Overview of Natural Language Processing Part II & ACS L90

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

Weiwei Sun

Department of Computer Science and Technology
University of Cambridge

Michaelmas 2023/24

Is FST suitable for the following phenomena?

Non-concatenative morphology, e.g. duplication in Chinese:

高兴	→	高高兴兴
happy		高兴高兴

$a^n b^n$??

Some yinkish dripners blorked quastofically into the nindin with the pidibs

words have classes

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

Labeling Words

Fish fish fish.

Fish fish fish.

fish

noun

US /fɪʃ/ **UK** /fɪʃ/

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

Lew Robertson/Photolibrary
/GettyImages

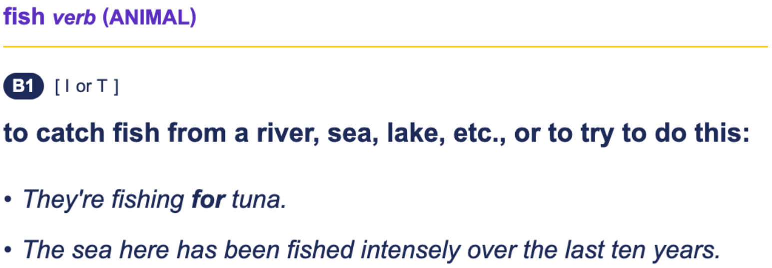
A1 [C or U]

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

Fish fish fish.



dictionary.cambridge.org/us/dictionary/english/fish

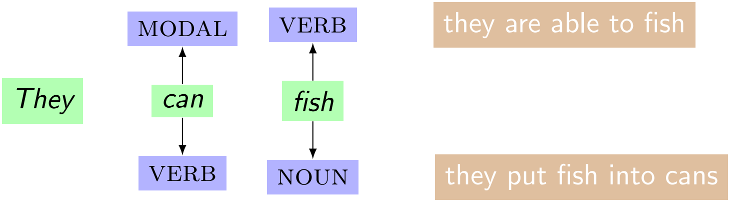
Part-of-speech tagging is useful

Fish/NOUN fish/VERB fish/NOUN

from FINDING NEMO MOVIE (2013)

photo: www.avforums.com/reviews/finding-nemo-movie-review.6237

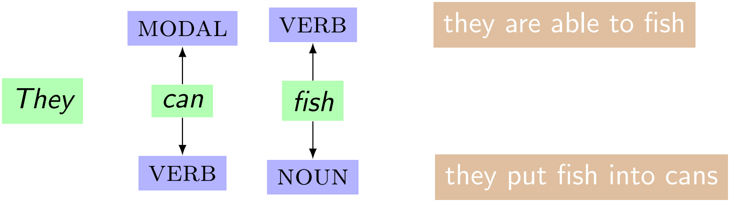
Global v local ambiguity



Ambiguity

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

Global v local ambiguity



Ambiguity

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

application-independent tags;
linguistic knowledge involved

from Ann Copestake's course

Information extraction (1)

Book a flight

- Leave London on 1st Dec 2020
- Arrive in London on 1st Dec 2020

FROM		
TO		
TIME		

Information extraction (1)

Book a flight

- Leave/O London/B-FROM on/O 1st/B-TIME Dec/I-TIME 2020/E-TIME
- Arrive/O in/O London/B-TO on/O 1st/B-TIME Dec/I-TIME 2020/E-TIME

FROM	London	
TO		London
TIME	1st Dec 2020	1st Dec 2020

Chunking

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

Information extraction (1)

Book a flight

- Leave/O London/B-FROM on/O 1st/B-TIME Dec/I-TIME 2020/E-TIME
- Arrive/O in/O London/B-TO on/O 1st/B-TIME Dec/I-TIME 2020/E-TIME

FROM	London	
TO		London
TIME	1st Dec 2020	1st 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.

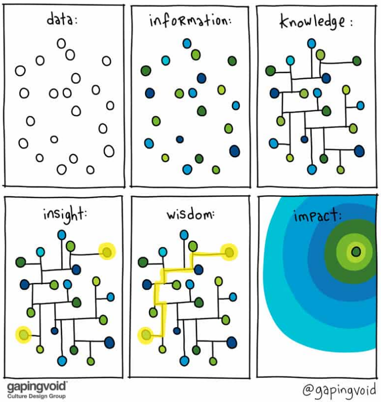
application-dependent tags; world knowledge involved

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

Data, Information, Knowledge, Wisdom

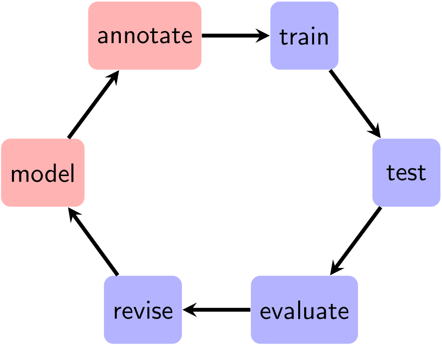


Last lecture

- Knowledge-driven approach: Finite-state machines
- Data-driven approach: Byte-pair encoding
  - Unsupervise learning, representation learning

Corpora

Annotations in NLP



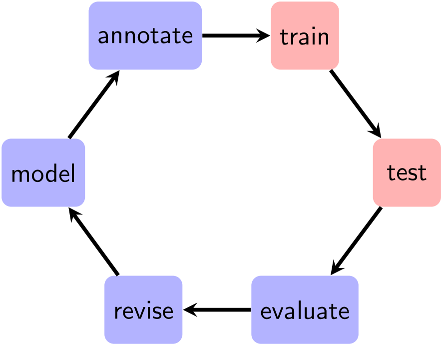
MATTER: the annotation development cycle

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



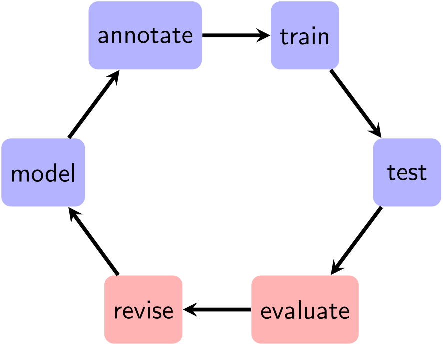
MATTER: the annotation development cycle

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)

Be careful

Data may be very *difficult to acquire*

- first language acquisition ▷ takes years to collect
- historical linguistics ▷ no longer exist
- brain activities ▷ wonderful machines, e.g. fMRI
- dolphin language ▷ ...

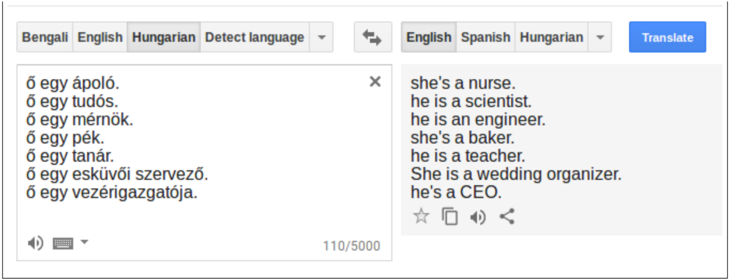
Data may be extremely *big*

- e.g. data from twitter

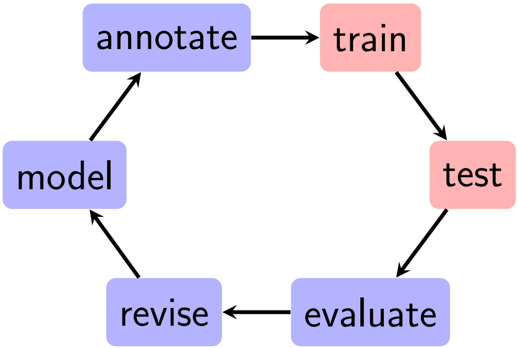
Data may be *private*

- the Cambridge Analytica/Facebook scandal

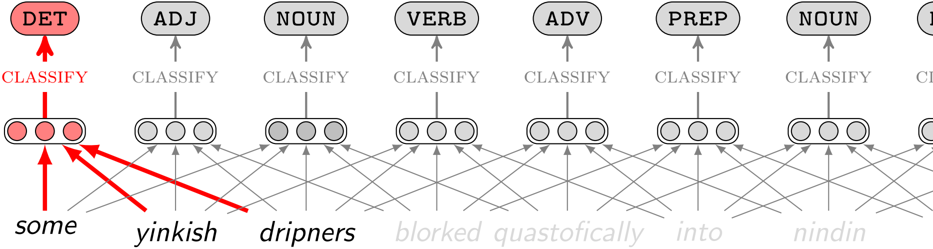
Data may be *biased* Prates et al. (2019) `https://arxiv.org/pdf/1809.02208.pdf`



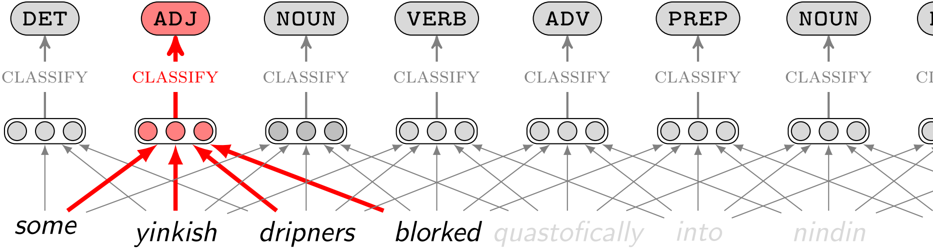
Log-Linear Models



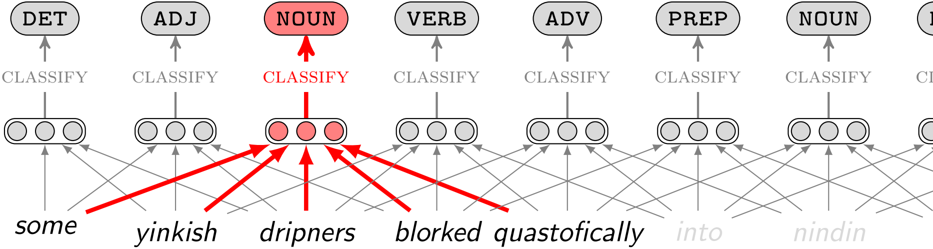
POS tagging and prediction



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$_{-3,-2}$=*er*
suf$_{-1}$=*s*

morphological features

Aspects of POS tagging

word$_{i-2}$=*some*
word$_{i-1}$=*yinkish*

word=*dripners*

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

suf$_{-3,-2}$=*er*
suf$_{-1}$=*s*

POS can be defined distributionally

Aspects of POS tagging

$\text{word}_{i-2}=$*some*
$\text{word}_{i-1}=$*yinkish*

$\text{word}_{i+2}=$*quastofically*
$\text{word}_{i+1}=$*blorked*

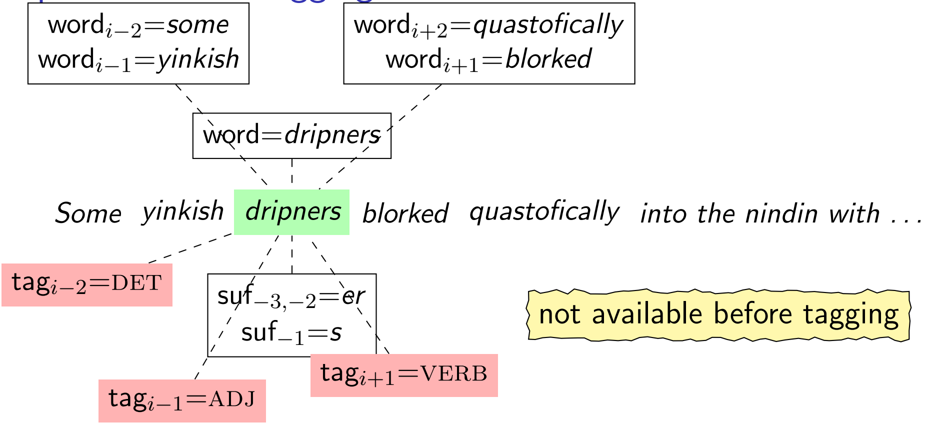
word=*dripners*

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

$\text{suf}_{-3,-2}=$*er*
$\text{suf}_{-1}=$*s*

POS can be defined distributionally

Aspects of POS tagging



Aspects of POS tagging

word$_{i-2}$=*some*
word$_{i-1}$=*yinkish*

word$_{i+2}$=*quastofically*
word$_{i+1}$=*blorked*

word=*dripners*

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

suf$_{-3,-2}$=*er*
suf$_{-1}$=*s*

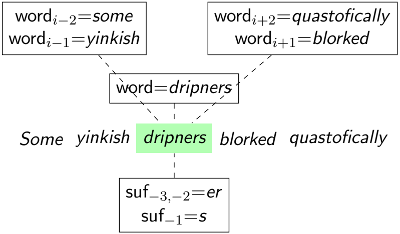
The task: model the distribution

$$p(t_i|w_1,\ldots,w_n) \Rightarrow p(t_i|\textsc{DeriveFeature}(w_{i-w}, w_{i-w+1}...w_{i+w}))$$

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

1-of-K encoding

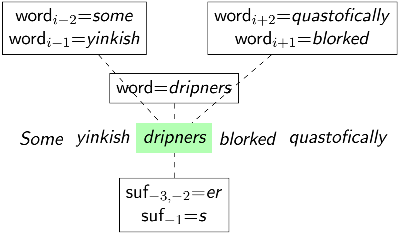
k is the index of current POS label;
D is the dimension of $f(x)$.

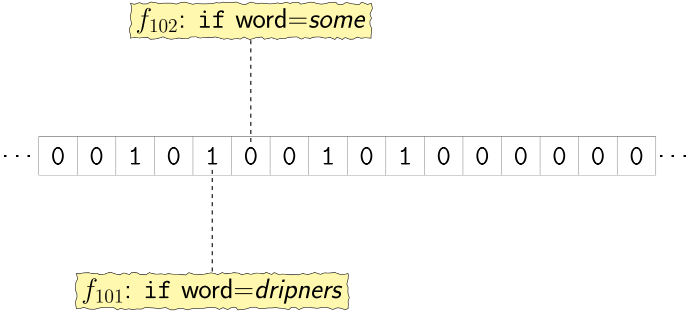


$\cdots$	0	0	1	0	1	0	0	1	0	1	0	0	0	0	0	0	$\cdots$

1-of-K encoding

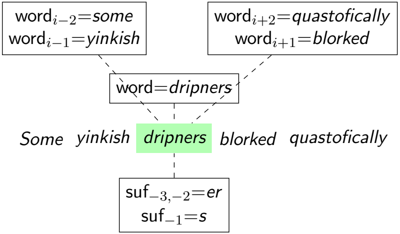
k is the index of current POS label;
D is the dimension of $f(x)$.

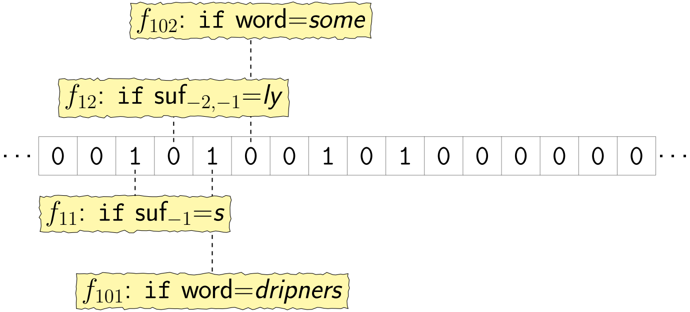




1-of-K encoding

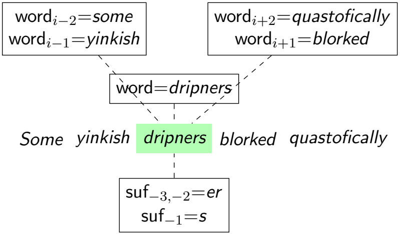
k is the index of current POS label;
D is the dimension of $f(x)$.

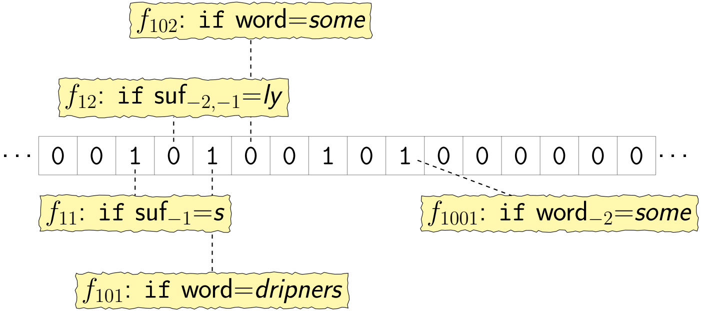




1-of-K encoding

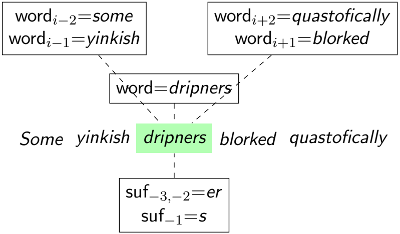
k is the index of current POS label;
D is the dimension of $f(x)$.

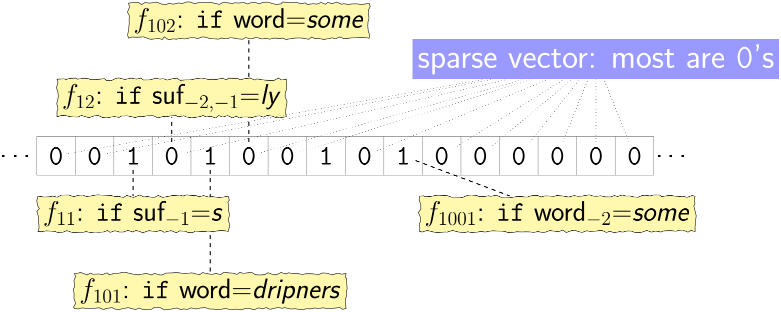




1-of-K encoding

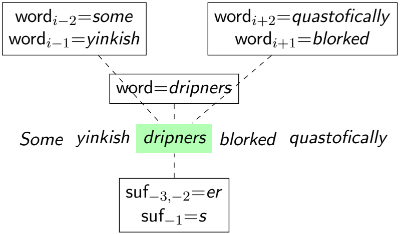
k is the index of current POS label;
D is the dimension of $f(x)$.

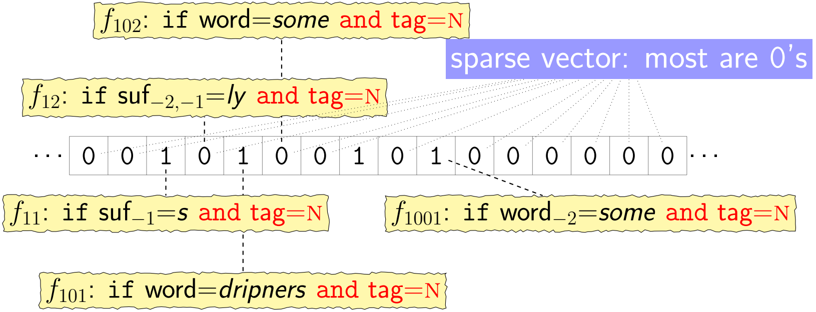




1-of-K encoding

k is the index of current POS label;
D is the dimension of $f(x)$.





1-of-K encoding

k is the index of current POS label;
D is the dimension of $f(x)$.

word$_{i-2}$=*some*
word$_{i-1}$=*yinkish*

word$_{i+2}$=*quastofically*
word$_{i+1}$=*blorked*

word=*dripners*

Some *yinkish* *dripners* *blorked* *quastofically*

suf$_{-3,-2}$=*er*
suf$_{-1}$=*s*

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

sparse vector: most are 0's

f_{12+kD}: if suf$_{-2,-1}$=*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$_{-1}$=*s* and tag=N

$f_{1001+kD}$: if word$_{-2}$=*some* and tag=N

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

$$f(x, y)$$
$$x = \langle w_1, \ldots, w_n, i \rangle$$
$$y = t_i$$

Log-linear models (multinomial logistic regression)

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

Log-linear models (multinomial logistic regression)

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 (multinomial logistic regression)

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

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

Log-linear models (multinomial logistic regression)

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

We define

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

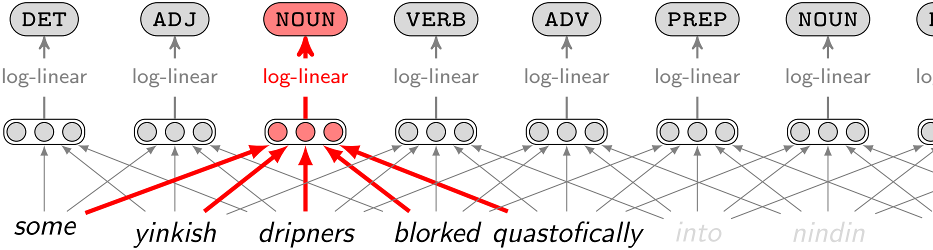
Why the name

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

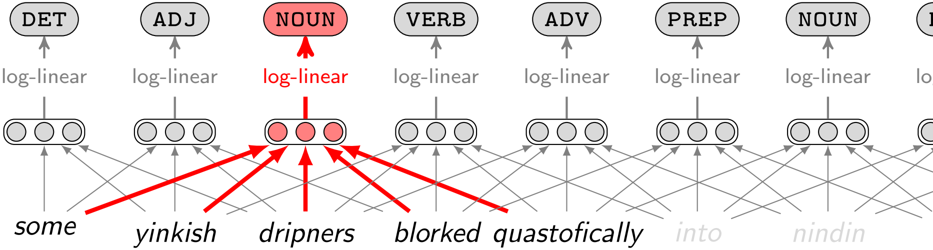
Prediction/ranking/scoring

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

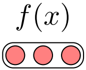
POS tagging and prediction



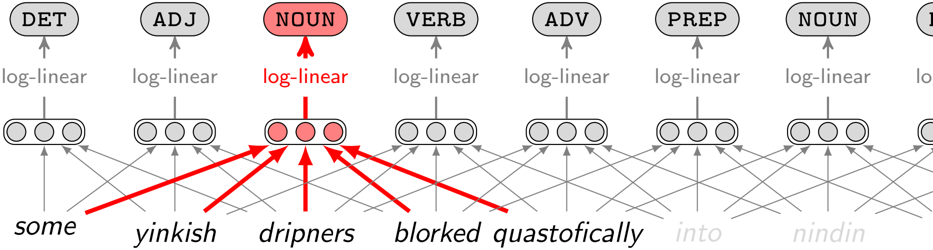
POS tagging and prediction



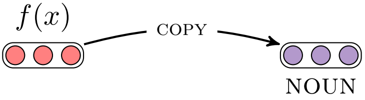
$f(x) \longrightarrow f(x, y)$



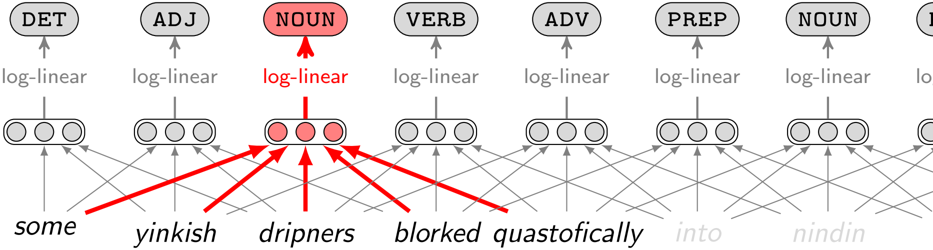
POS tagging and prediction



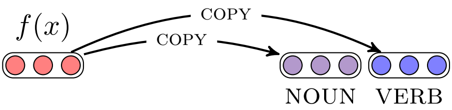
$$f(x) \longrightarrow f(x, y)$$



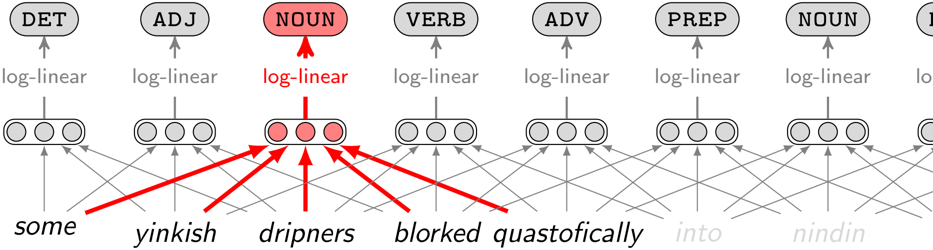
POS tagging and prediction



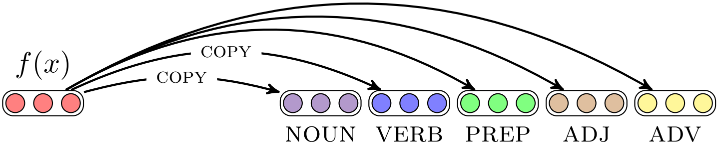
$f(x) \longrightarrow f(x, y)$



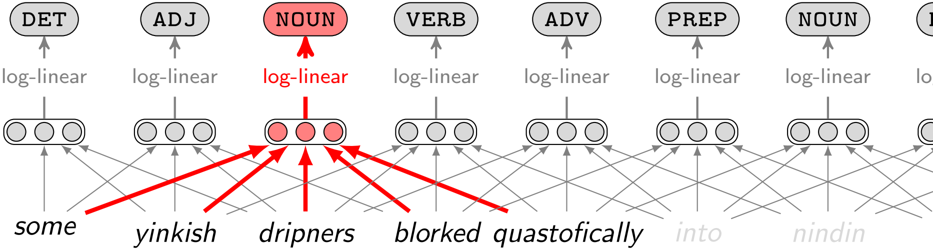
POS tagging and prediction



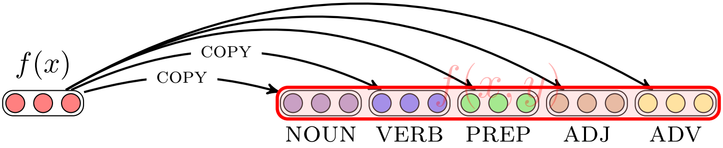
$$f(x) \longrightarrow f(x,y)$$



POS tagging and prediction



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



About weights

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

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

f_{12}: if suf$_{-2,-1}$=*ly* and tag=N

$\cdots$ | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | $\cdots$

f_{11}: if suf$_{-1}$=*s* and tag=N

f_{1001}: if word$_{-2}$=*some* and tag=N

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

About weights

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

$\cdots$ | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | $\cdots$

f_{1001}: `if` word$_{-2}$=*some* and tag=N

is θ_{1001} positively large?
vote for yes

Supervised learning

Assume there is a *good* annotated corpus

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

How can we get a *good* parameter vector?

Supervised learning

Assume there is a *good* annotated corpus

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

How can we get a *good* parameter vector?

Maximum-Likelihood Estimation

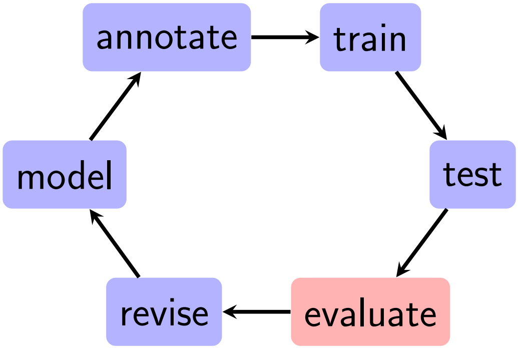
$$\hat{\theta} = \arg\max L(\theta)$$

where

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

To be continued in Lecture 9

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, development and test data* Training data is used for parameter estimation. Development data is used for tuning some hyperparameters. Test data must be kept unseen, e.g. 80% training, 10% devel and 10% test data.
- *Baseline*
- *Ceiling Human performance* on the task, often with 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

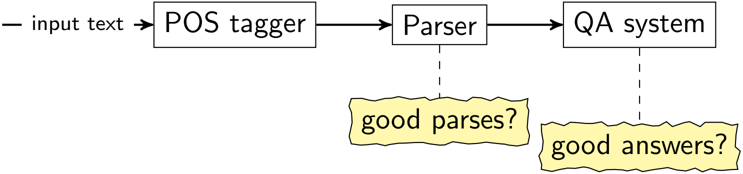
- It is common practice to compare the performance of multiple human annotators.
- If human beings cannot reach substantial agreement about what annotations are correct, it is likely either that the task is too difficult or that it is poorly defined.
- It is generally agreed that human inter-annotator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

Gale et al. (1992) observed that

> *our ability to measure performance is largely limited by our ability [to] obtain reliable judgments fromhuman 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.



Benchmarking and “fair” comparisons – fast science

- Test corpora have to be representative of the actual application

data-driven ☺ vs data set–driven ☹

based on Ann Copestake’s slides

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

data-driven ☺ vs data set–driven ☹

based on Ann Copestake's slides

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

data-driven ☺ vs data set–driven ☹

based on Ann Copestake’s slides

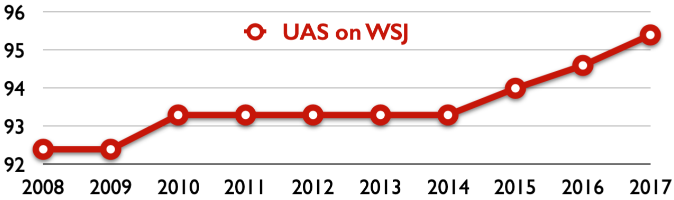
Benchmarking and "fair" comparisons – fast science

- 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 ☹

based on Ann Copestake's slides

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

Required

- Chapter 5. Logistic Regression. *Speech and Language Processing*. D Jurafsky and J Martin. `https://web.stanford.edu/~jurafsky/slp3/5.pdf`