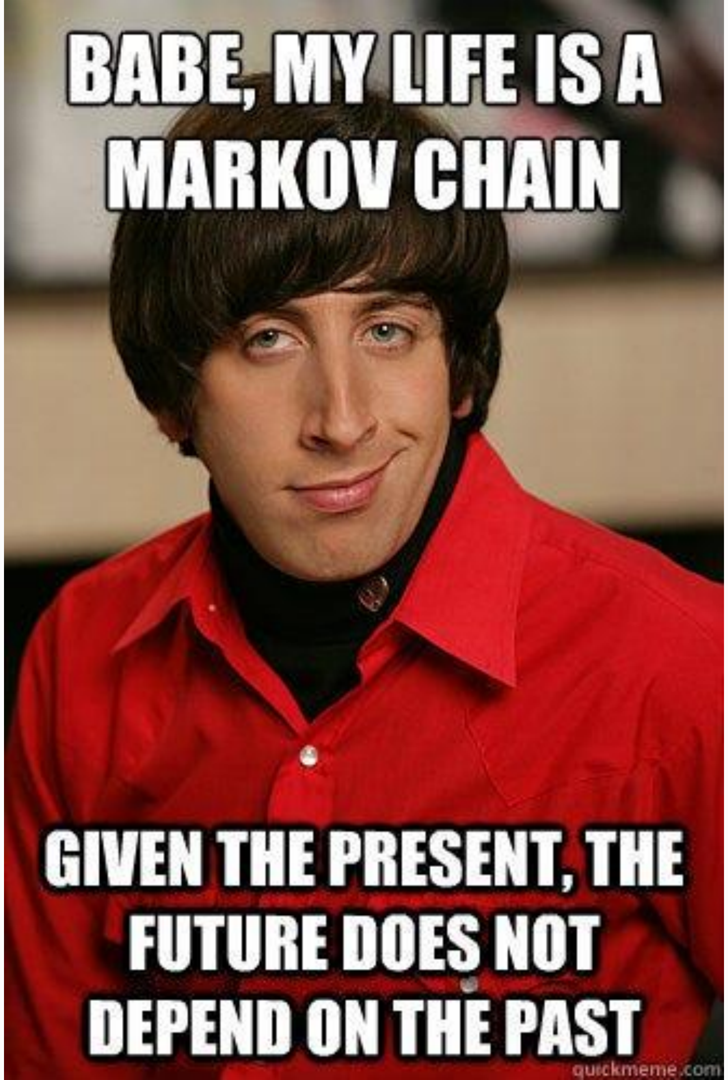


Source:
https://towardsdatascience.com/nlg-for-fun-automated-headlines-generator-6d0459f9588f

CS 2731 Introduction to Natural Language Processing

Session 18: POS tagging, NER, HMMs part 1

Michael Miller Yoder

March 18, 2024

University of Pittsburgh | School of Computing and Information

Course logistics: homeworks

- Homework 3 grades released this week or the next
- Homework 4 is **due Mon Mar 25**
  - Part 1: Do part-of-speech tagging manually with the Viterbi algorithm
  - Part 2: Fine-tune BERT-based models for part-of-speech tagging in English and Norwegian
    - Copy and fill in a skeleton Colab notebook

Course logistics

- Next project milestones
  - Peer review **due Wed Mar 27**
    - Will be released this Wed Mar 20
    - Form where you will review your own and your teammates' contributions so far
    - Will not be used for grading, just for addressing any issues
  - Basic working system **due Thu Apr 4**

Core tasks and applications of NLP

CORE TASKS

- text classification
- representation learning

MODULE 2

- language models
- conditional language models

MODULE 3

- sequence labeling

MODULE 4

- syntactic parsing

APPLICATIONS

- machine translation
- speech recognition & synthesis
- chatbots
- computational social science

Overview: POS tagging, NER, HMMs part 1

- 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.

Slide credit: *David Mortensen*

My cat who lives dangerously no longer has nine lives.

Slide credit: David Mortensen

My cat who lives dangerously no longer has nine lives.

lives	/lɪvz/	verb
lives	/lajvz/	noun

Slide credit: David Mortensen

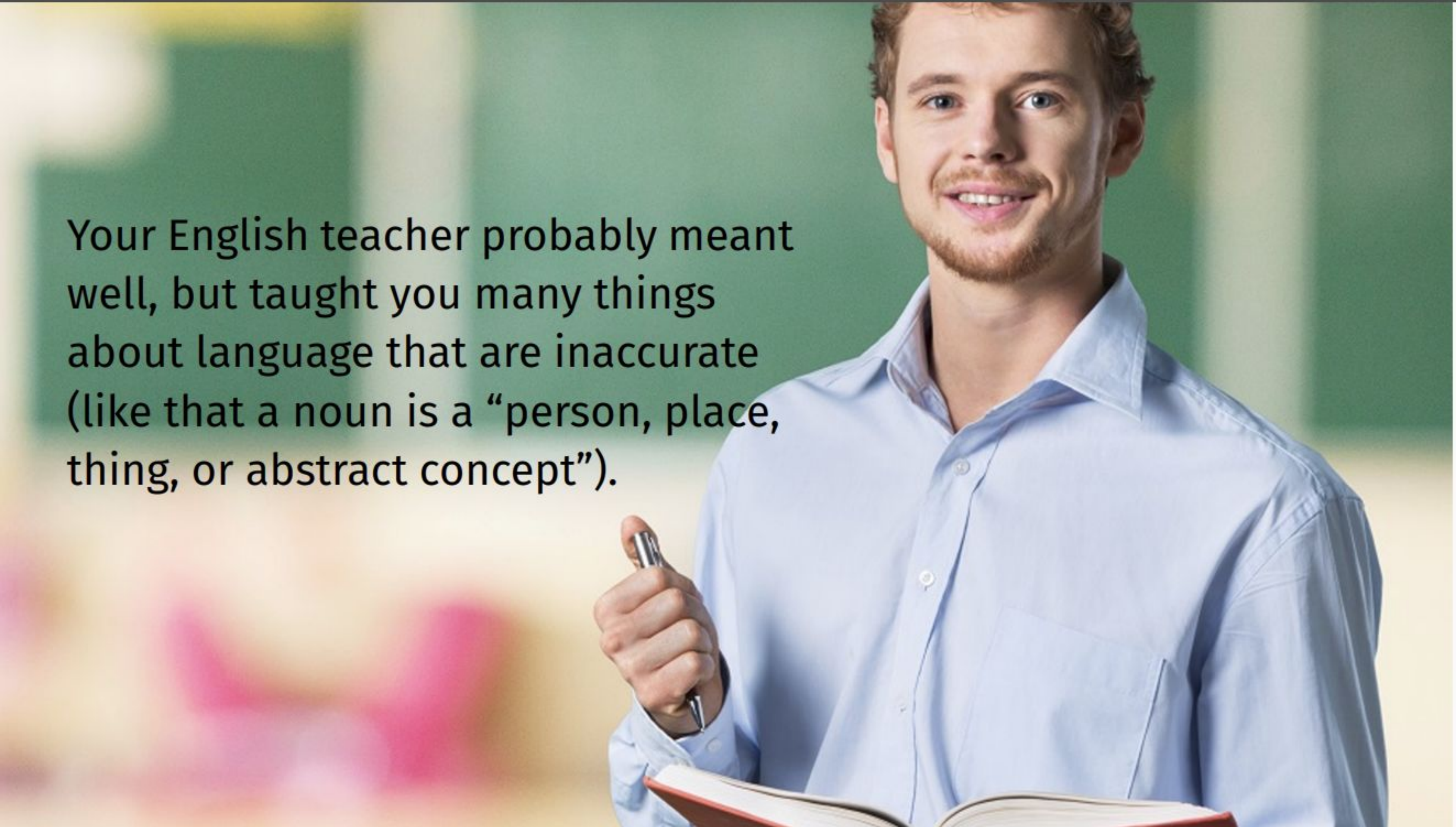

STOP
LETTING YOUR
DOG SHIT
In This Grass
Keep Your Dog Shit at HOME!

Examples of Parts of Speech

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

Slide credit: David Mortensen

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 the 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"

Slide credit: David Mortensen

Open Class Parts of Speech

Classes to which neologisms are readily added. In English:

nouns	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
verbs	can take noun phrases as arguments and tense morphology and can be modified by adverbs
adjectives	can modify nouns and take comparative and superlative morphology where allowed by prosody
adverbs	can modify verbs, adjectives, or other adverbs

Slide credit: David Mortensen

Closed Class Parts of Speech

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

prepositions	occur before noun phrases, connecting them syntactically to larger phrases
determiners	occur at the beginning of noun phrases
conjunction	join phrases, clauses, and sentences
auxiliary verbs	occur before (non-finite) main verbs
particles	are associated with a verb and are "moveable" (e.g. *He tore **off** his shirt* versus *He tore his shirt **off***
numerals	are distributed in some ways like nouns and in others like adjectives

Slide credit: David Mortensen

What about pronouns?

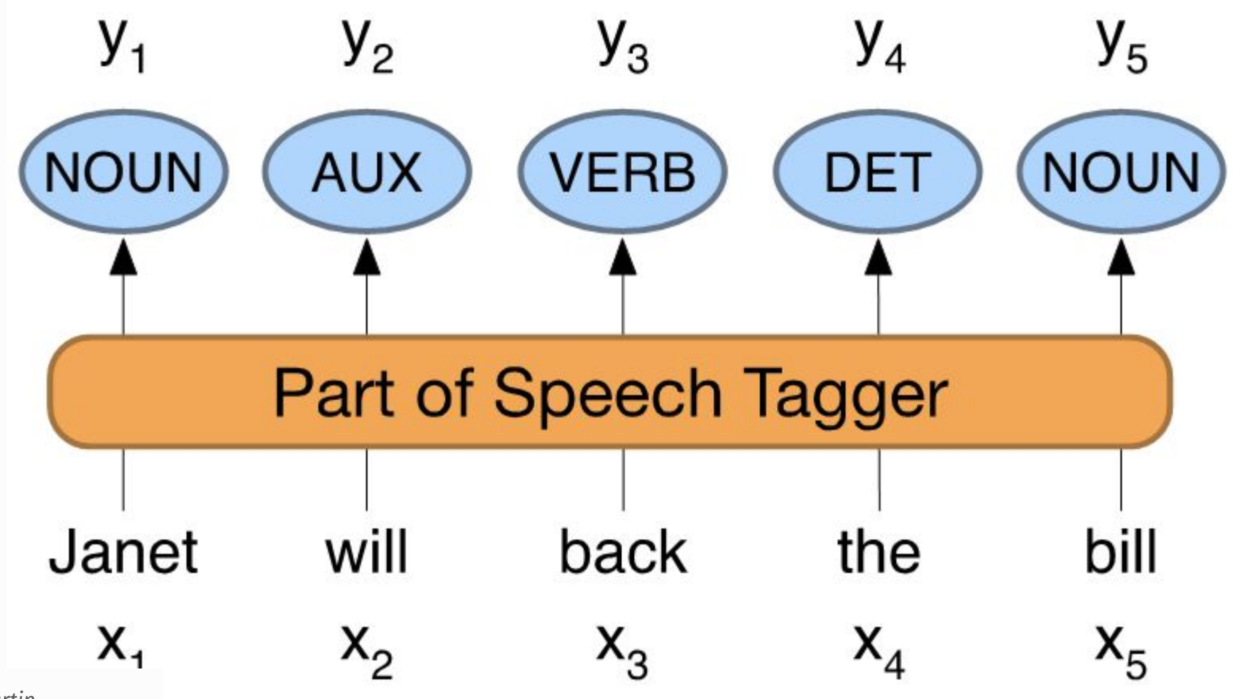
- Pronouns are generally considered, in English, to be a closed class—it is not easy to add new items to it.
- What are we to make of **neopronouns** like *xe* and *xem* or *ze* and *hir*?
- Their existence suggests that pronouns are not a completely closed class
  - Social movements can change grammar!
  - 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

Slide adapted from David Mortensen

Part of speech (POS) tagging

Part-of-speech tagging

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



Slide adapted from Jurafsky & Martin

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

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

Slide adapted from Jurafsky & Martin

Penn TreeBank tagset for English

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*			

Why part of speech tagging?

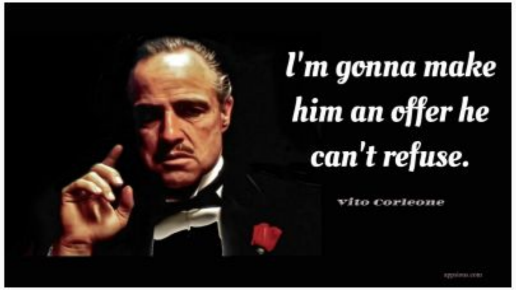
- Can be useful for other NLP tasks
  - Parsing: POS tagging can improve syntactic parsing
  - MT: reordering of adjectives and nouns (say from Spanish to English)
  - Sentiment or affective tasks: may want to distinguish adjectives or other POS
  - Text-to-speech (how do we pronounce “lead” or "object"?)
- 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

Slide adapted from Jurafsky & Martin

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**.

Slide adapted from David Mortensen

How difficult is POS tagging in English?

Roughly 15% of word types are ambiguous

- Hence 85% of word types are unambiguous
- *Janet* is always PROPN, *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

Slide adapted from Jurafsky & Martin

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 → PROPN

Slide adapted from Jurafsky & Martin

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

Slide adapted from Jurafsky & Martin

Named entity recognition (NER)

Named entities

- **Named entity** 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

Slide adapted from Jurafsky & Martin

Named entity tagging

The task of named entity recognition (NER):

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

Slide adapted from Jurafsky & Martin

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**].

Slide adapted from Jurafsky & Martin

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.

Slide adapted from Jurafsky & Martin

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

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

Slide adapted from Jurafsky & Martin

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] discussed 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

Slide adapted from Jurafsky & Martin

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,

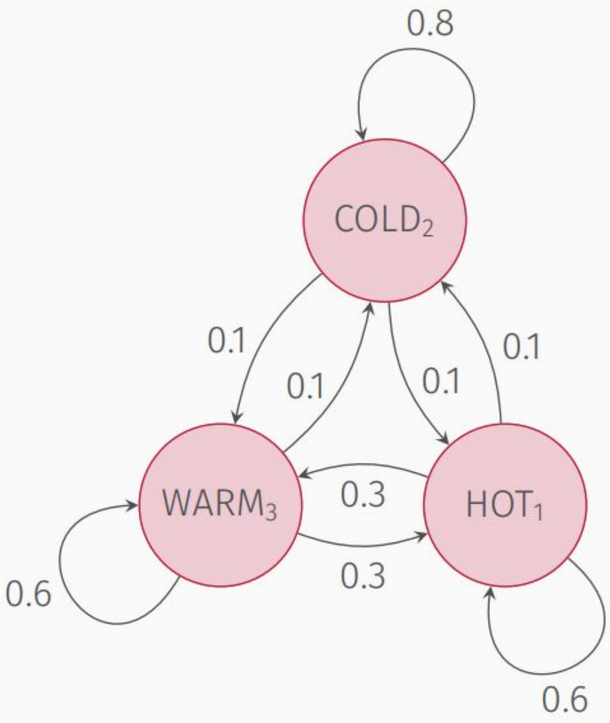
n I tags

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_2 at the timestep after COLD_2 is 0.8.
The probability of HOT_1 after COLD_2 is 0.1.
The probability of $\text{HOT}_1 \rightarrow \text{WARM}_3 \rightarrow \text{COLD}_2$ is $0.3 \times 0.1 = 0.03$

The Markov Assumption applies.

Slide credit: David Mortensen

The Markov Assumption

"When predicting the future, the past doesn't matter—only the present."

in other words

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

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

Slide credit: *David Mortensen*

In Hidden Markov Models, Markov Chains are Hidden

The basic idea of a Hidden Markov Model (or HMM) is like that of a Markov Chain, except that the states are never observed.

- Observations are “emitted” from the hidden states
- So, when seen as a graph, HMMs have two kinds of nodes and two kinds of edges
    - Hidden states and observations
    - Transitional probabilities and emission probabilities
- The hidden states and transitional probabilities represent the latent structure that “sits behind” the observed phenomena

Slide credit: David Mortensen

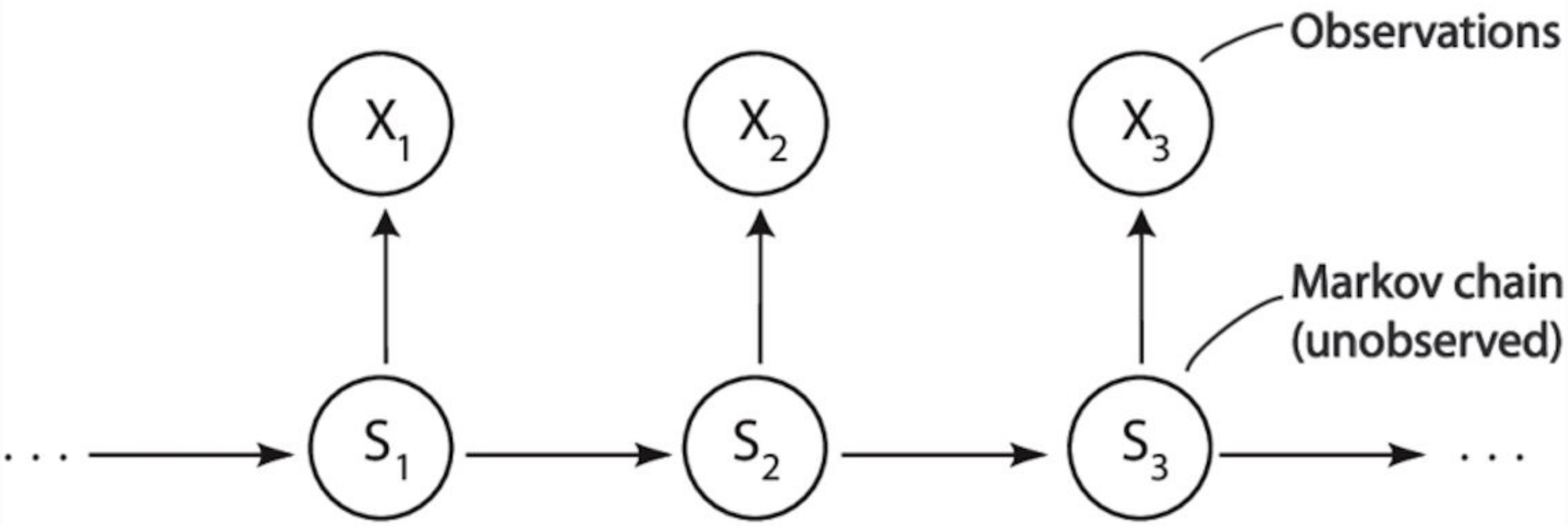


Figure credit: Amy Van Scoyoc
https://dlab.berkeley.edu/news/brief-primer-hidden-markov-models

A formal definition of the Hidden Markov Model (HMM)

$Q = q_1, \ldots, q_N$	a set of N states
$A = a_{1,1}, a_{1,2}, \ldots$	a transitional probability matrix of cells a_{ij}, where each cell is a probability of moving from state i to state j. $\sum_{j=1}^{N} a_{ij} = 1\ \forall i$
$O = o_1, \ldots, o_T$	a sequence of T observations, each drawn from a vocabulary V.
$B = b_1, \ldots, b_n$	a sequence of observation likelihoods (or emission probabilities). The probability that observation o_t is generated by state q_i.
$\pi = \pi_1, \ldots, \pi_N$	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). $\sum_{i=1}^{N} \pi_i = 1\ \forall i$

Slide credit: David Mortensen

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**)

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

Where $q_1, \ldots, q_T$ are the states at each time step and $o_1, \ldots, o_T$ 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)

Slide credit: David Mortensen

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

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

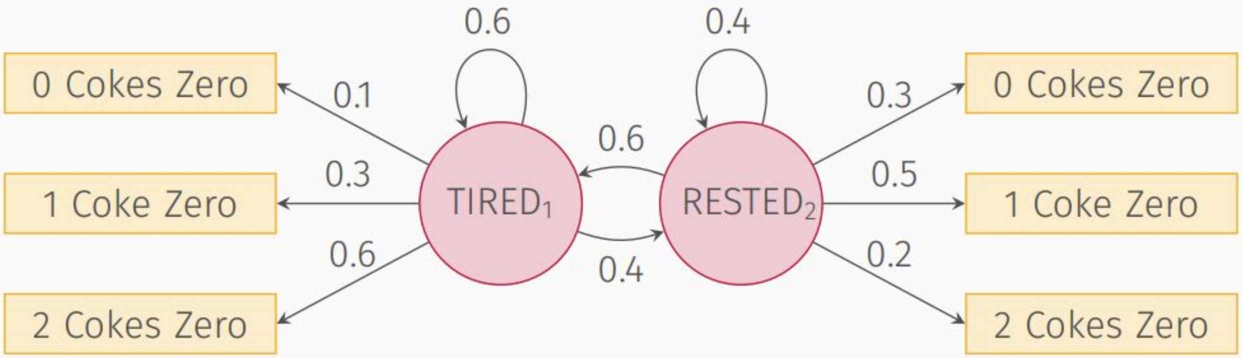
For timestep 1 through *n*:

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

Slide adapted from Diane Litman, Jurafsky & Martin

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?



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

Slide credit: David Mortensen

Conclusion

- 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?