nouns, verbs, and adjectives
- Part of speech (POS) tagging assigns a part of speech to every input word in context
- Named entity recognition (NER) is the task of identifying named entities like people, locations, and organizations
- NER can be framed as a sequence labeling task with a BIO framework
- HMMs can be used for sequence labeling tasks like POS tagging and NER
- Key parameters of HMMs are transition and emission probabilities

*Questions?*