

Some yinkish dripners blorked quastofically into the nindin with the pidibs

words have categories

Some/**DET** yinkish/**ADJ** dripners/**NOUN** blorked/**VERB** quastofically/**ADV**  
into/**PREP** the/**DET** nindin/**NOUN** with/**PREP** the/**DET** pidibs/**NOUN**

## Lecture 3: Part-of-Speech Tagging and Log-Linear Models

1. Labeling words
2. The statistical perspective
3. Corpora
4. Log-linear models
5. Evaluation

some slides  
are from  
Ann Copestake

# Labeling Words

Fish fish fish.

# Fish fish fish.

## fish

*noun*

US 🗣️ /fɪʃ/ UK 🗣️ /fɪʃ/

plural **fish** or **fishes**



Lew Robertson/Photolibrary  
/GettyImages

**A1** [ C or U ]

**an animal that lives in water, is covered with scales, and breathes by taking water in through its mouth, or the flesh of these animals eaten as food:**

- *Several large fish live in the pond.*
- *Sanjay **caught** the biggest fish I've ever seen.*
- *I don't like fish (= don't like to eat fish).*

[dictionary.cambridge.org/us/dictionary/english/fish](https://dictionary.cambridge.org/us/dictionary/english/fish)

# Fish fish fish.

**fish verb (ANIMAL)**

---

**B1** [ I or T ]

**to catch fish from a river, sea, lake, etc., or to try to do this:**

- *They're fishing **for** tuna.*
- *The sea here has been fished intensely over the last ten years.*

[dictionary.cambridge.org/us/dictionary/english/fish](https://dictionary.cambridge.org/us/dictionary/english/fish)

Part-of-speech tagging is useful

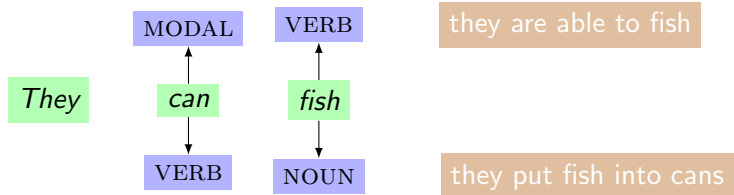
Fish/**NOUN** fish/**VERB** fish/**NOUN**



from FINDING NEMO MOVIE (2013)

photo: [www.avforums.com/reviews/finding-nemo-movie-review.6237](http://www.avforums.com/reviews/finding-nemo-movie-review.6237)

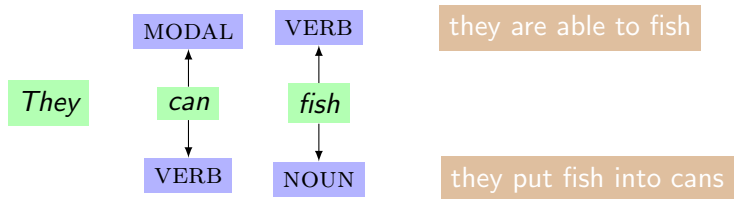
# Ambiguity



## Ambiguity

- *can*: modal verb, verb, singular noun
- *fish*: verb, singular noun, plural noun

# Ambiguity



## Ambiguity

- *can*: modal verb, verb, singular noun
- *fish*: verb, singular noun, plural noun

application-independent tags;  
linguistic knowledge involved



# Why POS tag?

*Coarse-grained syntax / word sense disambiguation*: fast, so applicable to very large corpora.

- Some linguistic research and lexicography: e.g., how often is *tango* used as a verb? *dog*?
- Named entity recognition and similar tasks (finite state patterns over POS tagged data).
- Features for machine learning e.g., sentiment classification. (e.g., *stink*/VERB vs *stink*/NOUN).
- Fast preliminary processing for full parsing: provide guesses at unknown words, cut down search space.

# Information extraction (1)

## Book a flight

- Leave London on 1<sup>st</sup> Dec 2020
- Arrive at London on 1<sup>st</sup> Dec 2020

FROM		
TO		
TIME		

# Information extraction (1)

## Book a flight

- Leave/○ London/B-FROM on/○ 1<sup>st</sup>/B-TIME Dec/I-TIME 2020/E-TIME
- Arrive/○ at/○ London/B-TO on/○ 1<sup>st</sup>/B-TIME Dec/I-TIME 2020/E-TIME

FROM	London	
TO		London
TIME	1 <sup>st</sup> Dec 2020	1 <sup>st</sup> Dec 2020

## Chunking

B	begin of $X$
I	inside $X$
E	end of $X$
O	outside $X$

# Information extraction (1)

## Book a flight

- Leave/○ London/B-FROM on/○ 1<sup>st</sup>/B-TIME Dec/I-TIME 2020/E-TIME
- Arrive/○ at/○ London/B-TO on/○ 1<sup>st</sup>/B-TIME Dec/I-TIME 2020/E-TIME

FROM	London	
TO		London
TIME	1 <sup>st</sup> Dec 2020	1 <sup>st</sup> Dec 2020

## Chunking

B	begin of $X$
I	inside $X$
E	end of $X$
O	outside $X$

application-dependent tags;  
contextual information matters

## Information extraction (2)

### Entity linking

from BBC news

*Time is running out for **Brussels** and **London** to reach a post-Brexit trade deal.*

***Downing Street** said **Johnson**, 55, is in extremely good spirits at the St Thomas' Hospital ward as his father, Stanley Johnson, called on his son to rest up.*

# Information extraction (2)

## Entity linking

from BBC news

*Time is running out for Brussels/European\_Council and London/Government\_of\_the\_United\_Kingdom to reach a post-Brexit trade deal.*

*Downing Street/Goverment\_of\_the\_United\_Kingdom said Johnson/Boris\_Johnson, 55, is in extremely good spirits at the St Thomas' Hospital ward as his father, Stanley Johnson, called on his son to rest up.*



application-dependent tags; world knowledge involved

# The Statistical Perspective

*The actual science of logic is conversant at present only with things either certain, impossible, or entirely doubtful, none of which (fortunately) we have to reason on. Therefore the true logic for this world is **the calculus of probabilities**, which takes account of the magnitude of the probability which is, or ought to be, in a reasonable man's mind.*



James C Maxwell



Corpora

# Data in NLP

- *Corpus*: text that has been collected for some purpose.
- balanced corpus: texts representing different genres  
genre is a type of text (vs domain)
- *Tagged corpus*: a corpus annotated with POS tags
- *Treebank*: a corpus annotated with parse trees
- *Specialist corpora*—e.g., collected to train or evaluate particular applications

Movie reviews for sentiment classification

Data collected from simulation of a dialogue system

# Be careful

Data may be very *difficult to acquire*

- first language acquisition
- historical linguistics
- brain activities
- dolphin language

▷ takes years to collect

▷ no longer exist

▷ wonderful machines, e.g. fMRI

▷ ...

Data may be extremely *big*

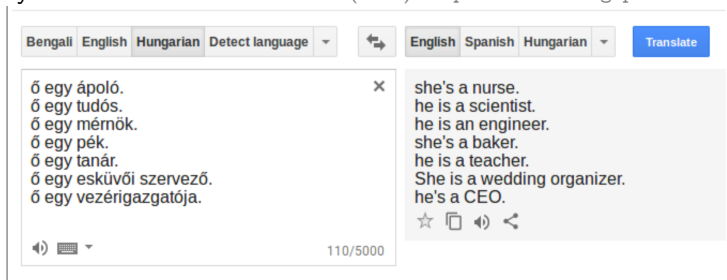
- e.g. data from twitter

Data may be *private*

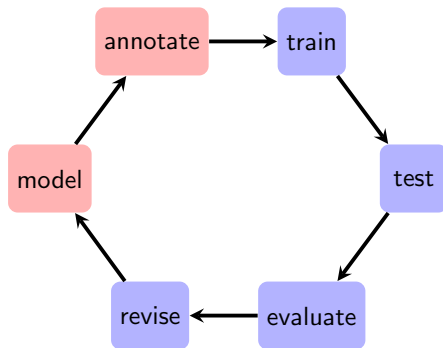
- the Cambridge Analytica/Facebook scandal

Data may be *biased*

Prates et al. (2019) <https://arxiv.org/pdf/1809.02208.pdf>



# Annotations in NLP



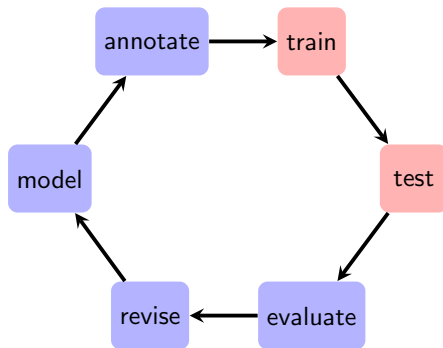
## MATTER: the annotation development cycle

**Model** Structural descriptions provide theoretically informed attributes derived from empirical observations over the data.

**Annotate** An annotation scheme assumes a feature set that encodes specific structural descriptions and properties of the input data.

Pustejovsky and Stubbs (2012)

# Annotations in NLP



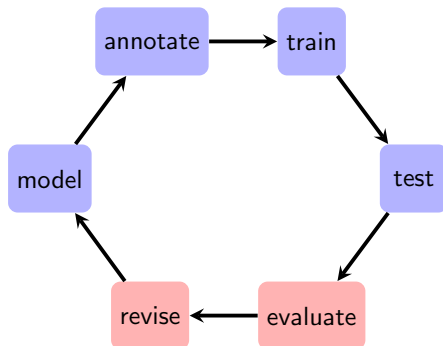
## MATTER: the annotation development cycle

**Train** The algorithm is trained over a corpus annotated with the target feature set.

**Test** The algorithm is tested against held-out data.

Pustejovsky and Stubbs (2012)

# Annotations in NLP



## MATTER: the annotation development cycle

**Evaluate** A standardized evaluation of results is conducted.

**Revise** The model and the annotation specification are revisited in order to make the annotation more robust and reliable with use in the algorithm.

Pustejovsky and Stubbs (2012)

# Tagset (CLAWS 5)

*tagset*: standardized codes for fine-grained parts of speech.

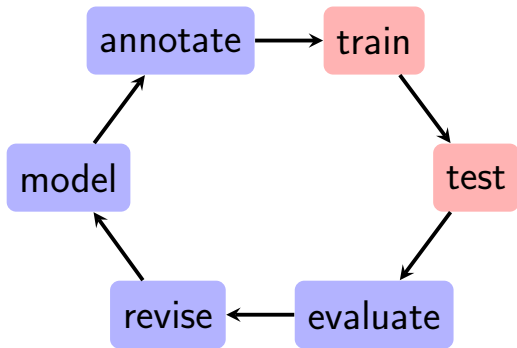
*CLAWS 5*: over 60 tags, including:

NN1	singular noun	NN2	plural noun
PNP	personal pronoun	VM0	modal auxiliary verb
VVB	base form of verb	VVI	infinitive form of verb

- They/*PNP* can/*VM0* fish/*VVI* ./PUN
- They/*PNP* can/*VVB* fish/*NN2* ./PUN
- They/*PNP* can/*VM0* fish/*NN2* ./PUN
- etc

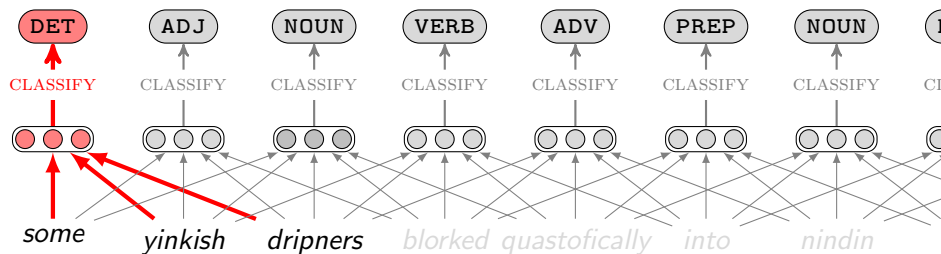
no full parse

## Log-Linear Models

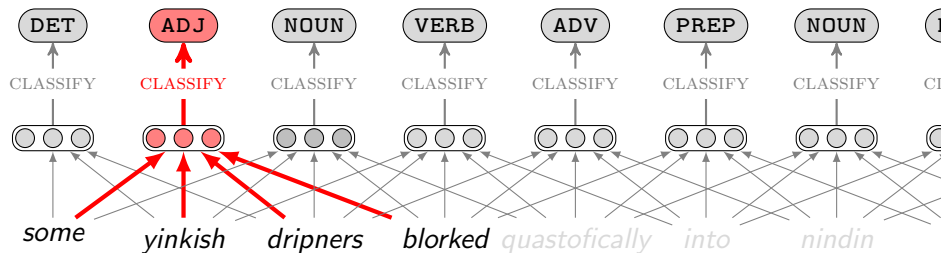




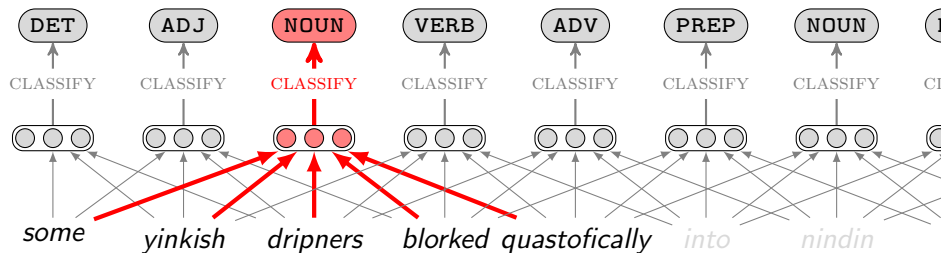
# POS tagging and prediction



# POS tagging and prediction



# POS tagging and prediction



## Aspects of POS tagging

*Some yinkish dripners blorked quastofically into the nindin with ...*

# Aspects of POS tagging

word=*dripners*

*Some yinkish dripners blorked quastofically into the nindin with ...*

the word itself

# Aspects of POS tagging

word=*dripners*

*Some yinkish dripners blorked quastofically into the nindin with ...*

suf<sub>-3,-2</sub>=*er*  
suf<sub>-1</sub>=*s*

morphological features

## Aspects of POS tagging

word<sub>i-2</sub>=some  
word<sub>i-1</sub>=yinkish

word=dripners

*Some yinkish dripners blorked quastofically into the nindin with ...*

suf<sub>-3,-2</sub>=er  
suf<sub>-1</sub>=s

POS can be defined distributionally

## Aspects of POS tagging

word<sub>i-2</sub>=some  
word<sub>i-1</sub>=yinkish

word<sub>i+2</sub>=quastofically  
word<sub>i+1</sub>=bloked

word=dripners

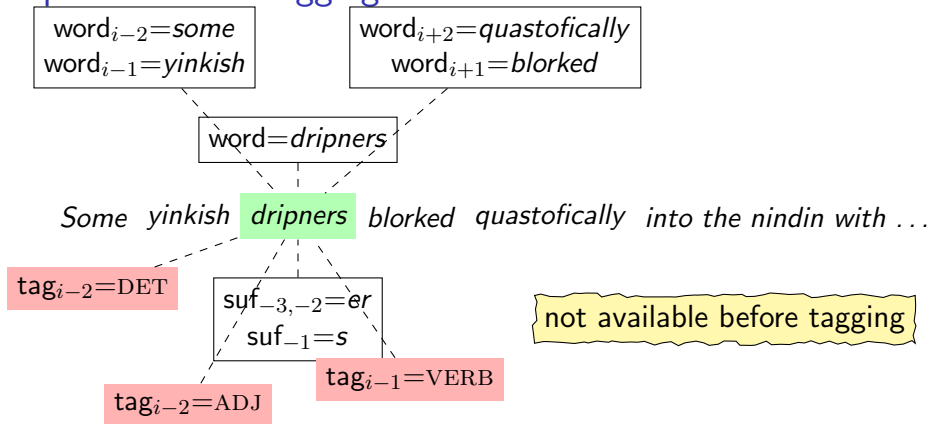
Some yinkish **dripners** bloked quastofically into the nindin with ...

suf<sub>-3,-2</sub>=er  
suf<sub>-1</sub>=s

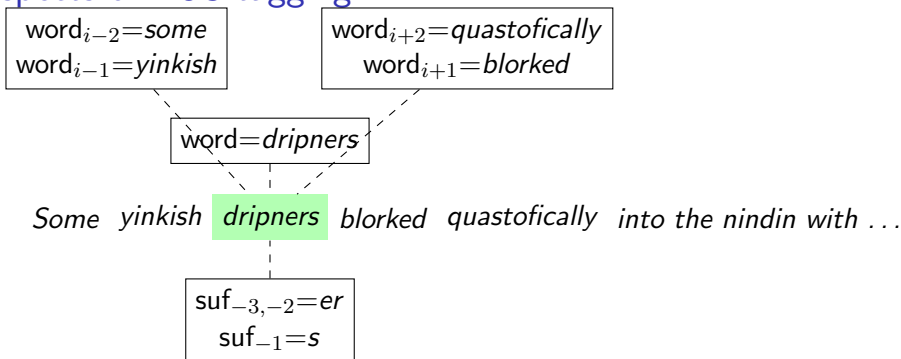
POS can be defined distributionally



## Aspects of POS tagging



## Aspects of POS tagging

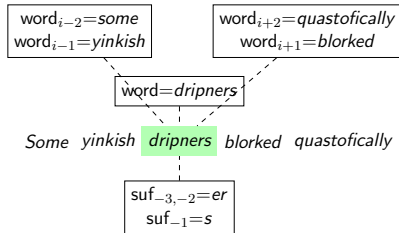


The task: model the distribution

$$p(t_i | w_1, \dots, w_n)$$

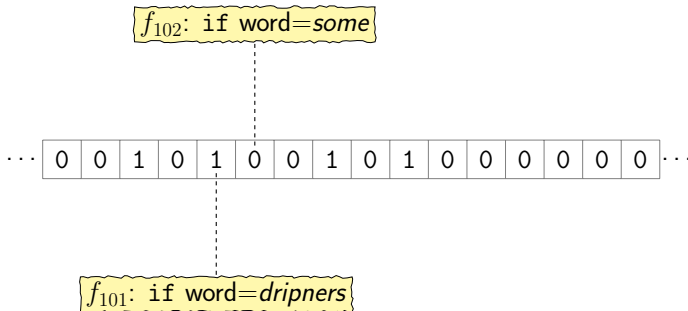
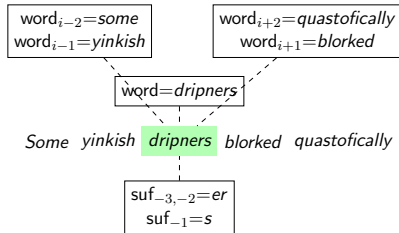
Many *features* may be relevant. Usually we only consider *local* features.

# Feature vector representations

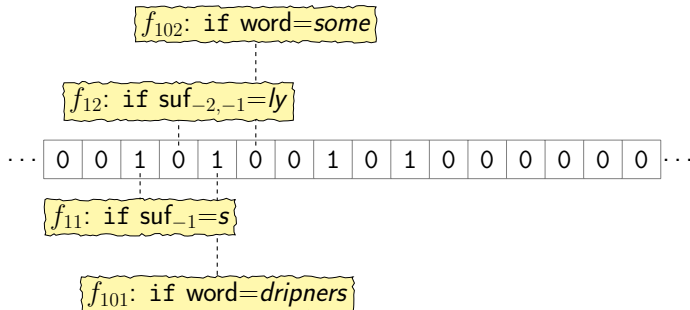
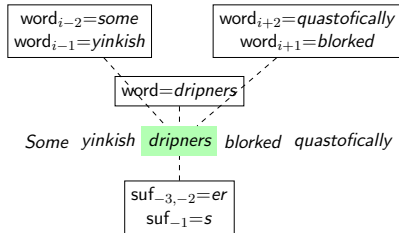


... 0 0 1 0 1 0 0 1 0 1 0 0 0 0 0 0 ...

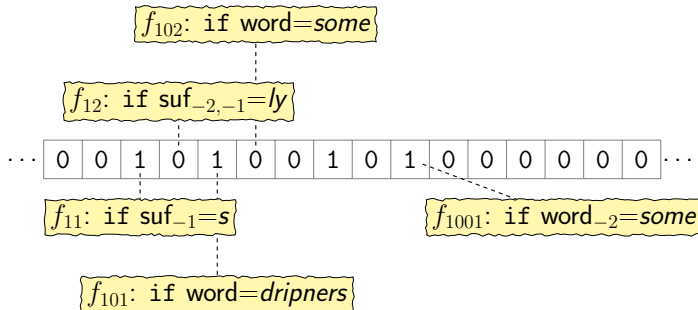
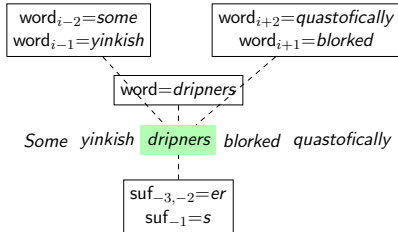
# Feature vector representations



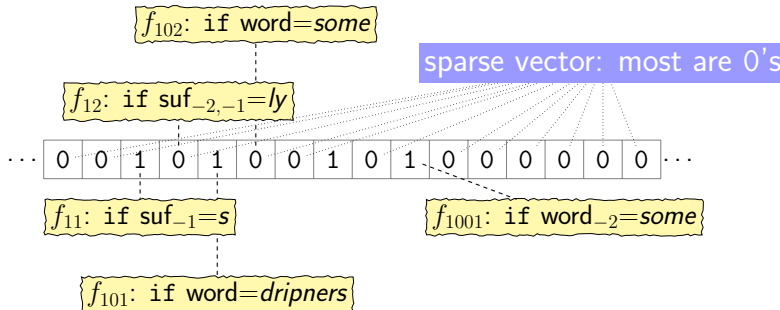
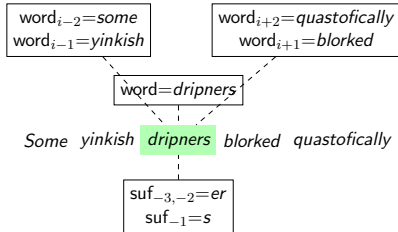
# Feature vector representations



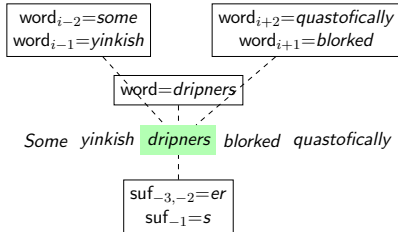
# Feature vector representations



# Feature vector representations



# Feature vector representations



$f_{102}$ : if word=some and tag=N

$f_{12}$ : if suf<sub>-2,-1</sub>=ly and tag=N

... 0 0 1 0 1 0 0 1 0 1 0 0 0 0 0 0 ...

$f_{11}$ : if suf<sub>-1</sub>=s and tag=N

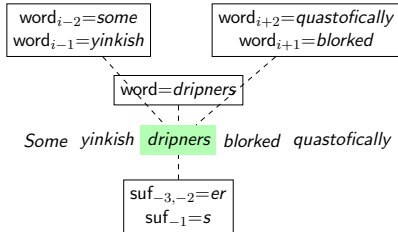
$f_{1001}$ : if word<sub>-2</sub>=some and tag=N

$f_{101}$ : if word=dripners and tag=N

sparse vector: most are 0's



# Feature vector representations



$f_{102+kD}$ : if word=some and tag=N

sparse vector: most are 0's

$f_{12+kD}$ : if suf<sub>-2,-1</sub>=ly and tag=N

... 0 0 1 0 1 0 0 1 0 1 0 0 0 0 0 0 ...

$f_{11+kD}$ : if suf<sub>-1</sub>=s and tag=N

$f_{1001+kD}$ : if word<sub>-2</sub>=some and tag=N

$f_{101+kD}$ : if word=dripners and tag=N

$$f(x, y)$$

$$x = \langle w_1, \dots, w_n, i \rangle$$

$$y = t_i$$

## Log-linear models

Assume we have a *parameter vector*  $\theta \in \mathbb{R}^m$ .

## Log-linear models

Assume we have a *parameter vector*  $\theta \in \mathbb{R}^m$ .

We define

$$p(y|x; \theta) = \frac{\exp(\theta^\top f(x, y))}{\sum_{y' \in \mathcal{Y}} \exp(\theta^\top f(x, y'))}$$

# Log-linear models

Assume we have a *parameter vector*  $\theta \in \mathbb{R}^m$ .

We define

$$p(y|x; \theta) = \frac{\exp(\theta^\top f(x, y))}{\sum_{y' \in \mathcal{Y}} \exp(\theta^\top f(x, y'))}$$

Why the name

$$\log p(y|x; \theta) = \underbrace{\theta^\top f(x, y)}_{\text{linear term}} - \underbrace{\log \sum_{y' \in \mathcal{Y}} \exp(\theta^\top f(x, y'))}_{\text{normalization term}}$$

# Log-linear models

Assume we have a *parameter vector*  $\theta \in \mathbb{R}^m$ .

We define

$$p(y|x; \theta) = \frac{\exp(\theta^\top f(x, y))}{\sum_{y' \in \mathcal{Y}} \exp(\theta^\top f(x, y'))}$$

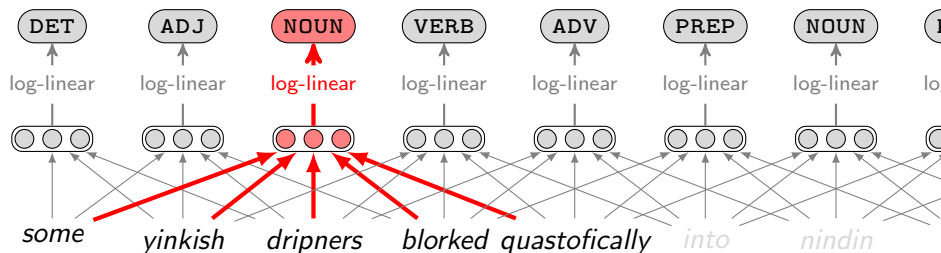
Why the name

$$\log p(y|x; \theta) = \underbrace{\theta^\top f(x, y)}_{\text{linear term}} - \underbrace{\log \sum_{y' \in \mathcal{Y}} \exp(\theta^\top f(x, y'))}_{\text{normalization term}}$$

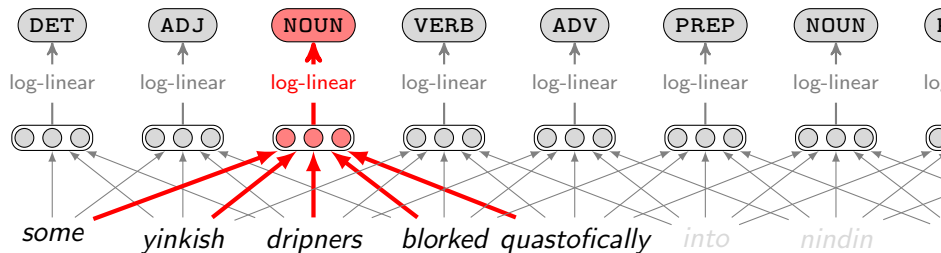
Prediction

$$\arg \max_{y^* \in \mathcal{Y}} p(y|x; \theta) = \arg \max_{y^* \in \mathcal{Y}} \log p(y|x; \theta) = \arg \max_{y^* \in \mathcal{Y}} \underbrace{\theta^\top f(x, y^*)}_{\text{linear function}}$$

# POS tagging and prediction



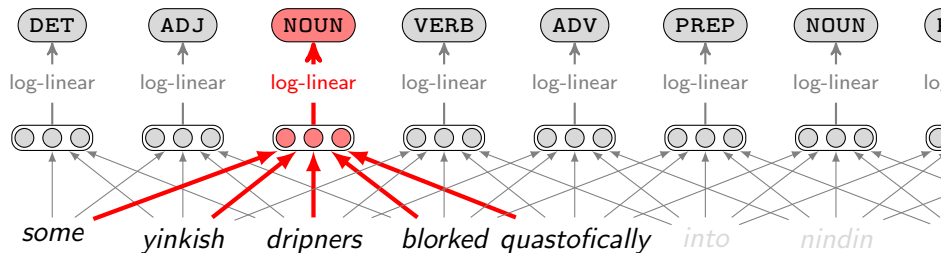
# POS tagging and prediction



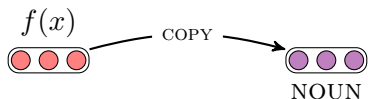
$$f(x) \longrightarrow f(x, y)$$

$$f(x)$$

# POS tagging and prediction

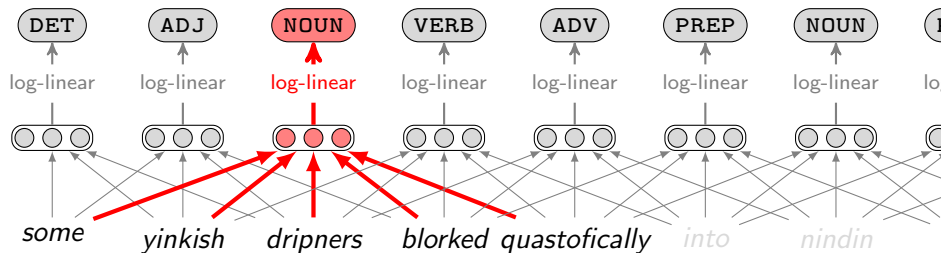


$$f(x) \longrightarrow f(x, y)$$

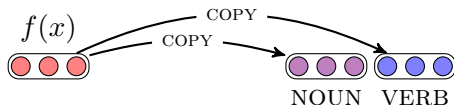




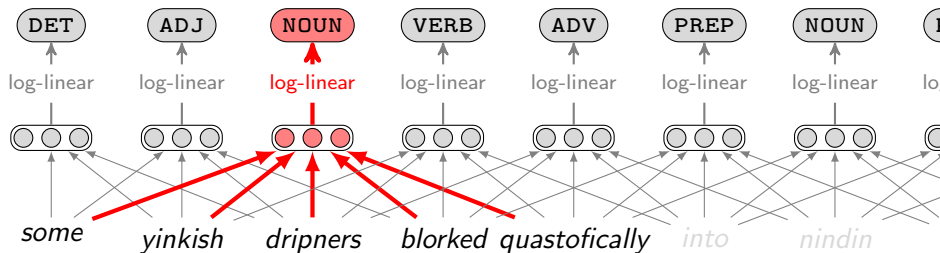
# POS tagging and prediction



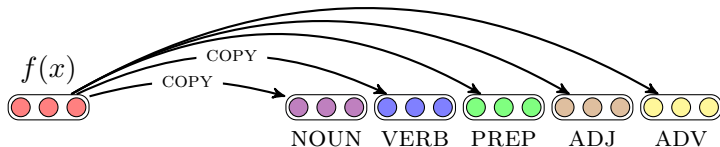
$$f(x) \longrightarrow f(x, y)$$



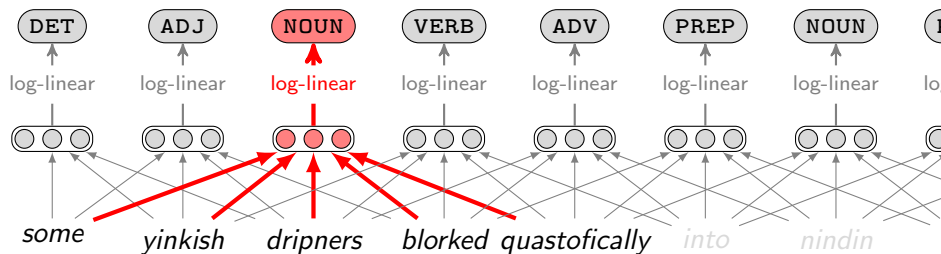
# POS tagging and prediction



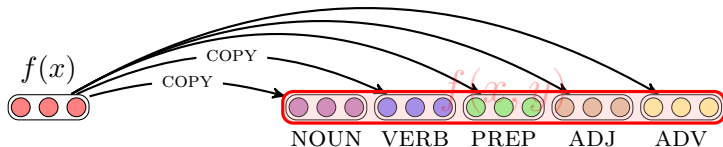
$$f(x) \longrightarrow f(x, y)$$



# POS tagging and prediction

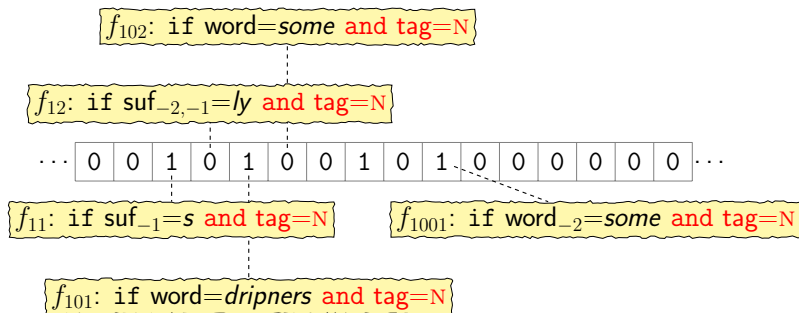


$$f(x) \longrightarrow f(x, y)$$



## About weights

$$p(y|x; \theta) = \frac{\exp(\theta^\top f(x, y))}{\sum_{y' \in \mathcal{Y}} \exp(\theta^\top f(x, y'))}$$



## About weights

$$p(y|x; \theta) = \frac{\exp(\theta^\top f(x, y))}{\sum_{y' \in \mathcal{Y}} \exp(\theta^\top f(x, y'))}$$

... 0 0 1 0 1 0 0 1 0 1 0 0 0 0 0 0 ...

$f_{1001}$ : if word<sub>-2</sub>=some and tag=N

is  $\theta_{1001}$  positive large?  
vote for yes

## About $\exp()$

$$\begin{aligned} p(y|x; \theta) &= \frac{\exp(\theta^\top f(x, y))}{\sum_{y' \in \mathcal{Y}} \exp(\theta^\top f(x, y'))} \\ &= \frac{1}{1 + \frac{\sum_{y' \in \mathcal{Y} \wedge y' \neq y} \exp(\theta^\top f(x, y'))}{\exp(\theta^\top f(x, y))}} \end{aligned}$$

$$\frac{\sum_{y' \in \mathcal{Y} \wedge y' \neq y} \exp(\theta^\top f(x, y'))}{\exp(\theta^\top f(x, y))} \quad \text{vs} \quad \frac{\sum_{y' \in \mathcal{Y} \wedge y' \neq y} \theta^\top f(x, y')}{\theta^\top f(x, y)}$$

# Supervised learning

Assume there is a *good* annotated corpus

$$\left\{ (x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \dots, (x^{(l)}, y^{(l)}) \right\}$$

How can we get a *good* parameter vector?

# Supervised learning

Assume there is a *good* annotated corpus

$$\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \dots, (x^{(l)}, y^{(l)})\}$$

How can we get a *good* parameter vector?

## Maximum-Likelihood Estimation

$$\hat{\theta} = \arg \max L(\theta)$$

where

$$\begin{aligned} L(\theta) &= \sum_{i=1}^l \log p(y^{(i)} | x^{(i)}; \theta) \\ &= \sum_{i=1}^l \left( \theta^\top f(x^{(i)}, y^{(i)}) - \log \sum_{y' \in \mathcal{Y}} \exp(\theta^\top f(x^{(i)}, y')) \right) \end{aligned}$$



# Parameter estimation

$$\arg \max L(\theta) = \arg \max \sum_{i=1}^l \left( \theta^\top f(x^{(i)}, y^{(i)}) - \log \sum_{y' \in \mathcal{Y}} \exp(\theta^\top f(x^{(i)}, y')) \right)$$

## Calculating gradients

$$\begin{aligned} \frac{dL(\theta)}{d\theta_k} &= \sum_{i=1}^m f_k(x^{(i)}, y^{(i)}) - \sum_{i=1}^m \frac{\sum_{y' \in \mathcal{Y}} (\exp(\theta^\top f(x^{(i)}, y')) f_k(x^{(i)}, y'))}{\sum_{y' \in \mathcal{Y}} \exp(\theta^\top f(x^{(i)}, y'))} \\ &= \sum_{i=1}^m f_k(x^{(i)}, y^{(i)}) - \sum_{i=1}^m \sum_{y' \in \mathcal{Y}} f_k(x^{(i)}, y') \frac{\exp(\theta^\top f(x^{(i)}, y'))}{\sum_{y' \in \mathcal{Y}} \exp(\theta^\top f(x^{(i)}, y'))} \end{aligned}$$

# Parameter estimation

$$\arg \max L(\theta) = \arg \max \sum_{i=1}^l \left( \theta^\top f(x^{(i)}, y^{(i)}) - \log \sum_{y' \in \mathcal{Y}} \exp(\theta^\top f(x^{(i)}, y')) \right)$$

## Calculating gradients

$$\begin{aligned} \frac{dL(\theta)}{d\theta_k} &= \sum_{i=1}^m f_k(x^{(i)}, y^{(i)}) - \sum_{i=1}^m \frac{\sum_{y' \in \mathcal{Y}} (\exp(\theta^\top f(x^{(i)}, y')) f_k(x^{(i)}, y'))}{\sum_{y' \in \mathcal{Y}} \exp(\theta^\top f(x^{(i)}, y'))} \\ &= \sum_{i=1}^m f_k(x^{(i)}, y^{(i)}) - \sum_{i=1}^m \sum_{y' \in \mathcal{Y}} f_k(x^{(i)}, y') \frac{\exp(\theta^\top f(x^{(i)}, y'))}{\sum_{y' \in \mathcal{Y}} \exp(\theta^\top f(x^{(i)}, y'))} \end{aligned}$$

$$\frac{dL(\theta)}{d\theta_k} = \underbrace{\sum_{i=1}^m f_k(x^{(i)}, y^{(i)})}_{\text{empirical counts}} - \underbrace{\sum_{i=1}^m \sum_{y' \in \mathcal{Y}} f_k(x^{(i)}, y') p(y'|x^{(i)}; \theta)}_{\text{expected counts}}$$

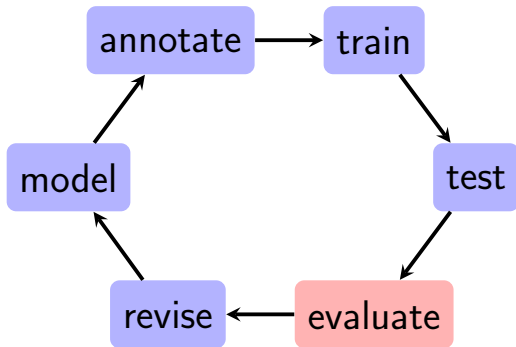
# Gradient Ascent Methods

Maximize  $L(\theta)$  where

$$\frac{dL(\theta)}{d\theta_k} = \underbrace{\sum_{i=1}^m f_k(x^{(i)}, y^{(i)})}_{\text{empirical counts}} - \underbrace{\sum_{i=1}^m \sum_{y' \in \mathcal{Y}} f_k(x^{(i)}, y') p(y'|x^{(i)}; \theta)}_{\text{expected counts}}$$

- 1 **Initialize**  $\theta^{[0]} \leftarrow 0$
- 2 **for**  $t = 1, \dots$
- 3     **calculate**  $\Delta = \frac{dL(\theta^{[t-1]})}{d\theta}$
- 4     **calculate**  $\beta_* = \arg \max_{\beta} L(\theta + \beta \Delta)$  ▷ line search
- 5     **update**  $\theta^{[t]} \leftarrow \theta^{[t-1]} + \beta_* \Delta$

# Evaluation



# Experimental Science

- Experiments are run to test hypotheses
- Hypotheses are tentative theoretical explanations

morphological segmentation facilitates syntactic parsing  
system A outperforms system B on data set C
- Validating hypotheses requires repeated testing

slide from J Nivre's ACL Presidential Address 2017—*Challenges for ACL*

# Intrinsic evaluation

- Creating a test set that contains a sample of test sentences for input, along with the ground truth.
- Quantifying the system's agreement with the ground truth.
- *Training data and test data* Test data must be kept unseen, often 90% training and 10% test data.
- *Baseline*
- *Ceiling Human performance* on the task, where the ceiling is the percentage agreement found between two annotators (inter annotator agreement)
- *Error analysis* Error rates are nearly always unevenly distributed.
- *Replicability* and *reproducibility*

# Inter-annotator agreement

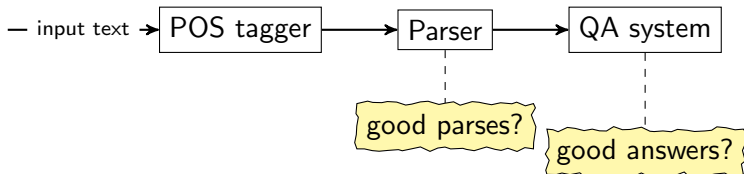
- It is common practice to compare the performance of multiple human annotators.
- If human beings cannot reach substantial agreement about what annotations are correct, it is likely either that the task is too difficult or that it is poorly defined.
- It is generally agreed that human inter-annotator agreement defines the upper limit on our ability to measure automated performance.

Gale et al. (1992) observed that

*our ability to measure performance is largely limited by our ability [to] obtain reliable judgments from human informants*

# Extrinsic evaluation

- Measuring the quality of the system by looking at its impact on the effectiveness of downstream applications.
- Can be applied to compare *heterogeneous* resources.





# Evaluation of POS tagging

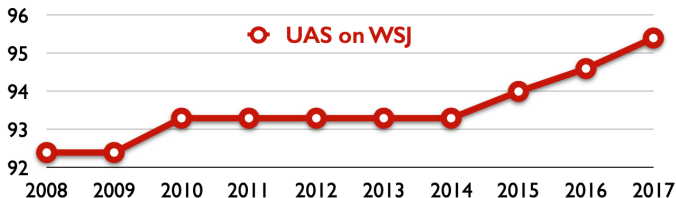
- Tested against held-out data: percentage of correct tags
- One tag per word (some systems give multiple tags when uncertain)
- Over 95% for English on normal corpora (but note punctuation is unambiguous)
- Performance plateau about 97% on most commonly used test set for English
- Baseline of taking the most common tag gives 90% accuracy
- Different tagsets give slightly different results: utility of tag to end users vs predictive power

## Benchmarking and “fair” comparisons – fast science

- Test corpora have to be representative of the actual application
- POS tagging and similar techniques are not always very robust to differences in domain
- Balanced corpora may be better, but still don't cover all text types
- Communication aids: extreme difficulty in obtaining data, text corpora don't give good prediction for real data

data-driven 😊 vs data set-driven ☹️

# Good Science



“Measurement as a virtue in itself”

“Lots of numbers with very small differences”

“What are the research questions?”

slide from J Nivre’s ACL Presidential Address 2017—*Challenges for ACL*

# Readings

- Ann's note
- M Collins' note

`www.cs.columbia.edu/~mcollins/loglinear.pdf`