Lecture 2: Probability and Language Models

Intro to NLP, CS585, Fall 2014

Brendan O'Connor (http://brenocon.com)

Admin

- Waitlist
- Moodle access: Email me if you don't have it
- Did you get an announcement email?
- Piazza vs Moodle?
- Office hours today

Things today

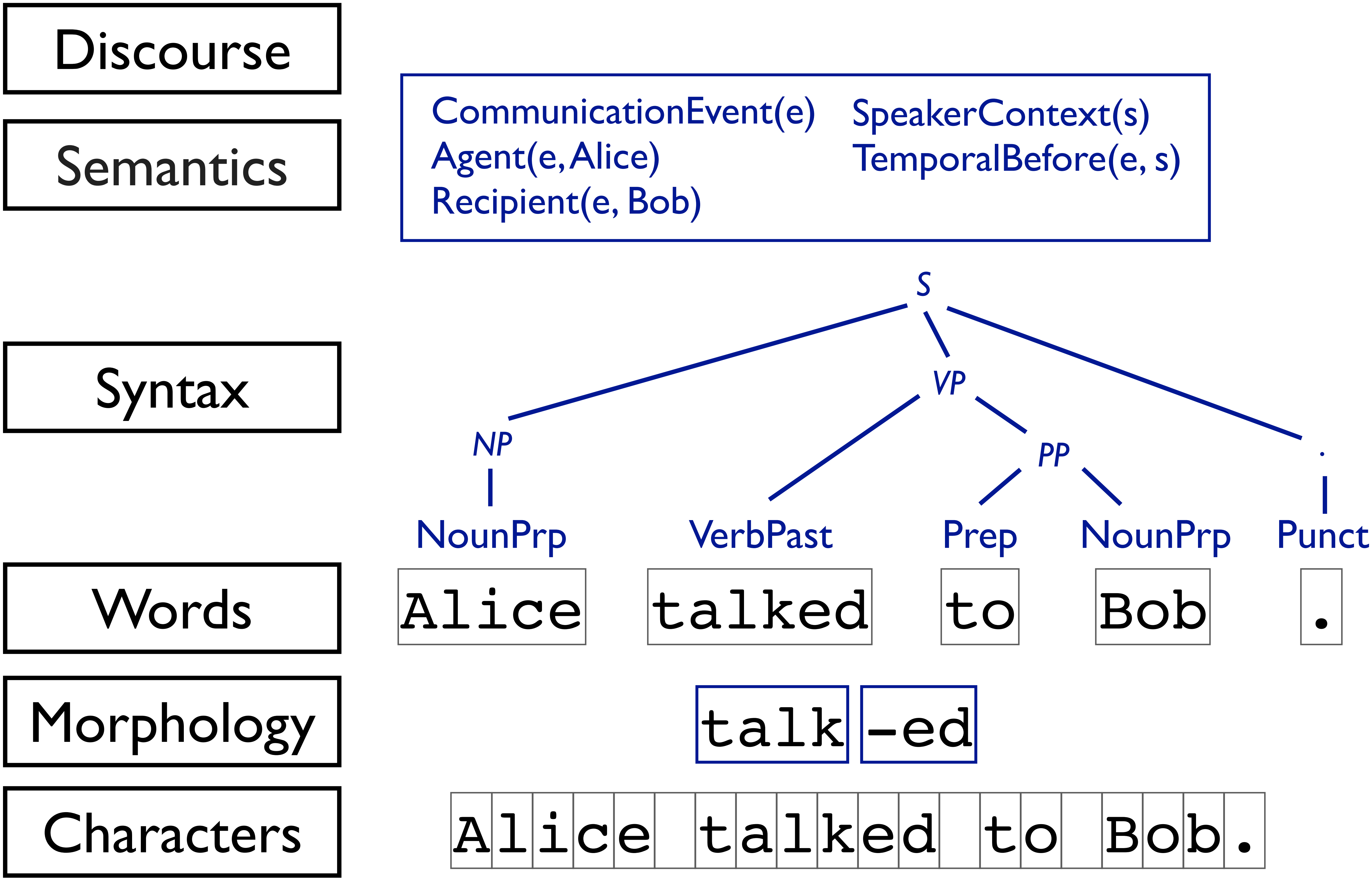
- Homework: ambiguities
- Python demo
- Probability Review
- Language Models

Python demo

- [TODO link ipython-notebook demo]

- For next week, make sure you can run
  - Python 2.7 (Built-in on Mac & Linux)
  - IPython Notebook http://ipython.org/notebook.html
    - Please familiarize yourself with it.
    - Python 2.7, IPython 2.2.0
  - Nice to have: Matplotlib
- Python interactive interpreter
- Python scripts

Levels of linguistic structure



Levels of linguistic structure

Words are fundamental units of meaning

and easily identifiable*

*in some languages

Words	`Alice`	`talked`	`to`	`Bob`	`.`

Characters `Alice talked to Bob.`

Probability theory

Review: definitions/laws

$$= \sum_a P(A = a)$$

Conditional Probability $$= \frac{P(AB)}{P(B)}$$

Chain Rule $$= P(A|B)P(B)$$

Law of Total Probability
$$= \sum_b P(A, B = b)$$
$$= \sum_b P(A|B = b)P(B = b)$$

Disjunction (Union) $$P(A \vee B) =$$

Negation (Complement) $$P(\neg A) =$$

Bayes Rule

Want P(H|D) but only have P(D|H)
e.g. H causes D, or P(D|H) is easy to measure...

H: who wrote this document?

Model: authors' word probs

D: words

Bayesian inference

Likelihood

Prior

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)}$$

Posterior

Normalizer

Rev. Thomas Bayes
c. 1701-1761

Bayes Rule and its pesky denominator

Likelihood

Prior

$$P(h|d) = \frac{P(d|h)P(h)}{P(d)} = \frac{P(d|h)P(h)}{\sum_{h'} P(d|h')P(h')}$$

$$P(h|d) = \frac{1}{Z} P(d|h)P(h)$$

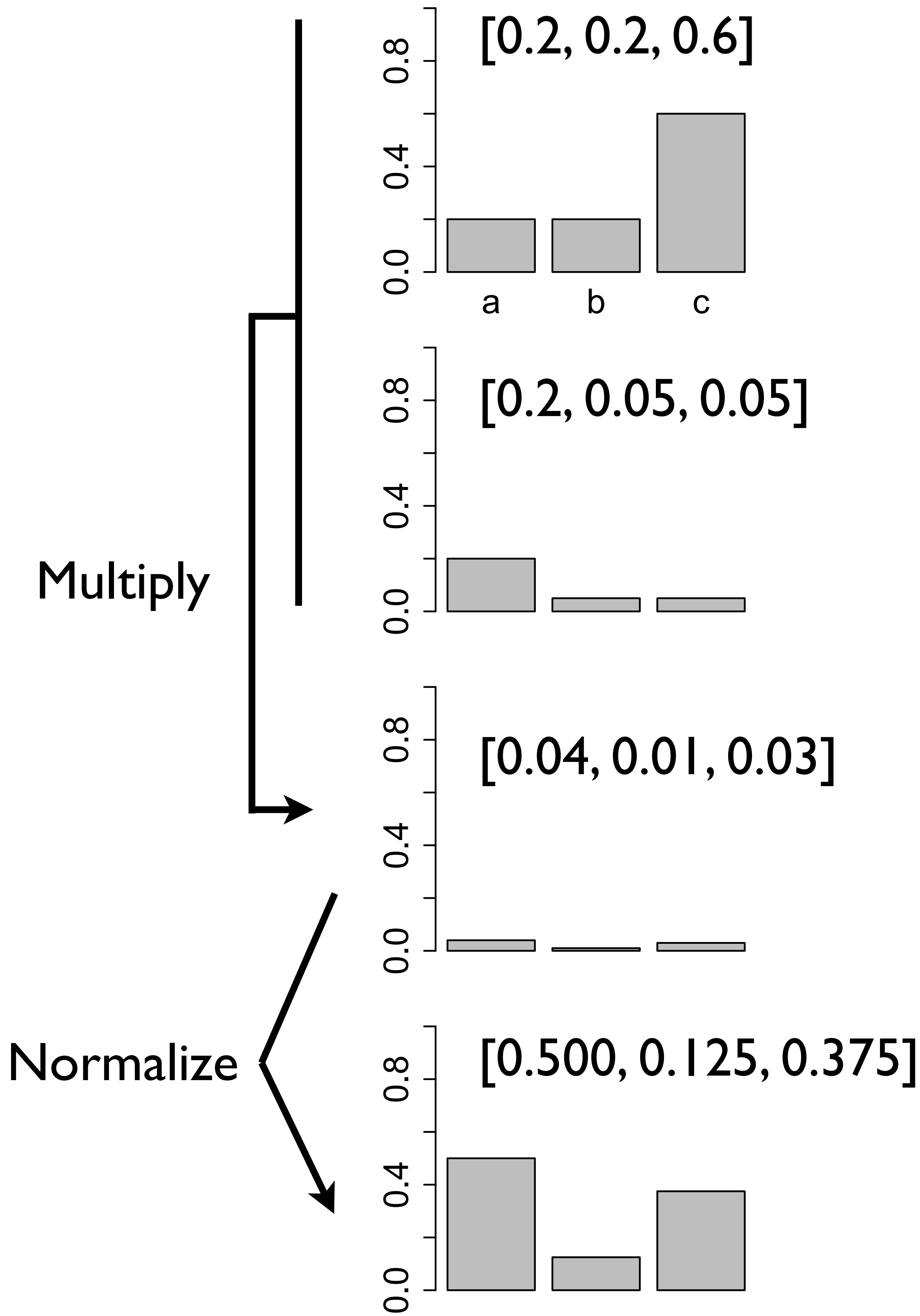
Z: whatever lets the posterior, when summed across *h*, to sum to 1
Zustandssumme, "sum over states"

$$P(h|d) \propto \underline{P(d|h)P(h)}$$

"Proportional to"
(implicitly for varying H.
This notation is very common, though slightly ambiguous.)

Unnormalized posterior
By itself does not sum to 1!

Bayes Rule: Discrete



Sum to 1?

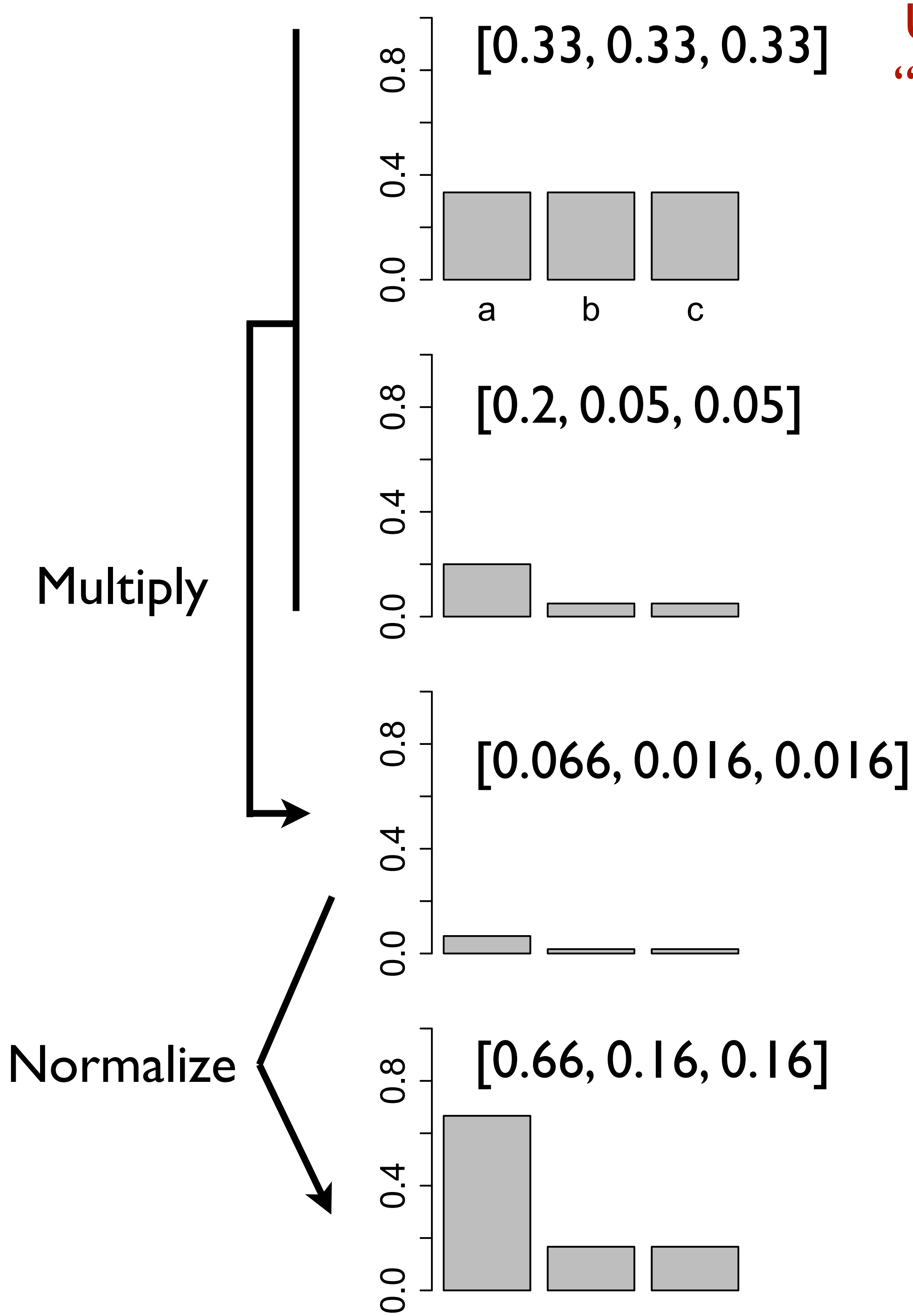
$P(H = h)$
Prior — Yes

$P(E|H = h)$
Likelihood — No

$P(E|H = h)P(H = h)$
Unnorm. Posterior — No

$\frac{1}{Z}P(E|H = h)P(H = h)$
Posterior — Yes

Bayes Rule: Discrete, uniform prior



Uniform distribution:
"Uninformative prior"

Sum to 1?

$$P(H = h)$$

Prior — Yes

$$P(E|H = h)$$

Likelihood — No

$$P(E|H = h)P(H = h)$$

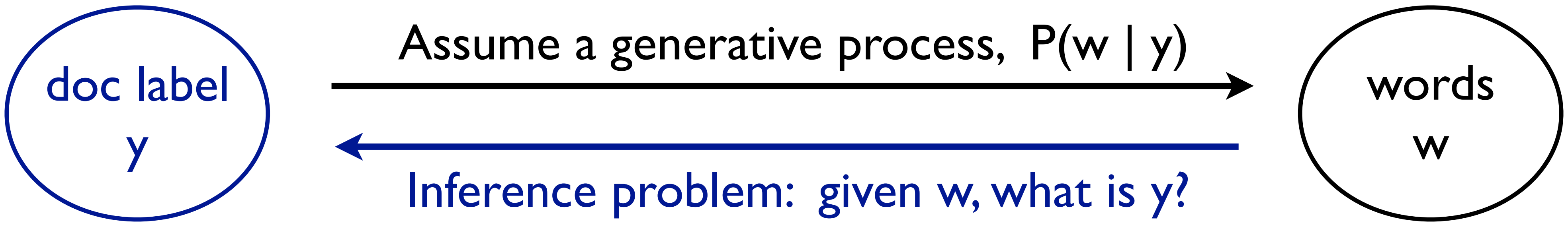
Unnorm. Posterior — No

Uniform prior implies that
posterior is just renormalized likelihood

$$\frac{1}{Z}P(E|H = h)P(H = h)$$

Posterior — Yes

Bayes Rule for doc classification



If we knew $P(w|y)$

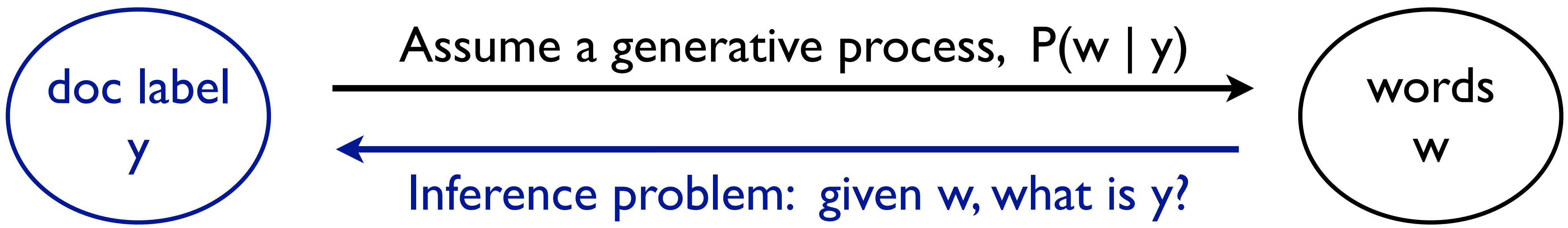
We could estimate $P(y|w) \propto P(y)P(w|y)$

	abracadabra	*gesundheit*
Anna	5 per 1000 words	6 per 1000 words
Barry	10 per 1000 words	1 per 1000 words

Look at random word.
It is *abracadabra*

Assume 50% prior prob
Prob author is Anna?

Bayes Rule for doc classification



If we knew $P(w|y)$

We could estimate $P(y|w) \propto P(y)P(w|y)$

	abracadabra	*gesundheit*
Anna	5 per 1000 words	6 per 1000 words
Barry	10 per 1000 words	1 per 1000 words

Look at two random words.

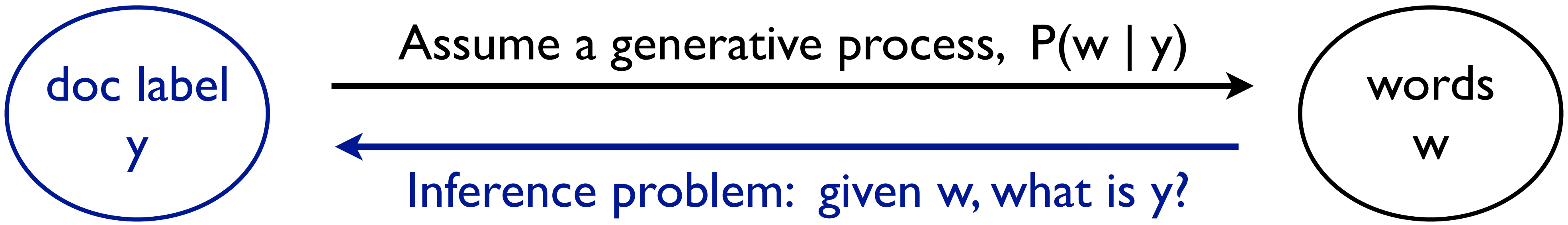
w_1 = *abracadabra*

w_2 = *gesundheit*

Assume 50% prior prob

Prob author is Anna?

Bayes Rule for doc classification



If we knew $P(w|y)$

We could estimate $P(y|w) \propto P(y)P(w|y)$

	abracadabra	*gesundheit*
Anna	5 per 1000 words	6 per 1000 words
Barry	10 per 1000 words	1 per 1000 words

Look at two random words.

w_1 = *abracadabra*

w_2 = *gesundheit*

Chain rule:

$P(w_1, w_2 \mid y) = P(w_1 \mid w_2 \ y) \ P(w_2 \mid y)$

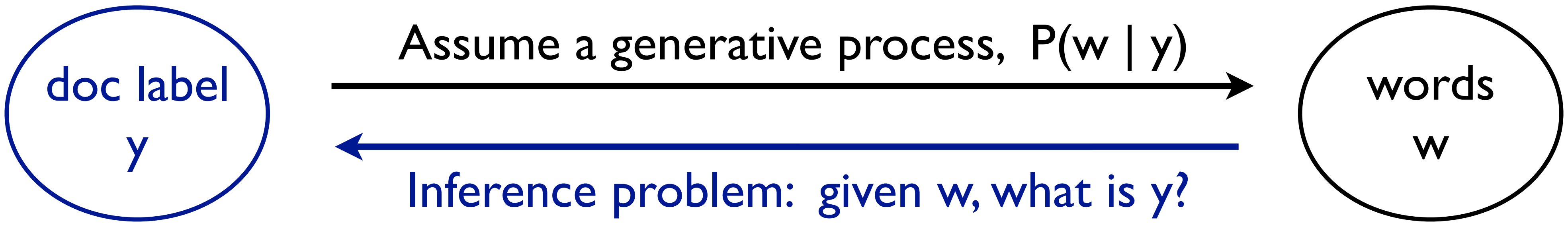
ASSUME conditional independence:

$P(w_1, w_2 \mid y) = P(w_1 \mid y) \ P(w_2 \mid y)$

Assume 50% prior prob

Prob author is Anna?

Cond indep. assumption: "Naive Bayes"



$$P(w_1 \ldots w_T \mid y) = \prod_{t=1}^{T} P(w_t \mid y)$$

$$\text{each } w_t \in 1..V$$ *V* = vocabulary size

Generative story ("Multinom NB" [McCallum & Nigam 1998]):

- For each token *t* in the document,
  - Author chooses a word by rolling the same weighted V-sided die

This model is wrong!

How can it possibly be useful for doc classification?

Bayes Rule for *text* inference

Noisy channel model

Original text → Hypothesized transmission process → Observed data

Original text ← Inference problem ← Observed data

Codebreaking

P(plaintext | encrypted text) $\propto$ P(encrypted text | plaintext) P(plaintext)

Bletchley Park (WWII)

Enigma machine

Bayes Rule for *text* inference

Noisy channel model

Original text

Hypothesized transmission process

Inference problem

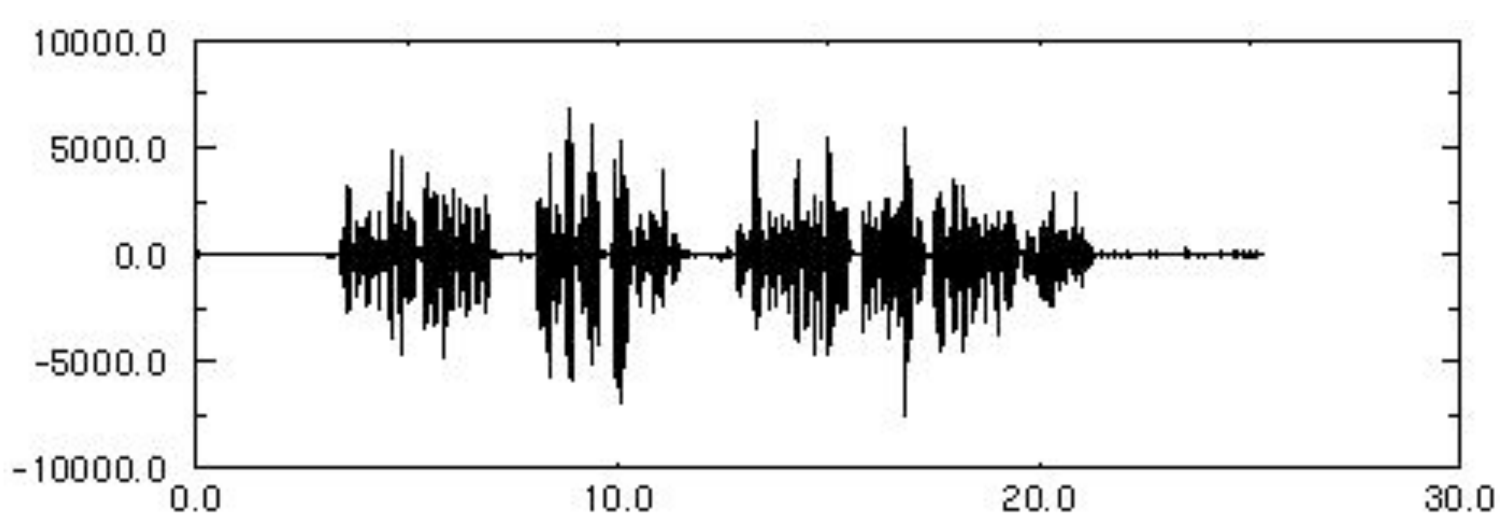
Observed data

Codebreaking

P(plaintext | encrypted text) $\propto$ P(encrypted text | plaintext) P(plaintext)

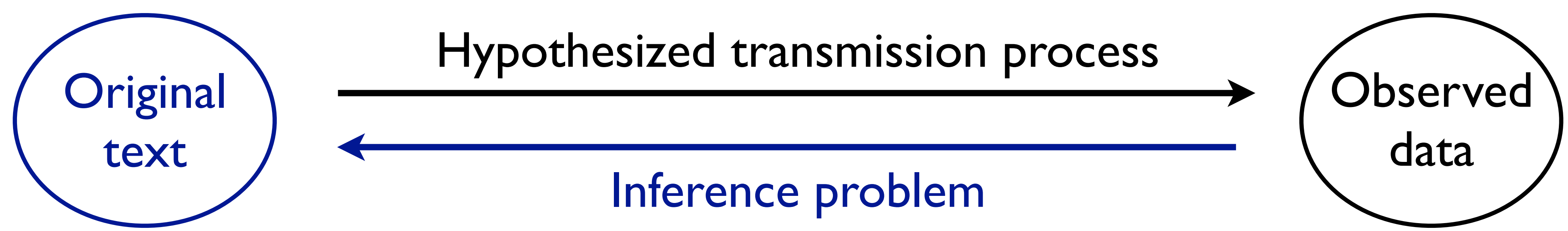
Speech recognition

P(text | acoustic signal) $\propto$ P(acoustic signal | text) P(text)



Bayes Rule for *text* inference

Noisy
channel
model



Codebreaking

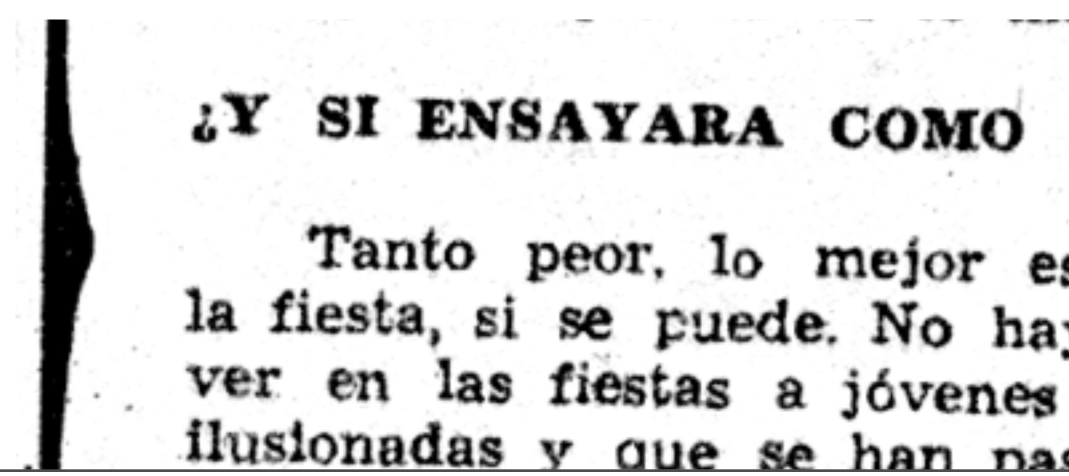
P(plaintext | encrypted text) $\propto$ P(encrypted text | plaintext) P(plaintext)

Speech recognition

P(text | acoustic signal) $\propto$ P(acoustic signal | text) P(text)

Optical character recognition

P(text | image) $\propto$ P(image | text) P(text)

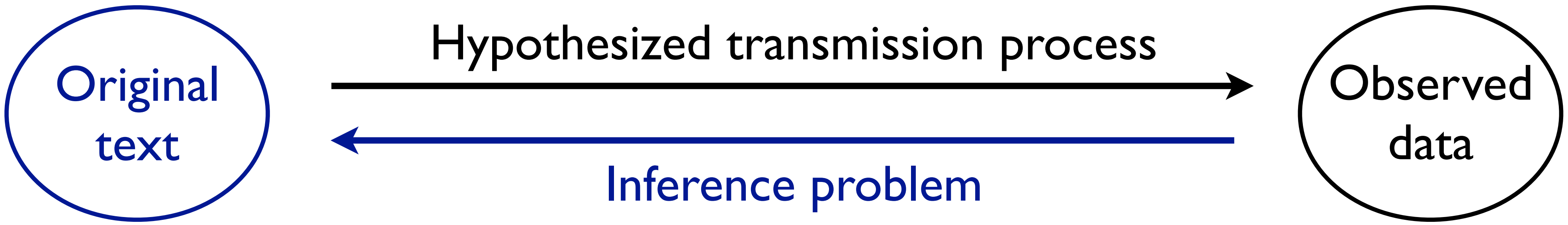


¿Y SI ENSAYARA COMO

Tanto peor, lo mejor es
la fiesta, si se puede. No hay
ver en las fiestas a jóvenes
ilusionadas y que se han pas

Bayes Rule for *text* inference

Noisy channel model



Codeb

P(plaintex text)

Speech

P(text | a

Optical

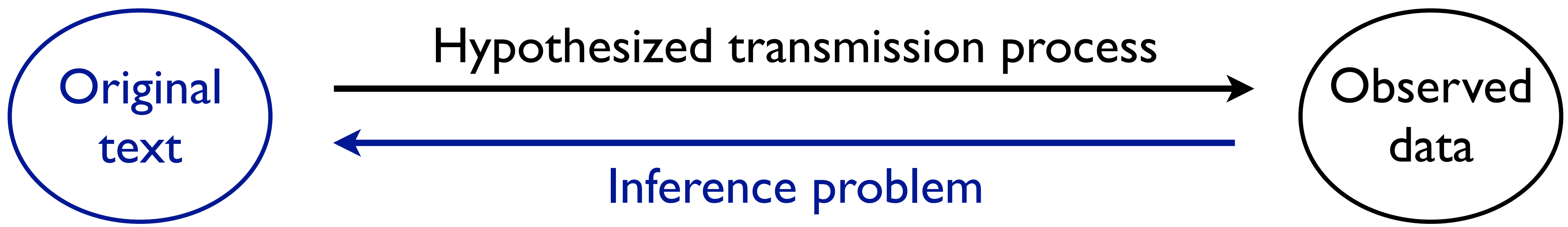
P(text | i

One naturally wonders if the problem of translation could conceivably be treated as a problem in cryptography. When I look at an article in Russian, I say: 'This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode.'

-- Warren Weaver (1955)

Bayes Rule for *text* inference

Noisy channel model



Codebreaking

P(plaintext | encrypted text) $\propto$ P(encrypted text | plaintext) P(plaintext)

Speech recognition

P(text | acoustic signal) $\propto$ P(acoustic signal | text) P(text)

Optical character recognition

P(text | image) $\propto$ P(image | text) P(text)

Machine translation?

P(target text | source text) $\propto$ P(source text | target text) P(target text)