# POS Tagging

context

| | | |
|---|---|---|
| everyone likes | ______________ | |
| a bottle of | ______________ | is on the table |
| | ______________ | makes you drunk |
| a cocktail with | ______________ | and seltzer |

from last time

# Distribution

- Words that appear in similar contexts have similar representations (and similar meanings, by the distributional hypothesis).

# Parts of speech

- Parts of speech are categories of words defined **distributionally** by the morphological and syntactic contexts a word appears in.

# Morphological distribution

- POS often defined by distributional properties; verbs = the class of words that each combine with the same set of affixes

| | -s | -ed | -ing |
|---|---|---|---|
| walk | walks | walked | walking |
| slice | slices | sliced | slicing |
| believe | believes | believed | believing |

| | -s | -ed | -ing |
|---|---|---|---|
| walk | walks | walked | walking |
| sleep | sleeps | slept | sleeping |
| eat | eats | ate | eating |
| give | gives | gave | giving |

Bender 2013

# Syntactic distribution

- Substitution test: if a word is replaced by another word, does the sentence remain grammatical?

| Kim saw the | elephant | before we did |
|---|---|---|
| | dog | |
| | idea | |
| | *of | |
| | *goes | |

# Syntactic distribution

- These can often be too strict; some contexts admit substitutability for some pairs but not others.

| | | |
|---|---|---|
| Kim saw the | elephant | before we did |
| | *Sandy | |

both nouns but
common vs. proper

| | | |
|---|---|---|
| Kim *arrived the | elephant | before we did |

both verbs but
transitive vs. intransitive

Bender 2013

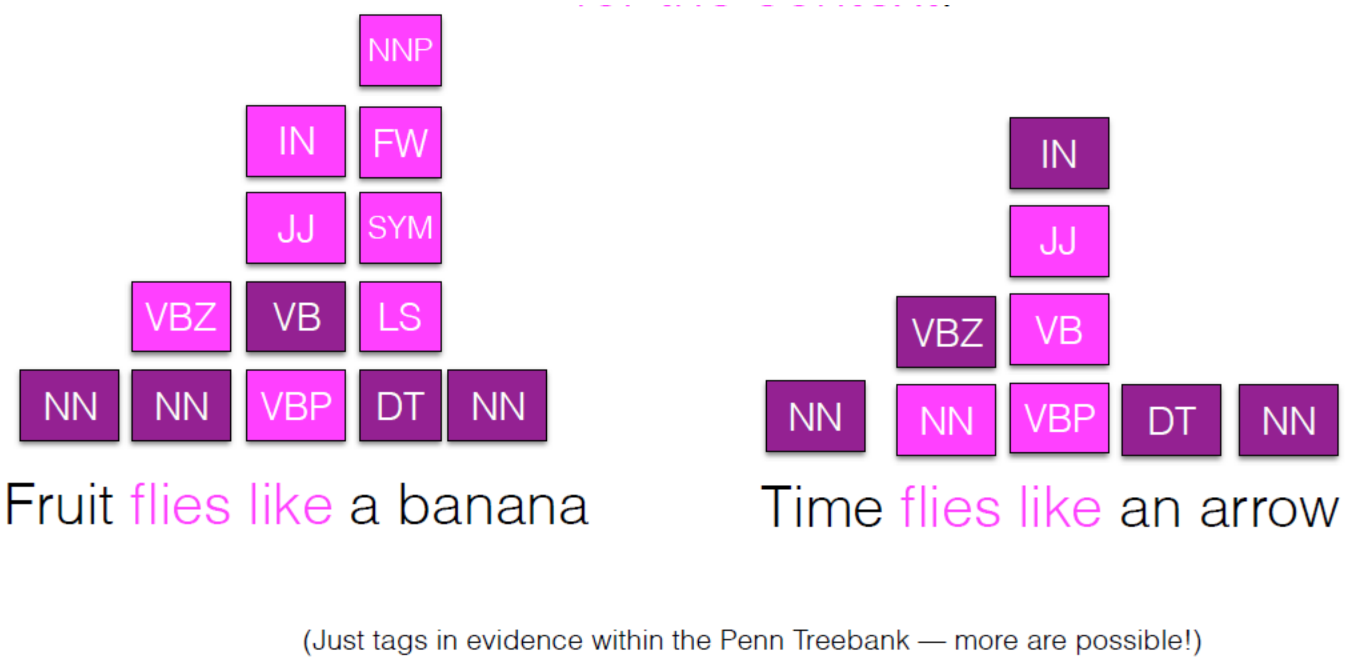
| | |
|---|---|
| Nouns | People, places, things, actions-made-nouns ("I like swimming"). Inflected for singular/plural |
| Verbs | Actions, processes. Inflected for tense, aspect, number, person |
| Adjectives | Properties, qualities. Usually modify nouns |
| Adverbs | Qualify the manner of verbs ("She ran downhill extremely quickly yesteray") |
| Determiner | Mark the beginning of a noun phrase ("a dog") |
| Pronouns | Refer to a noun phrase (he, she, it) |
| Prepositions | Indicate spatial/temporal relationships (on the table) |
| Conjunctions | Conjoin two phrases, clauses, sentences (and, or) |

| | | |
|---|---|---|
| Open class | Nouns | fax, affluenza, subtweet, bitcoin, cronut, emoji, listicle, mocktail, selfie, skort |
| | Verbs | text, chillax, manspreading, photobomb, unfollow, google |
| | Adjectives | crunk, amazeballs, post-truth, woke |
| | Adverbs | hella, wicked |
| Closed class | Determiner | |
| | Pronouns | |
| | Prepositions | English has a new preposition, because internet [Garber 2013; Pullum 2014] |
| | Conjunctions | |

OOV? Guess Noun

# POS tagging

- Words often have more than one PO
    - *The **back** door* (adjective)
    - *On my **back*** (noun)
    - *Win the voters **back*** (particle)
    - *Promised to **back** the bill* (verb)




- The POS tagging task: Determine the POS tag for all tokens in a sentence.
- Due to ambiguity (and unknown words), we cannot rely on a dictionary to look up the correct POS tags.

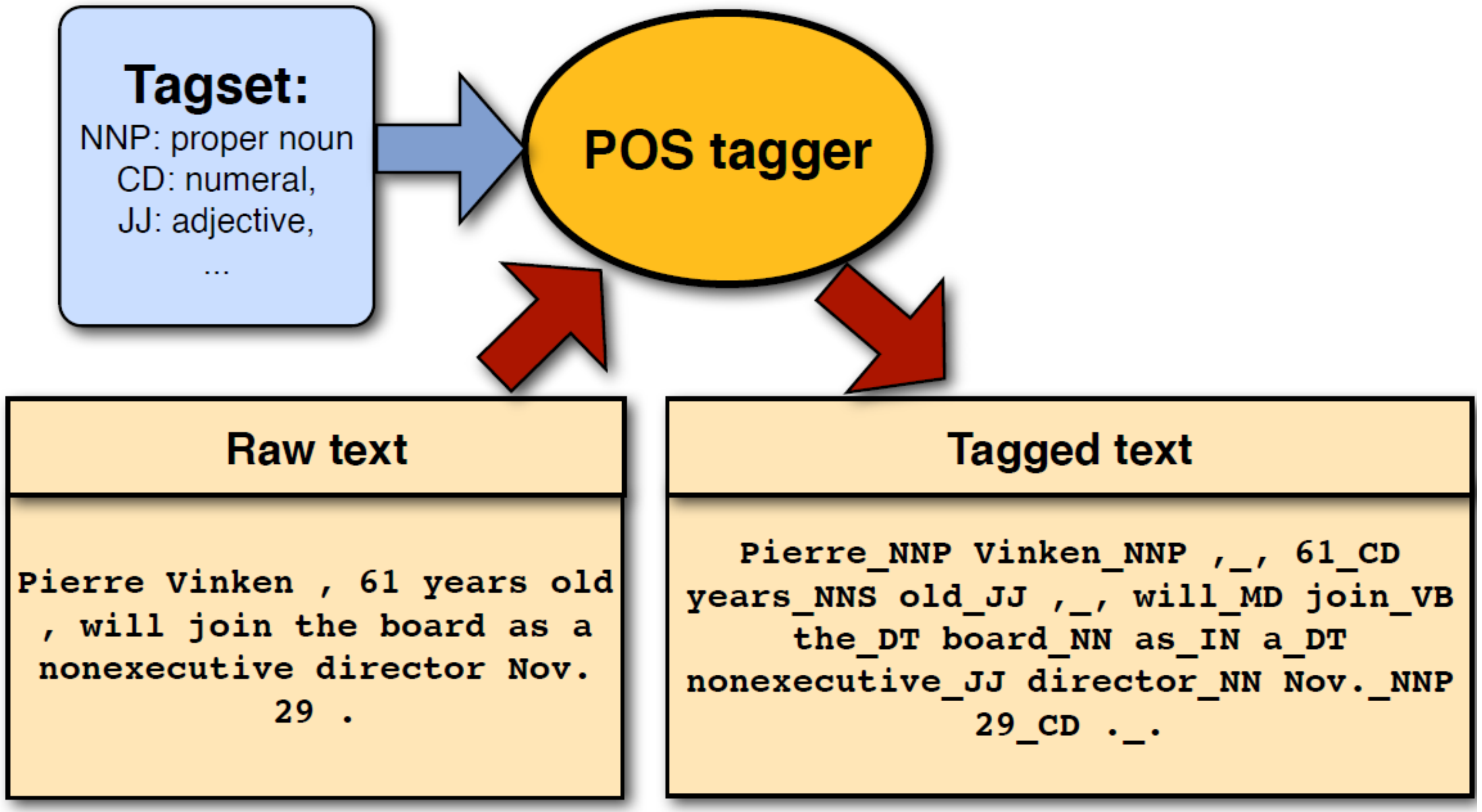
# How Much Ambiguity is There?

- Most **word *types*** appear with only one POS tag….
  - Brown corpus with 87-tag set: 3.3% of word types are ambiguous,
  - Brown corpus with 45-tag set: 18.5% of word types are ambiguous
- … but a large fraction of **word *tokens*** are ambiguous Original Brown corpus: 40% of tokens are ambiguous

# Creating a POS Tagger

- To handle ambiguity and coverage, POS taggers rely on learned models.
- For a **new language** (or domain)
  - Step 0: Define a POS tag set
  - Step 1: Annotate a corpus with these tags
- For a **well-studied language** (and domain):
  - Step 1: Obtain a POS-tagged corpus
- For any language….:
  - Step 2: Choose a POS tagging model (e.g. an HMM)
  - Step 3: Train your model on your training corpus
  - Step 4: Evaluate your model on your test corpus

# POS Tagging


Tagset:
NNP: proper noun
CD: numeral,
JJ: adjective,
...
POS tagger
Raw text
Pierre Vinken , 61 years old , will join the board as a nonexecutive director Nov. 29 .
Tagged text
Pierre_NNP Vinken_NNP ,_, 61_CD years_NNS old_JJ ,_, will_MD join_VB the_DT board_NN as_IN a_DT nonexecutive_JJ director_NN Nov._NNP 29_CD ._.

# Evaluation Metric: Test Accuracy

- **How many *words* in the unseen test data can you tag correctly?**
  - State of the art on Penn Treebank: around 97%
  - ⇨ **How many *sentences* can you tag correctly?**
- Compare your model against a **baseline**
  - Standard: assign to each word its most likely tag
  - (use training corpus to estimate P(t|w) )
  - Baseline performance on Penn Treebank: around 93.7%
- … and a **(human) ceiling**
  - How often do human annotators agree on the same tag?
  - Penn Treebank: around 97%

# Qualitative evaluation

- Generate a confusion matrix (for development data): How often was a word with tag *i* mistagged as tag *j*:

Correct Tags

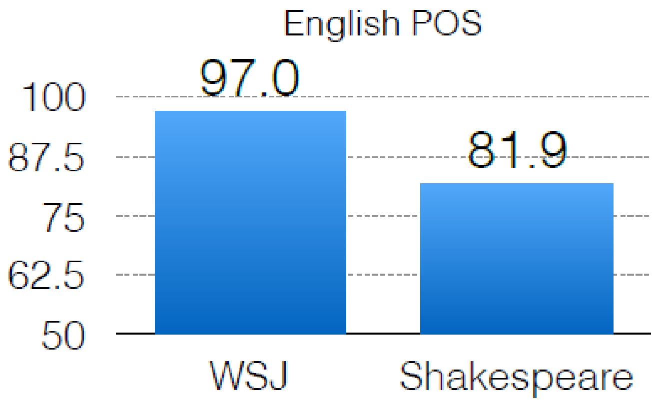
Predicted Tags

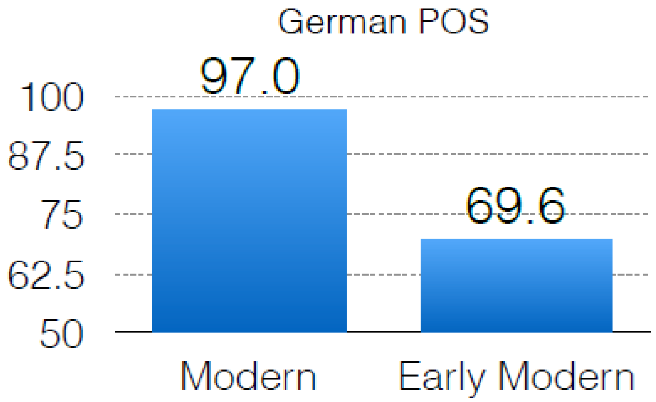
| | IN | JJ | NN | NNP | RB | VBD | VBN |
|---|---|---|---|---|---|---|---|
| IN | — | .2 | | | .7 | | |
| JJ | .2 | — | 3.3 | 2.1 | 1.7 | .2 | 2.7 |
| NN | | 8.7 | — | | | | .2 |
| NNP | .2 | 3.3 | 4.1 | — | .2 | | |
| RB | 2.2 | 2.0 | .5 | | — | | |
| VBD | | .3 | .5 | | | — | 4.4 |
| VBN | | 2.8 | | | | 2.6 | — |

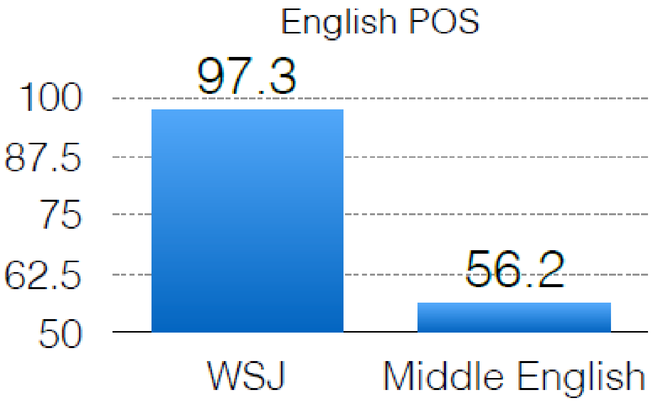
% of errors caused by mistagging VBN as JJ

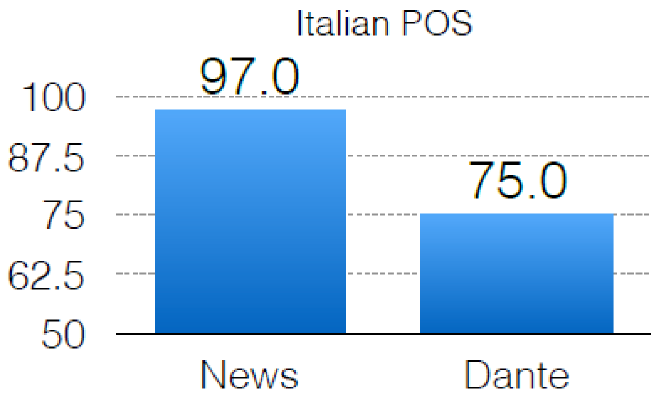
- See what errors are causing problems

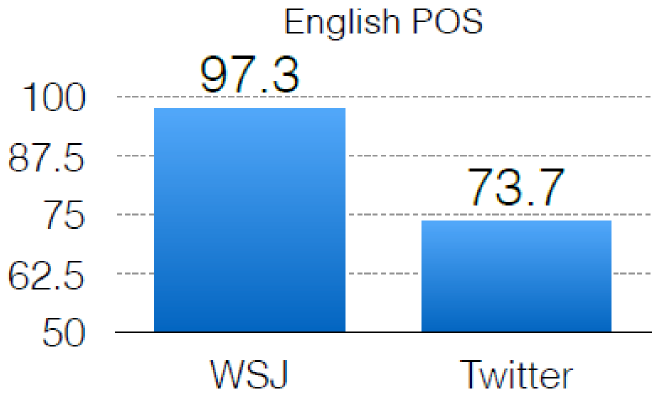
# Domain difference



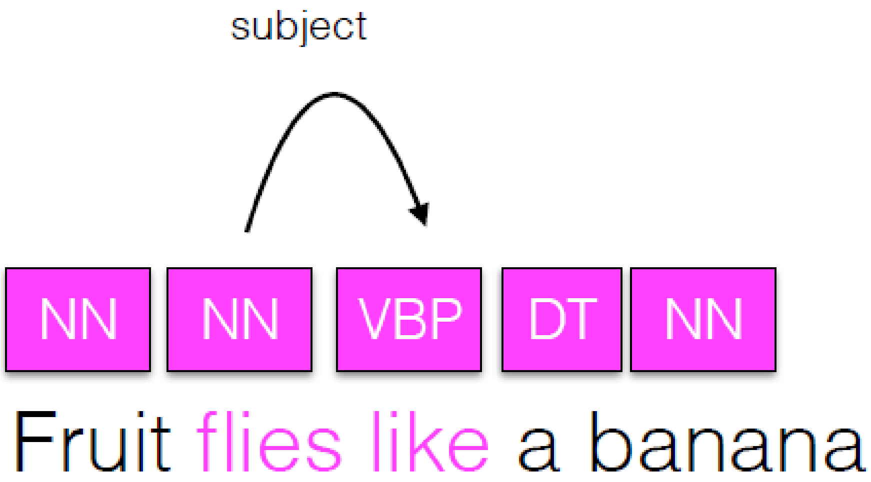



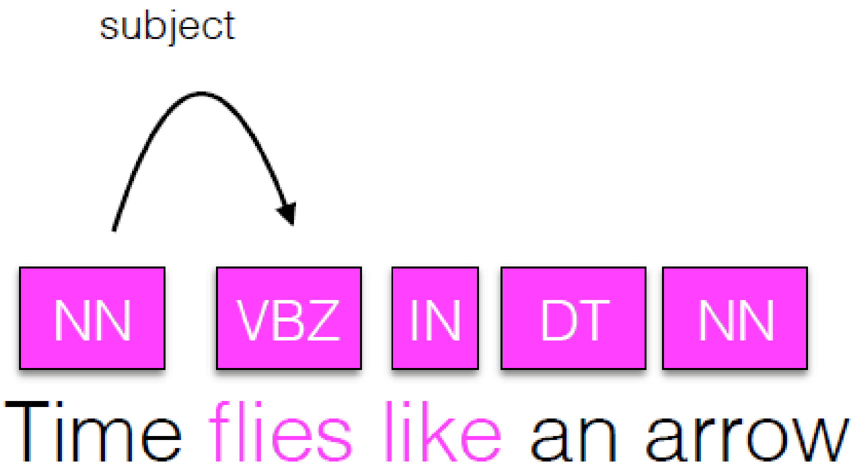



# Why is part of speech tagging useful?

# POS indicative of syntax



# POS indicative of MWE

at least one adjective/noun or noun phrase

and definitely one noun

((A | N)+ | ((A | N)*(NP))(A | N)*)N

| | |
|---|---|
| *AN*: | linear function; lexical ambiguity; mobile phase |
| *NN*: | regression coefficients; word sense; surface area |
| *AAN*: | Gaussian random variable; lexical conceptual paradigm; aqueous mobile phase |
| *ANN*: | cumulative distribution function; lexical ambiguity resolution; accessible surface area |
| *NAN*: | mean squared error; domain independent set; silica based packing |
| *NNN*: | class probability function; text analysis system; gradient elution chromatography |
| *NPN*: | degrees of freedom; [*no example*]; energy of adsorption |

Justeson and Katz 1995

# POS is indicative of pronunciation

- Content:
  - Noun: CONtent
  - Adjective: conTENT
- Object
  - Noun: OBject
  - Verb: obJECT

# Defining a Tag Set

**Tag sets have different granularities:**

- Brown corpus (Francis and Kucera 1982): 87 tags
- Penn Treebank (Marcus et al. 1993): 45 tags
- Simplified version of Brown tag set (de facto standard for English now)

  NN: common noun (singular or mass): *water*, *book*

  NNS: common noun (plural): *books*

# Verbs

| tag | description | example |
|---|---|---|
| VB | base form | I want to like |
| VBD | past tense | I/we/he/she/you liked |
| VBG | present participle | He was liking it |
| VBN | past participle | I had liked it |
| VBP | present (non 3rd-sing) | I like it |
| VBZ | present (3rd-sing) | He likes it |
| MD | modal verbs | He can go |

# Nouns

| | tag | description | example |
|---|---|---|---|
| non-proper | NN | non-proper, singular or mass | the company |
| | NNS | non-proper, plural | the companies |
| proper | NNP | proper, singular | Carolina |
| | NNPS | proper, plural | Carolinas |

# JJ (Adjectives)

- General adjectives
  - *happy person*
  - *new mail*
- Ordinal numbers
  - *fourth person*

```
2002 other/jj
1925 new/jj
1563 last/jj
1174 many/jj
1142 such/jj
1058 first/jj
 824 major/jj
 715 federal/jj
 698 next/jj
 644 financial/jj
```

Santorini 1990

# RB (Adverb)

- Most words that end in -ly
- Degree words (quite, too, very)
- Negative markers: not, n't, never

```
4410 n't/rb
2071 also/rb
1858 not/rb
1109 now/rb
1070 only/rb
1027 as/rb
 961 even/rb
 839 so/rb
 810 about/rb
 804 still/rb
```

Santorini 1990

# IN (preposition, subordinating conjunction)

- All prepositions (except *to*) and subordinating conjunctions
  - He jumped on the table because he was excited

```
31111 of/in
22967 in/in
11425 for/in
 7181 on/in
 6684 that/in
 6399 at/in
 6229 by/in
 5940 from/in
 5874 with/in
 5239 as/in
```

Santorini 1990

# Sequence labeling

$$x = \{x_1, \ldots, x_n\}$$

$$y = \{y_1, \ldots, y_n\}$$

- For a set of inputs x with n sequential time steps, one corresponding label $y_i$ for each $x_i$

# Named entity recognition

B-PERS I-PERS O O O O ORG

tim cook is the ceo of apple

3 or 4-class:

- person
- location
- organization
- (misc)

7-class:

- person
- location
- organization
- time
- money
- percent
- date

# Majority class

- Pick the label each word is seen most often with in the training data

| fruit | flies | like | a | banana |
|---|---|---|---|---|
| NN 12 | VBZ 7 | VB 74 | FW 8 | NN 3 |
| | NNS 1 | VBP 31 | SYM 13 | |
| | | JJ 28 | LS 2 | |
| | | IN 533 | JJ 2 | |
| | | | IN 1 | |
| | | | DT 25820 | |
| | | | NNP 2 | |

# Naive Bayes

- Treat each prediction as independent of the others

$$P(\mathrm{y} \mid x) = \frac{P(\mathrm{y})P(x \mid \mathrm{y})}{\sum_{y' \in \mathcal{Y}} P(y')P(x \mid y')}$$

$$P(\mathrm{VBZ} \mid flies) = \frac{P(\mathrm{VBZ})P(flies \mid \mathrm{VBZ})}{\sum_{y' \in \mathcal{Y}} P(y')P(flies \mid y')}$$

# Logistic regression

- Treat each prediction as independent of the others but condition on much more expressive set of features

$$P(y \mid x; \beta) = \frac{\exp\left(x^\top \beta_y\right)}{\sum_{y' \in \mathcal{Y}} \exp\left(x^\top \beta_{y'}\right)}$$

$$P(\mathrm{VBZ} \mid flies) = \frac{\exp\left(x^\top \beta_{\mathrm{VBZ}}\right)}{\sum_{y' \in \mathcal{Y}} \exp\left(x^\top \beta_{y'}\right)}$$

# Sequence

- Models that make independent predictions for elements in a sequence can reason over expressive representations of the input x (including correlations among inputs at different time steps $x_i$ and $x_j$.
- But they don’t capture another important source of information: correlations in the labels y.

# Sequences

- Most common tag bigrams inPenn Treebank training

| | | |
|---|---|---|
| DT | NN | 41909 |
| NNP | NNP | 37696 |
| NN | IN | 35458 |
| IN | DT | 35006 |
| JJ | NN | 29699 |
| DT | JJ | 19166 |
| NN | NN | 17484 |
| NN | , | 16352 |
| IN | NNP | 15940 |
| NN | . | 15548 |
| JJ | NNS | 15297 |
| NNS | IN | 15146 |
| TO | VB | 13797 |
| NNP | , | 13683 |
| IN | NN | 11565 |

# Sequences

| x | time | flies | like | an | arrow |
|---|---|---|---|---|---|
| y | NN | VBZ | IN | DT | NN |

$$P(y = \text{NN VBZ IN DT NN} \mid x = \text{time flies like an arrow})$$

# Generative vs. Discriminative models

- Generative models specify a joint distribution over the labels and the data. With this you could generate new data

$$P(x, y) = P(y)\, P(x \mid y)$$

- Discriminative models specify the conditional distribution of the label y given the data x. These models focus on how to discriminate between the classes

$$P(y \mid x)$$

# Generative

$$P(y \mid x) = \frac{P(x \mid y)P(y)}{\sum_{y' \in \mathcal{Y}} P(x \mid y)P(y)}$$

$$P(y \mid x) \propto P(x \mid y)P(y)$$

$$\max_y P(x \mid y)P(y)$$

How do we parameterize these probabilities when x and y are sequences?

# Hidden Markov Model

Prior probability of label sequence

$$P(y) = P(y_1, \ldots, y_n)$$

$$P(y_1, \ldots, y_n) \approx \prod_{i=1}^{n+1} P(y_i \mid y_{i-1})$$

- We’ll make a first-order Markov assumption and calculate the joint probability as the product the individual factors conditioned only on the previous tag.

# Hidden Markov Model

- Remember: a Markov assumption is an approximation to this
- exact decomposition (the chain rule of probability)

$$
\begin{aligned}
P(y_i, \ldots, y_n) = & \; P(y_1) \\
& \times P(y_2 \mid y_1) \\
& \times P(y_3 \mid y_1, y_2) \\
& \ldots \\
& \times P(y_n \mid y_1, \ldots, y_{n-1})
\end{aligned}
$$

# Hidden Markov Model

- Here again we’ll make a strong assumption: the probability of
- the word we see at a given time step is only dependent on its
- Label

$$P(x \mid y) = P(x_1, \ldots, x_n \mid y_1, \ldots, y_n)$$

$$P(x_1, \ldots, x_n \mid y_1, \ldots, y_n) \approx \prod_{i=1}^{N} P(x_i \mid y_i)$$

# Hidden Markov Model

NNP VBZ

| | |
|---|---|
| is | 1121 |
| has | 854 |
| says | 420 |
| does | 77 |
| plans | 50 |
| expects | 47 |
| ‘s | 40 |
| wants | 31 |
| owns | 30 |
| makes | 29 |
| hopes | 24 |
| remains | 24 |
| claims | 19 |
| seems | 19 |
| estimates | 17 |

NN VBZ

| | |
|---|---|
| is | 2893 |
| has | 1004 |
| does | 128 |
| says | 109 |
| remains | 56 |
| ‘s | 51 |
| includes | 44 |
| continues | 43 |
| makes | 40 |
| seems | 34 |
| comes | 33 |
| reflects | 31 |
| calls | 30 |
| expects | 29 |
| goes | 27 |

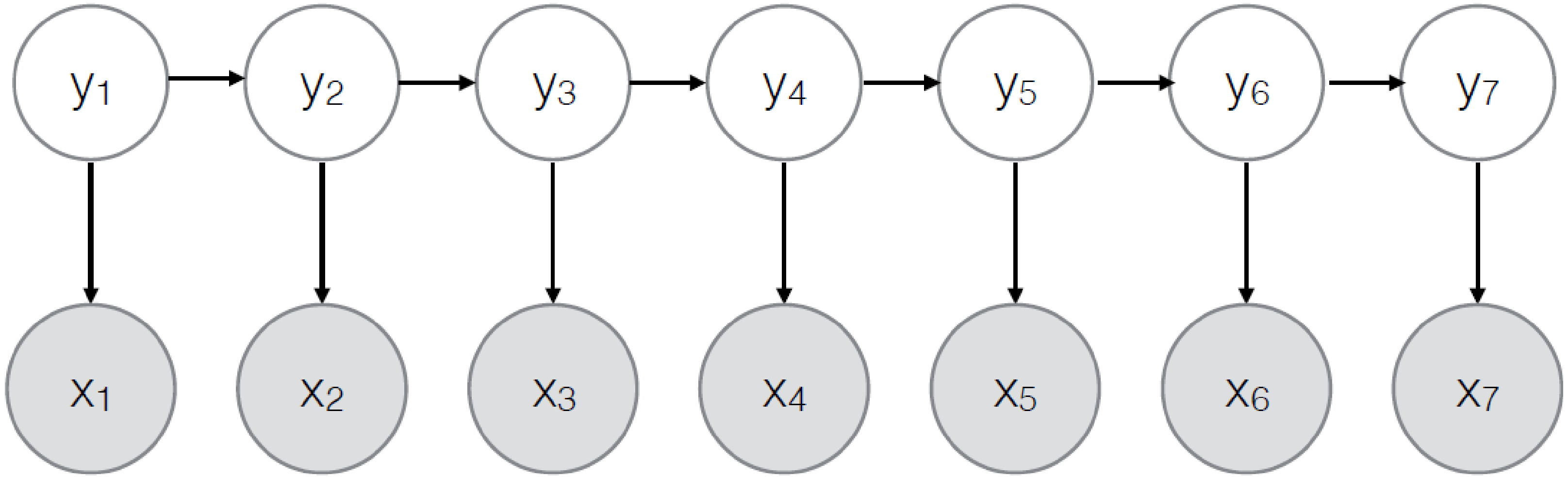
$$P(x_i \mid y_i, y_{i-1})$$

# HMM

$$P(x_1, \ldots, x_n, y_1, \ldots, y_n) \approx \prod_{i=1}^{n+1} P(y_i \mid y_{i-1}) \prod_{i=1}^{n} P(x_i \mid y_i)$$

# HMM

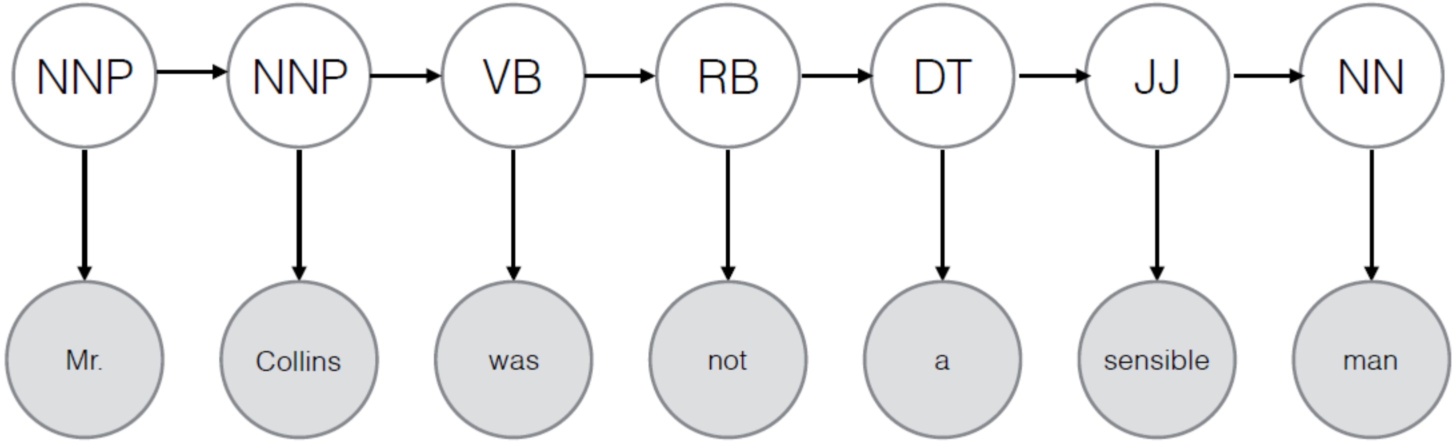
$P(y_3 \mid y_2)$




$P(x_3 \mid y_3)$

# HMM

$P(VB \mid NNP)$




$P(was \mid VB)$

# Parameter estimation

$$P(y_t \mid y_{t-1}) \qquad \frac{c(y_1, y_2)}{c(y_1)}$$

MLE for both is just counting
(as in Naive Bayes)

$$P(x_t \mid y_t) \qquad \frac{c(x, y)}{c(y)}$$