Outline of POS/HMM unit:

- The POS-tagging problem
  - The idea of parts of speech or word classes
  - Our set of POS categories
  - Formal definition of the problem
- Hidden Markov Models definition
  - Informal explanation of HMMs
  - Formal definition
  - Statement of the three (or four) HMM problems
- HMM Problem 1 and the forward algorithm
- HMM Problem 2 and the Viterbi algorithm, applied to POS-tagging
- HMM Problem 3 and the Baum-Welch algorithm, with other linguistic applications

English parts of speech:

| | | | |
|---|---|---|---|
| Noun | Adjective | Pronoun | Conjunction |
| Verb | Adverb | Preposition | Interjection |

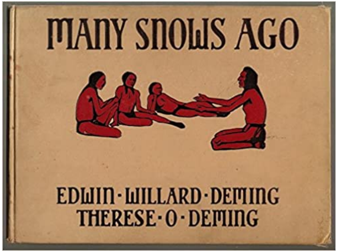
Ancient Indian (Sanskrit) parts of speech:

| | | | |
|---|---|---|---|
| Noun | Verb | Preverb | Particle |

Ancient Greek parts of speech:

| | | | |
|---|---|---|---|
| Noun | Participle | Pronoun | Conjunction |
| Verb | Adverb | Preposition | Article |

Many languages, including English, divide common nouns into count nouns and mass nouns. Count nouns can occur in the singular and plural and can be counted. Mass nouns are used when something is conceptualized as a homogenous group. So *snow*, *salt*, and *communism* are not counted (i.e., ∗*two snows* or ∗*two communisms*). J&M §17.1




*Karl Marx and Josef Stalin represent two very different communisms.*

# Nouns and adjectives

Nouns can be used attributively:

> *There is of course nothing new in putting a noun to this use when no convenient adjective is available; examples abound in everyday speech—***government department**, **nursery school**, **television set**, **test match**, *and innumerable others. But the noun-adjective, useful in its proper place, is now running riot and corrupting the language.* *H.W. Fowler, Modern English Usage*

Adjectives can be used substantively:

> *Do not let the* **perfect** *be the enemy of the* **good**.

# Word classes

*Linguists group the words of a language into classes (sets) which show similar syntactic behavior, and often a typical semantic type. These word classes are otherwise called* **syntactic** *or* **grammatical categories***, but more commonly still by the traditional name* **parts of speech** *(POS). Three important parts of speech are* **noun***,* **verb***, and* **adjective***. ... The most basic test for words belonging to the same class is the* **substitution test***. Adjectives can be picked out as words that occur in the frame:*

$$\textit{The} \left\{ \begin{array}{c} \textit{sad} \\ \textit{intelligent} \\ \textit{green} \\ \textit{fat} \\ \ldots \end{array} \right\} \textit{one is in the corner}$$

*Manning and Schütze, Foundations of Statistical NLP, pg 81*

# Computed word classes

| |
|---|
| Friday Monday Thursday Wednesday Tuesday Saturday Sunday weekends Sundays Saturdays |
| June March July April January December October November September August |
| people guys folks fellows CEOs chaps doubters commies unfortunates blokes |
| down backwards ashore sideways southward northward overboard aloft downwards adrift |
| water gas coal liquid acid sand carbon steam shale iron |
| great big vast sudden mere sheer gigantic lifelong scant colossal |
| man woman boy girl lawyer doctor guy farmer teacher citizen |
| American Indian European Japanese German African Catholic Israeli Italian Arab |
| pressure temperature permeability density porosity stress velocity viscosity gravity tension |
| mother wife father son husband brother daughter sister boss uncle |
| machine device controller processor CPU printer spindle subsystem compiler plotter |
| John George James Bob Robert Paul William Jim David Mike |
| anyone someone anybody somebody |
| feet miles pounds degrees inches barrels tons acres meters bytes |
| director chief professor commissioner commander treasurer founder superintendent dean custodian |
| liberal conservative parliamentary royal progressive Tory provisional separatist federalist PQ |
| had hadn't hath would've could've should've must've might've |
| asking telling wondering instructing informing kidding reminding bothering thanking deposing |
| that tha theat |
| head body hands eyes voice arm seat eye hair mouth |

**Table 2**
Classes from a 260,741-word vocabulary.

Brown et al, "Class-Based *n*-gram Models of Natural Language"

# Computed word classes

little prima moment's trifle tad Litle minute's tinker's hornet's teammate's
6
ask remind instruct urge interrupt invite congratulate commend warn applaud
object apologize apologise avow whish
cost expense risk profitability deferral earmarks capstone cardinality mintage reseller
B dept. AA Whitey CL pi Namerow PA Mgr. LaRose
# Rel rel. #S Shree
S Gens nai Matsuzawa ow Kageyama Nishida Sumit Zollner Mallik
research training education science advertising arts medicine machinery Art AIDS
rise focus depend rely concentrate dwell capitalize embark intrude typewriting
Minister mover Sydneys Minster Miniter
3
running moving playing setting holding carrying passing cutting driving fighting
court judge jury slam Edelstein magistrate marshal Abella Scalia larceny
annual regular monthly daily weekly quarterly periodic Good yearly convertible
aware unaware unsure cognizant apprised mindful partakers
force ethic stoppage force's conditioner stoppages conditioners waybill forwarder Atonabee
systems magnetics loggers products' coupler Econ databanks Centre inscriber correctors
industry producers makers fishery Arabia growers addiction medalist inhalation addict
brought moved opened picked caught tied gathered cleared hung lifted

**Table 3**
Randomly selected word classes.

Brown et al, "Class-Based *n*-gram Models of Natural Language"

| Universal | | Penn Treebank | | |
|---|---|---|---|---|
| ADJ | Adjective | JJ | Adjective | *yellow* |
| | | JJR | Comparative adjective | *bigger* |
| | | JJS | Superlative adjective | *wildest* |
| ADP | Adposition | IN | Preposition | *of, in , by* |
| | | RP | Particle | *up, off* |
| ADV | Adverb | RB | Adverb | *quickly* |
| | | RBR | Comparative adverb | *faster* |
| | | RBS | Superlative adverb | *fastest* |
| | | WRB | Wh-adverb | *how, where* |
| CONJ | Conjunction | CC | Coordinating conjunction | *and, but, or* |

| | Universal | | Penn Treebank | |
|---|---|---|---|---|
| DET | Determiner, article | DT | Determiner | *a, the* |
| | | PDT | Predeterminer | *all, both* |
| | | PRP$ | Posessive pronoun | *your, one's* |
| | | WDT | Wh-determiner | *which, that* |
| | | WP$ | Wh-possessive | *whose* |
| NOUN | Noun | NN | Singular or mass noun | *llama* |
| | | NNP | Proper noun, singular | *IBM* |
| | | NNPS | Noun, plural | *llamas* |
| NUM | Numeral | CD | Cardinal number | *one, two* |
| PRT | Particle | POS | Possessive ending | *'s* |
| | | TO | "to" [Infinitive marker] | *to* |
| PRON | Pronoun | EX | Existential "there" | *there* |
| | | PRP | Personal pronoun | *I, you, he* |
| | | WP | Wh-pronoun | *what, who* |

| | Universal | | Penn Treebank | |
|---|---|---|---|---|
| VERB | Verb | MD | Modal *can, should* | |
| | | VB | Verb base | *eat* |
| | | VBD | Verb past tense | *ate* |
| | | VBG | Verb gerund | *eating* |
| | | VBN | Verb past participle | *eaten* |
| | | VBP | Verb non-3sp | *eat* |
| | | VBZ | Verb 3sp | *eats* |
| . | Puntuation mark | (none) | | |
| X | Other | FW | Foreign word | *mea culpa* |
| | | LS | List item marker | *1, 2, One* |
| | | SYM | Symbol | *+, %, &* |
| | | UH | Interjection | *ah, oops* |

*Suppose we want to determine the average annual temperature at a particular location on earth over a series of years.*

*To simplify the problem, we consider only two annual temparatures, "hot" and "cold." Suppose that evidence indicates that the probability of a hot year followed by another hot year is 0.7 and the probability that a cold year is followed by another cold year is 0.6.*

*Also suppose that research indicates a correlation between the size of tree growth rings and temparature. For simplicity, we consider only three different tree ring sizes: small, medium, and large. Finally suppose hot years are more likely to result in large tree rings, cold years in small.*

| | H | C | S | M | L |
|---|---|---|---|---|---|
| H | 0.7 | 0.3 | 0.1 | 0.4 | 0.5 |
| C | 0.4 | 0.6 | 0.7 | 0.2 | 0.1 |

*Mark Stamp, "A Revealing Introduction to Hidden Markov Models". Abridged.*

Let $Q$ be a set of $N$ states types. Use $i, j, ii, jj \in [0, N)$ to index into $Q$.
Let $V$ be a set of $M$ symbols types. Use $k \in [0, M)$ to index into $V$.

Let $\bar{S}$ be a sequence of $T$ state tokens and $\bar{\mathcal{O}}$ be a sequence of $T$ observation tokens.
Use $t \in [0, T)$ to index into $\bar{\mathcal{O}}$ and $\bar{S}$

Thus $\bar{\mathcal{O}} = \langle \mathcal{O}_0, \mathcal{O}_1, \ldots \mathcal{O}_{T-1} \rangle$ is a sequence of observation tokens, e.g., $\mathcal{O}_t = v_k$,
and $\bar{S} = \langle S_0, S_1, \ldots S_{T-1} \rangle$ is a sequence of state tokens, e.g., $S_t = q_j$.

A **hidden Markov model** is a triple $\lambda = (A, B, \boldsymbol{\pi})$ where

- $A$ is an $N \times N$ matrix of state transition probabilities: $a_{ij} = P(S_{t+1} = q_j \mid S_t = q_i)$
- $B$ is an $N \times M$ matrix of emission (or observation) probabilities: $b_j(k) = P(\mathcal{O}_t = v_k \mid S_t = q_j)$
- $\boldsymbol{\pi}$ is the initial state distribution. $\pi_i = P(S_0 = q_i)$

Four HMM problems:

Problem 0. Given $\bar{\mathcal{O}}$ together with $\bar{S}$, compute $\lambda = (A, B, \boldsymbol{\pi})$ most likely to have produced those sequences.
[Solution: MLE, possibly with smoothing.]

Problem 1. Given $\lambda = (A, B, \boldsymbol{\pi})$ and $\bar{\mathcal{O}}$, compute the probability that $\lambda$ assigns to $\bar{\mathcal{O}}$.
[Solution: The forward algorithm.]

Problem 2. Given $\lambda = (A, B, \boldsymbol{\pi})$ and $\bar{\mathcal{O}}$, find $\bar{S}$ that maximizes the probability that $\lambda$ assigns to $\bar{\mathcal{O}}$.
[Solution: The Viterbi algorithm.]

Problem 3. Given $\bar{\mathcal{O}}$, $M$ (or $V$), and $N$, find $\lambda = (A, B, \boldsymbol{\pi})$ that maximizes the likelihood of $\bar{\mathcal{O}}$.
[Solution: The Baum-Welch algorithm, a version of EM.]

Coming up:

- Language model programming assignment (Fri, Sept 26)
- Reading from J&M, Sections 8.(0–4) (Wed, Sept 24) At least through Section 1, by the due date; you may spread out the rest.
- POS quiz (Thurs, Sept 5)
- HMMs quiz (Tues, Sept 30)
- Swift reading (Wed, Oct 1)
- HMM/POS programming assignment (Wed, Oct 8)