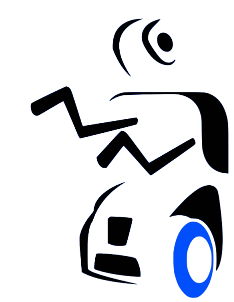
Semi-Supervised Learning of Sequence Models via Method of Moments

EMNLP - Empirical Methods for Natural Language Processing
November 1-6, 2016 Austin, Texas



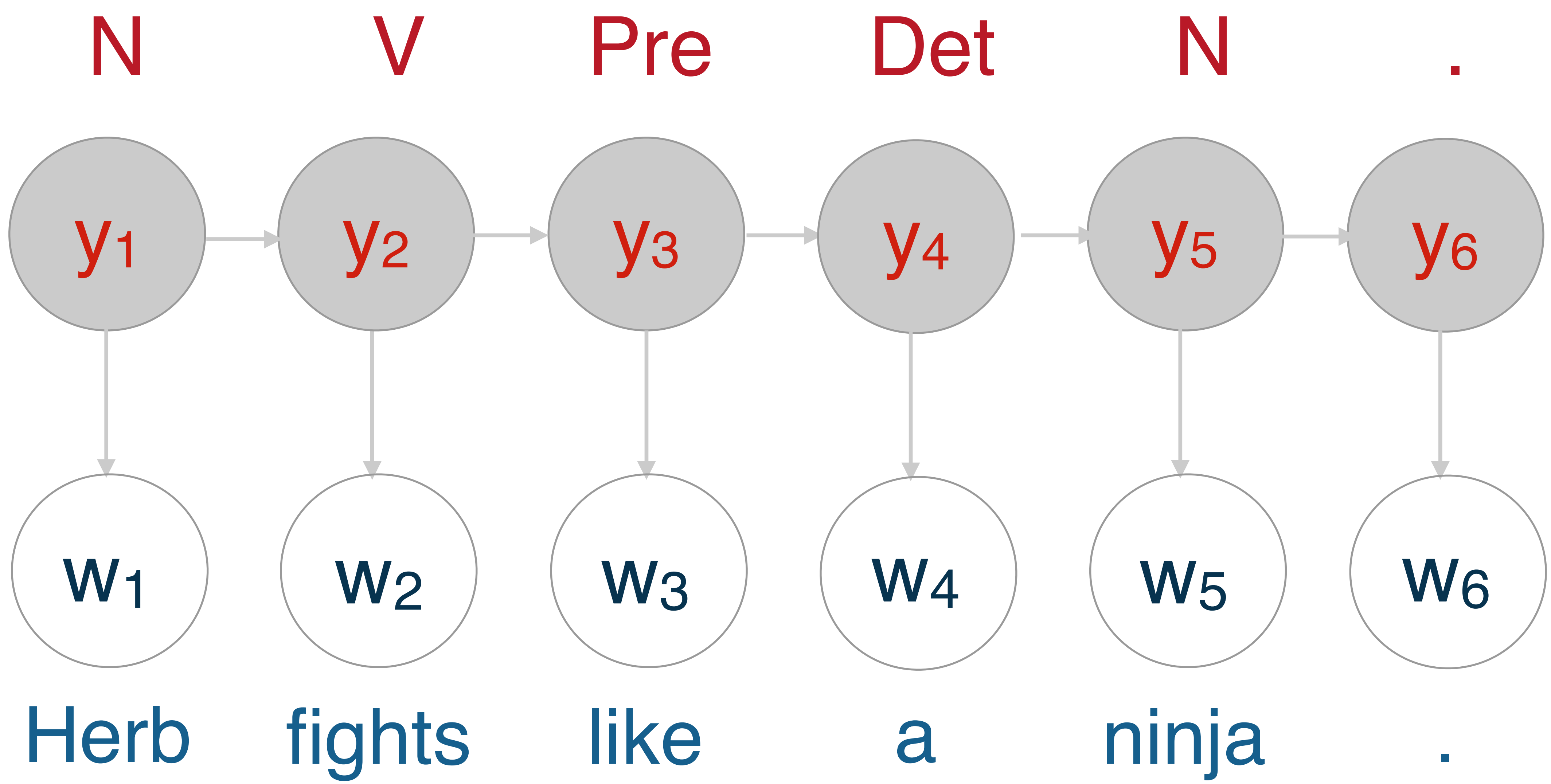
Zita Marinho
IST, University of Lisbon
Robotics Institute, CMU
zmarinho@cmu.edu

André F. T. Martins
IT, IST, University of Lisbon
Unbabel
andre.martins@unbabel.com

Shay B. Cohen
School of Informatics
University of Edinburgh
scohen@inf.ed.ac.uk

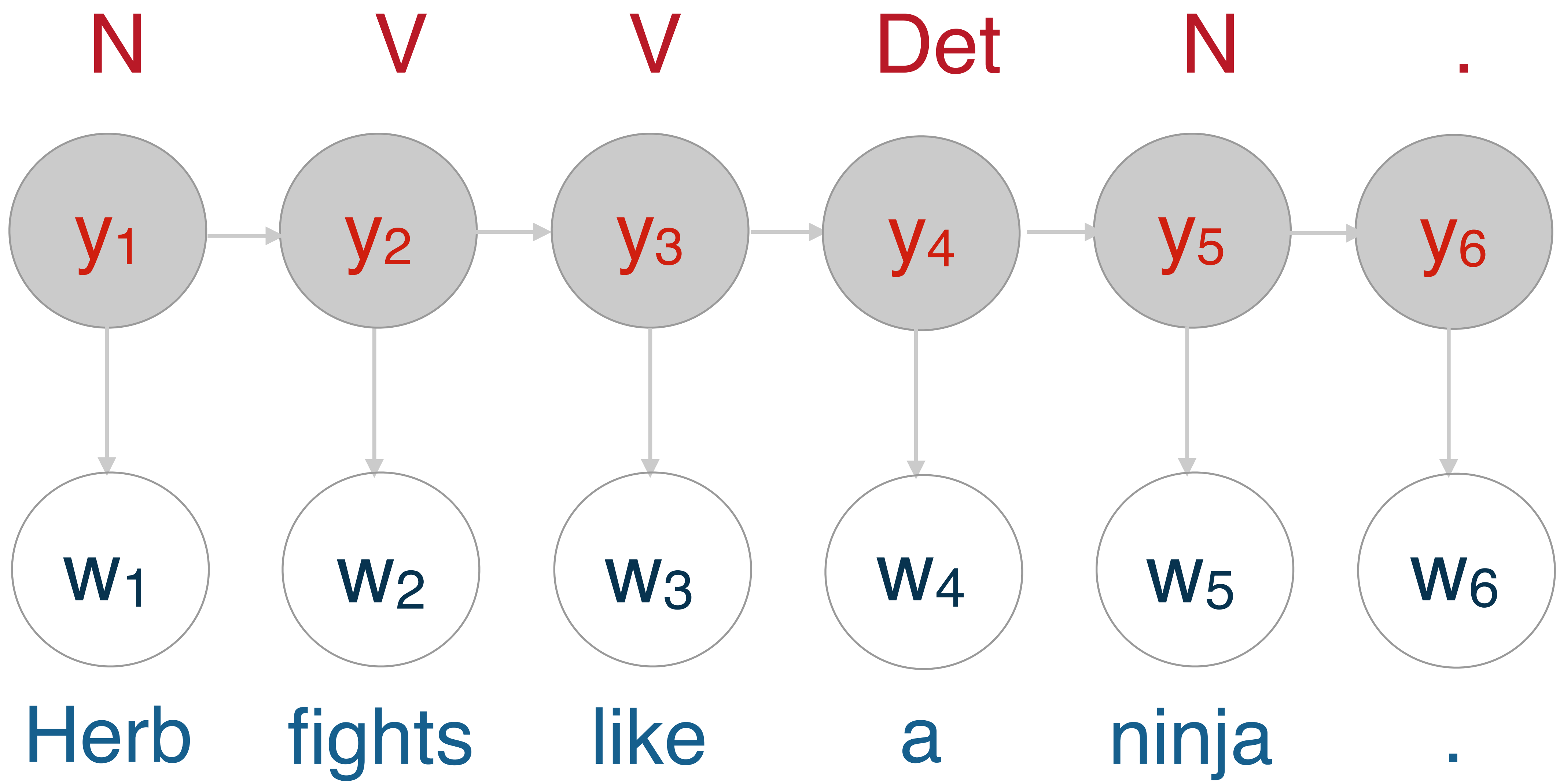
Noah A. Smith
Computer Science & Eng.
University of Washington
nasmith@cs.washington.edu

Sequence Labeling



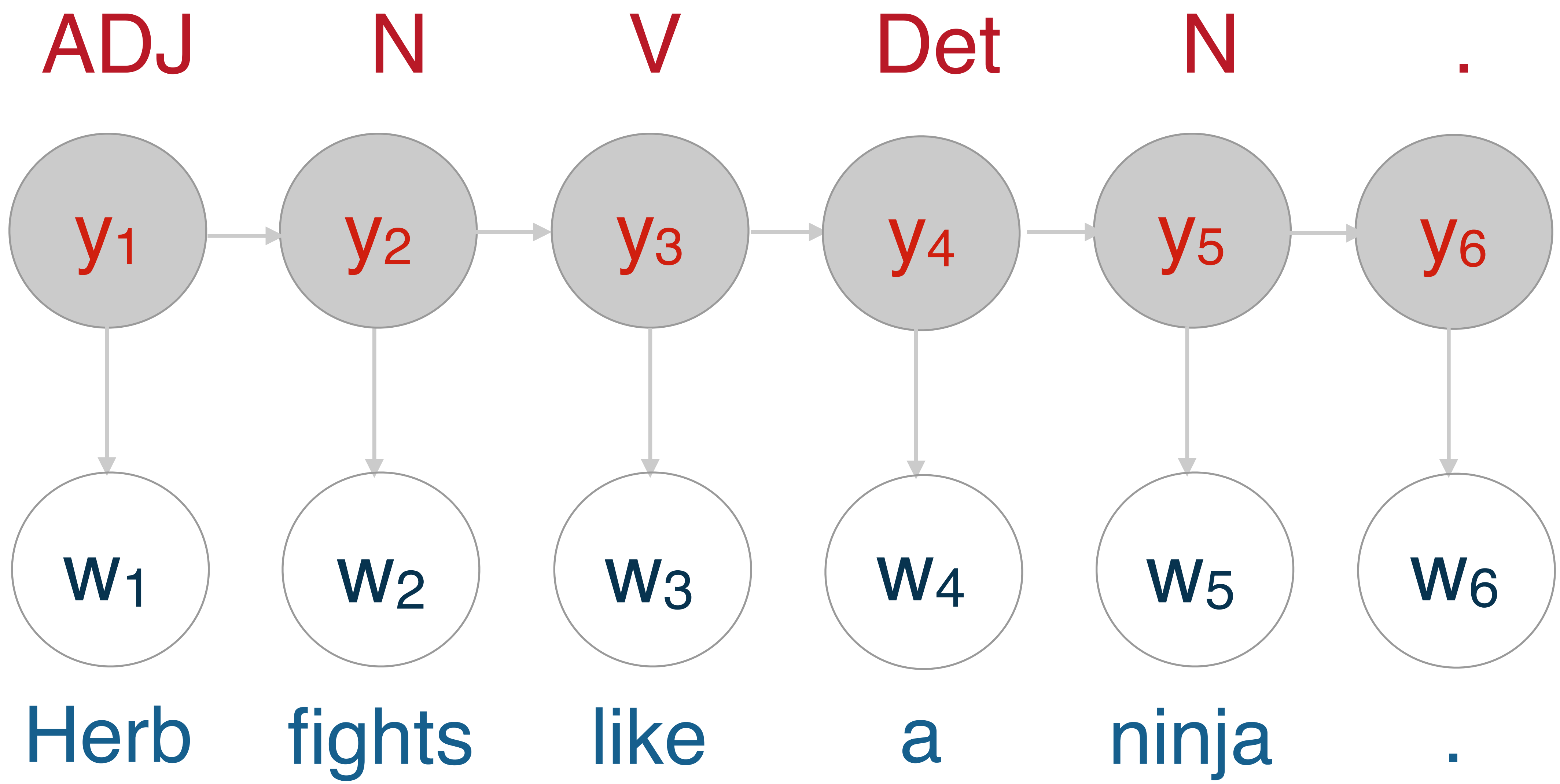
observed data $\{w_1, w_2, w_3, \ldots, w_6\}$
labels $\{y_1, y_2, y_3, \ldots, y_6\}$

Sequence Labeling



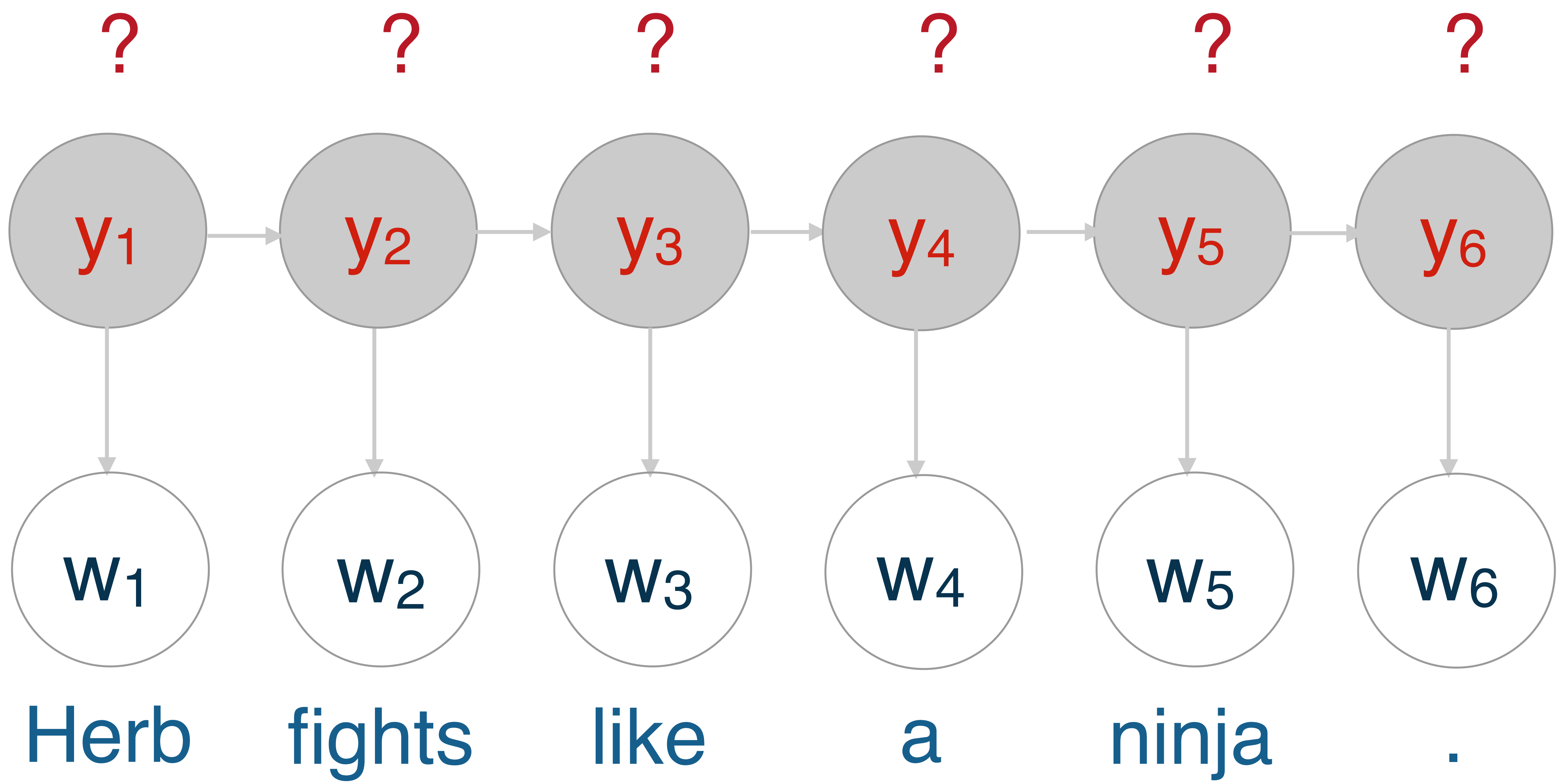
observed data $\{w_1, w_2, w_3, \ldots, w_6\}$
labels $\{y_1, y_2, y_3, \ldots, y_6\}$

Sequence Labeling



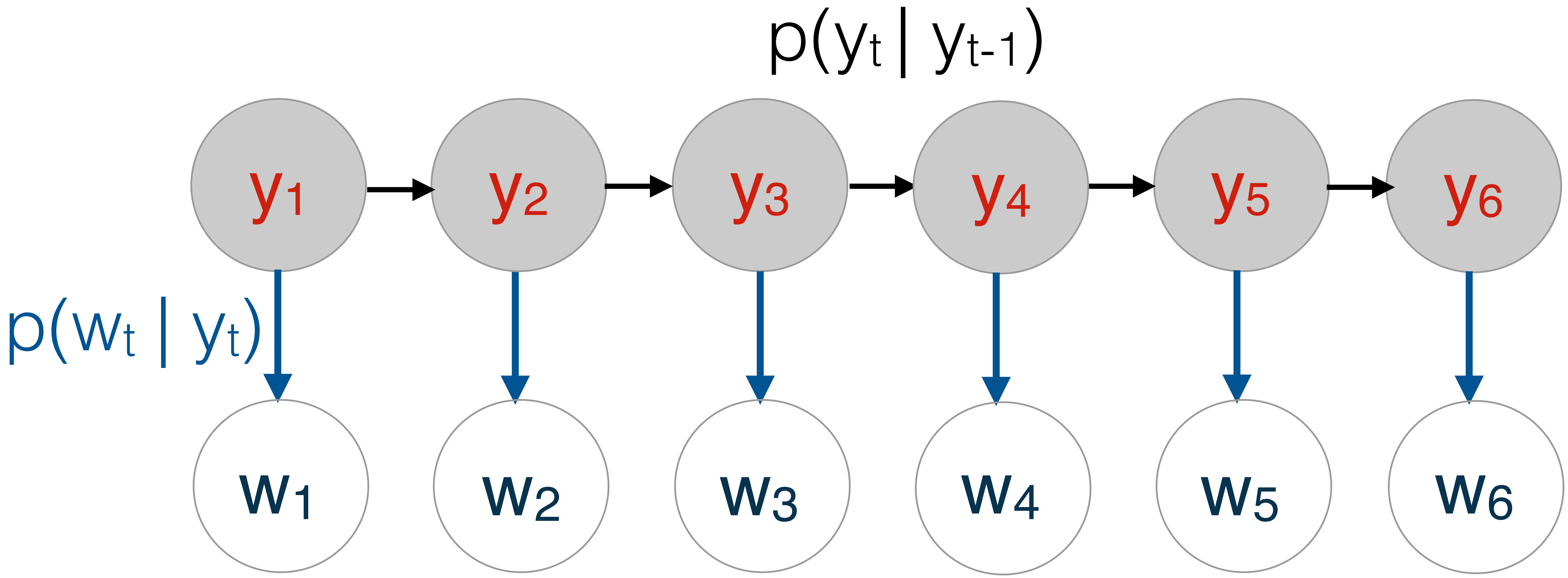
observed data $\{w_1, w_2, w_3, \ldots, w_6\}$
labels $\{y_1, y_2, y_3, \ldots, y_6\}$

K^6 possible assignments



observed data $\{w_1, w_2, w_3, \ldots, w_6\}$
labels $\{y_1, y_2, y_3, \ldots, y_6\}$

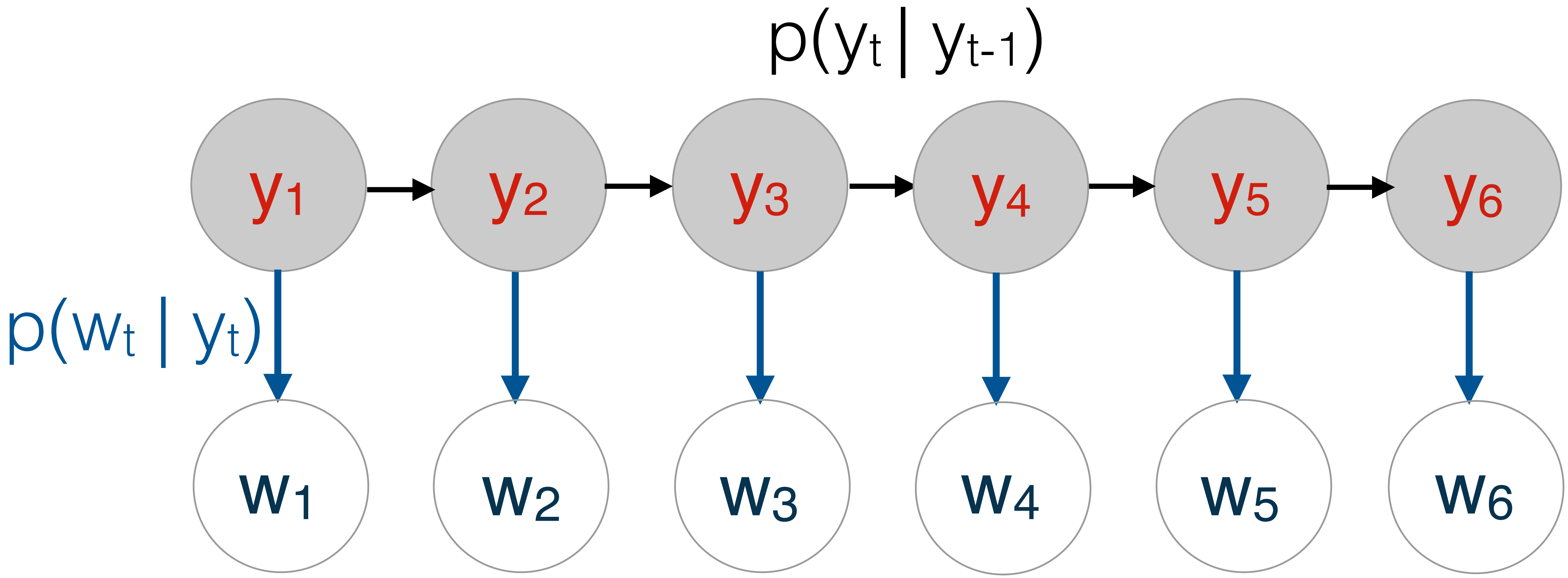
Learn parameters?



- supervised learning
- unsupervised/semi-supervised (this talk)

Hidden Markov Model

Learn parameters?



- supervised learning
- unsupervised/semi-supervised (this talk)
- model can be extended to include features

Berg-Kirkpatrick, et al, Painless unsupervised learning with features. NAACL HLT, 2010.

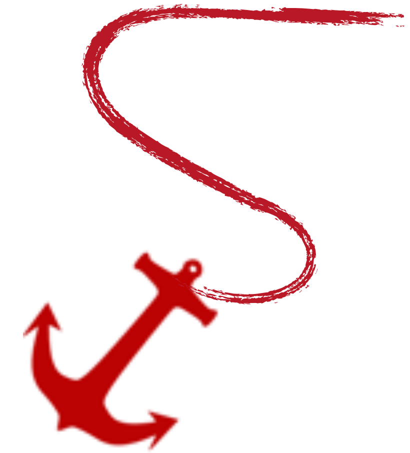
Maximum Likelihood estimation (MLE)		Method of Moments estimation (MoM)
• exact inference is hard	→	**computationally efficient**
• EM sensitive to local optima (depends on initialization)	→	**no local optima**
• EM expensive in large datasets (several inference passes)	→	**one pass over data**

Hidden Markov Model

	via Maximum Likelihood Estimation		via Method of Moments	
	MLE HMM	MLE feature HMM	MoM HMM	MoM feature HMM
semi-supervised learning	✓	✓	?	?
unsupervised learning	✓	✓	✓	?

Shay B. Cohen, Karl Stratos, Michael Collins, Dean P. Foster and Lyle Ungar, *Spectral Learning of Latent-Variable PCFGs: Algorithms and Sample Complexity*, JMLR 2014

Arora et al., *A Practical Algorithm for Topic Modeling with Provable Guarantees,* ICML 2013



Outline

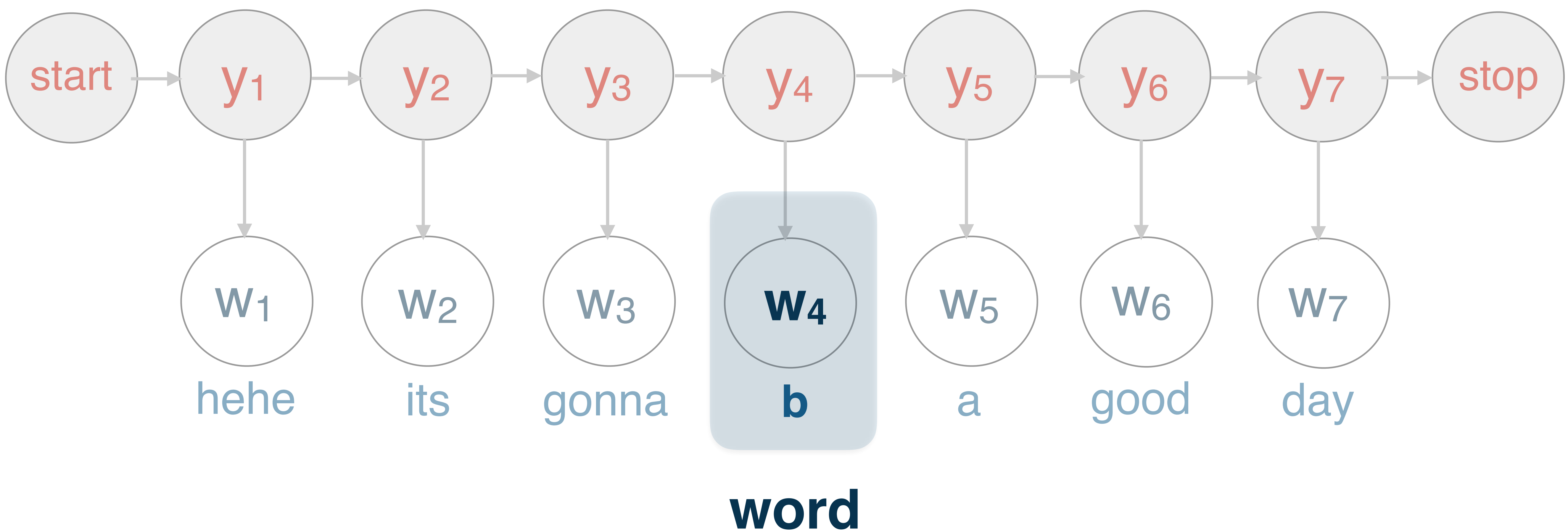
1. Learn HMM models via MoM
2. Solve a QP
3. Extend to feature-based model
4. Experiments

Key insight:

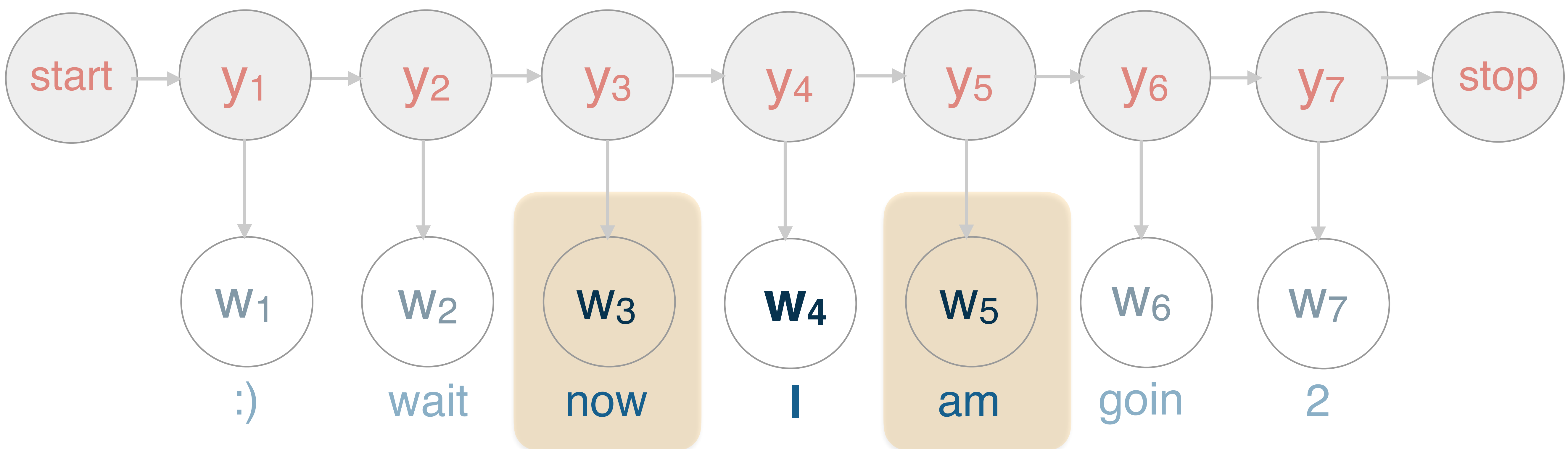
1. Conditional Independence:
   infer label by looking at context

2. Anchor Trick:
   learn a proxy for labels with anchors

1. Conditional Independence

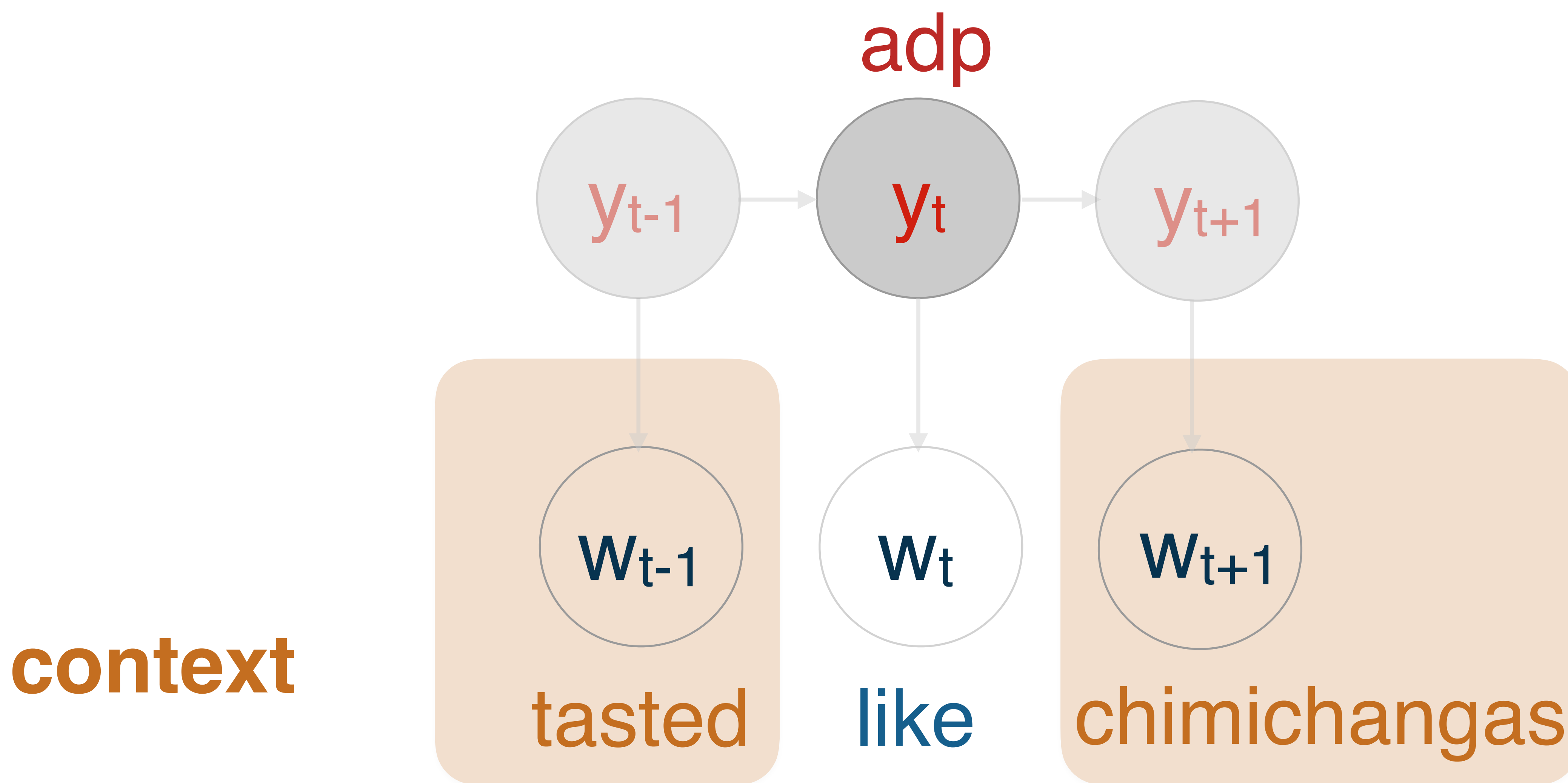


1. Conditional Independence

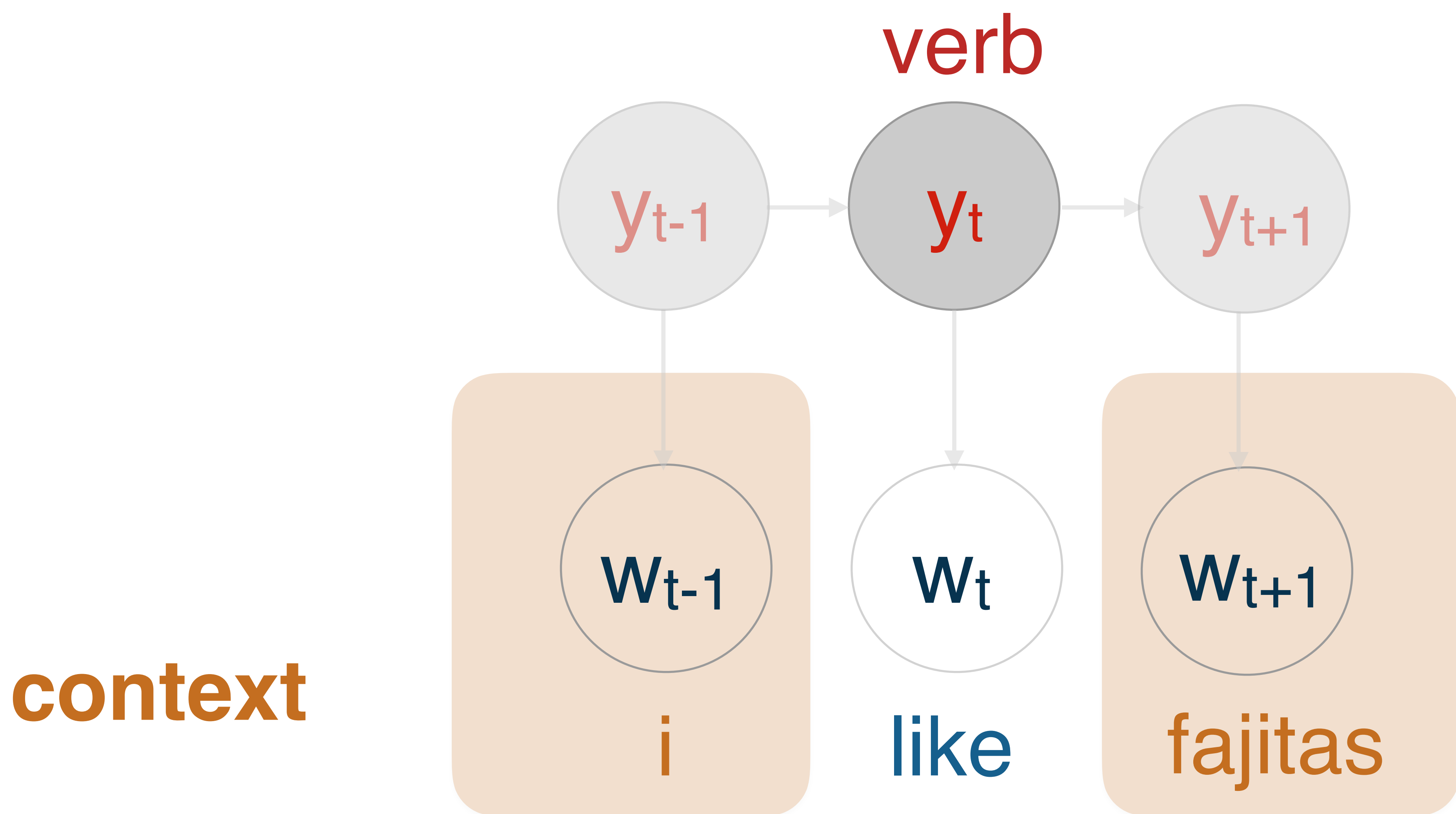


context = { w_{-1} , w_{+1} }

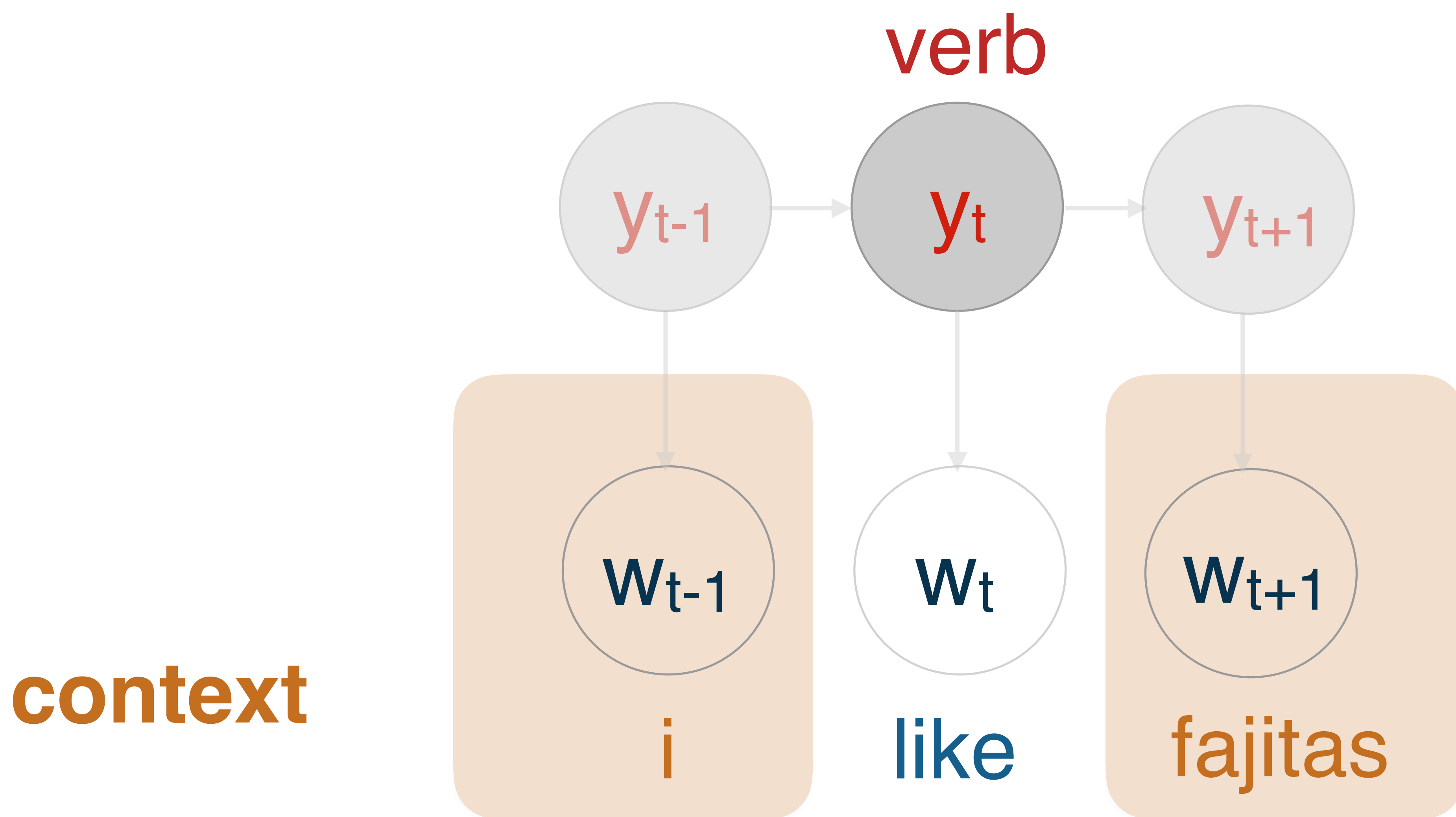
1. Conditional Independence



1. Conditional Independence



1. Conditional Independence

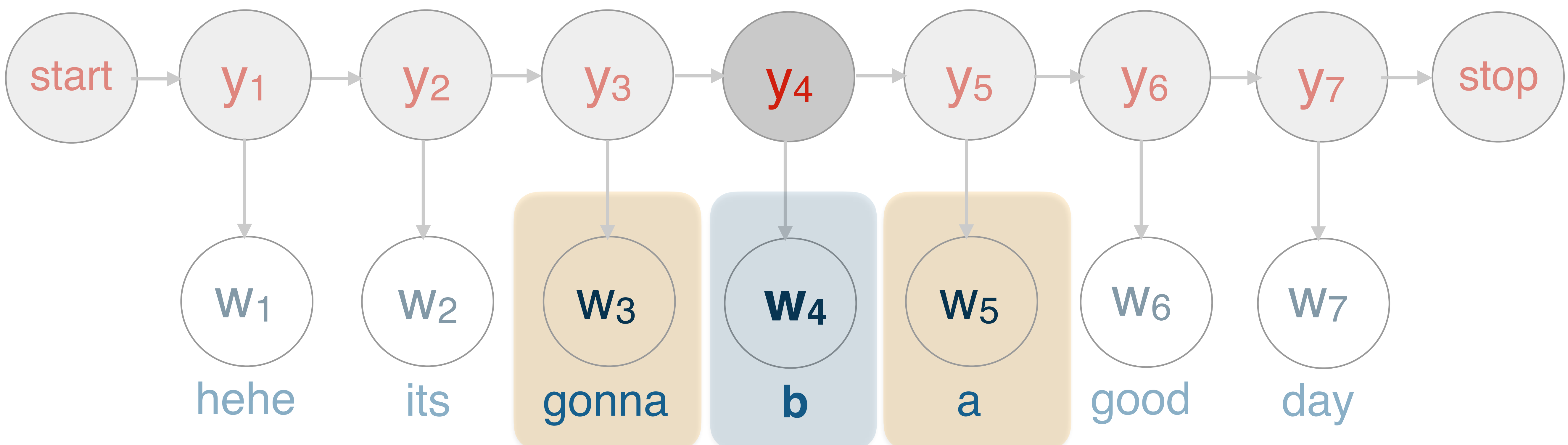


"You shall know a word by the company it keeps."

Firth, 1957

1. Conditional Independence

word $\perp$ context | label



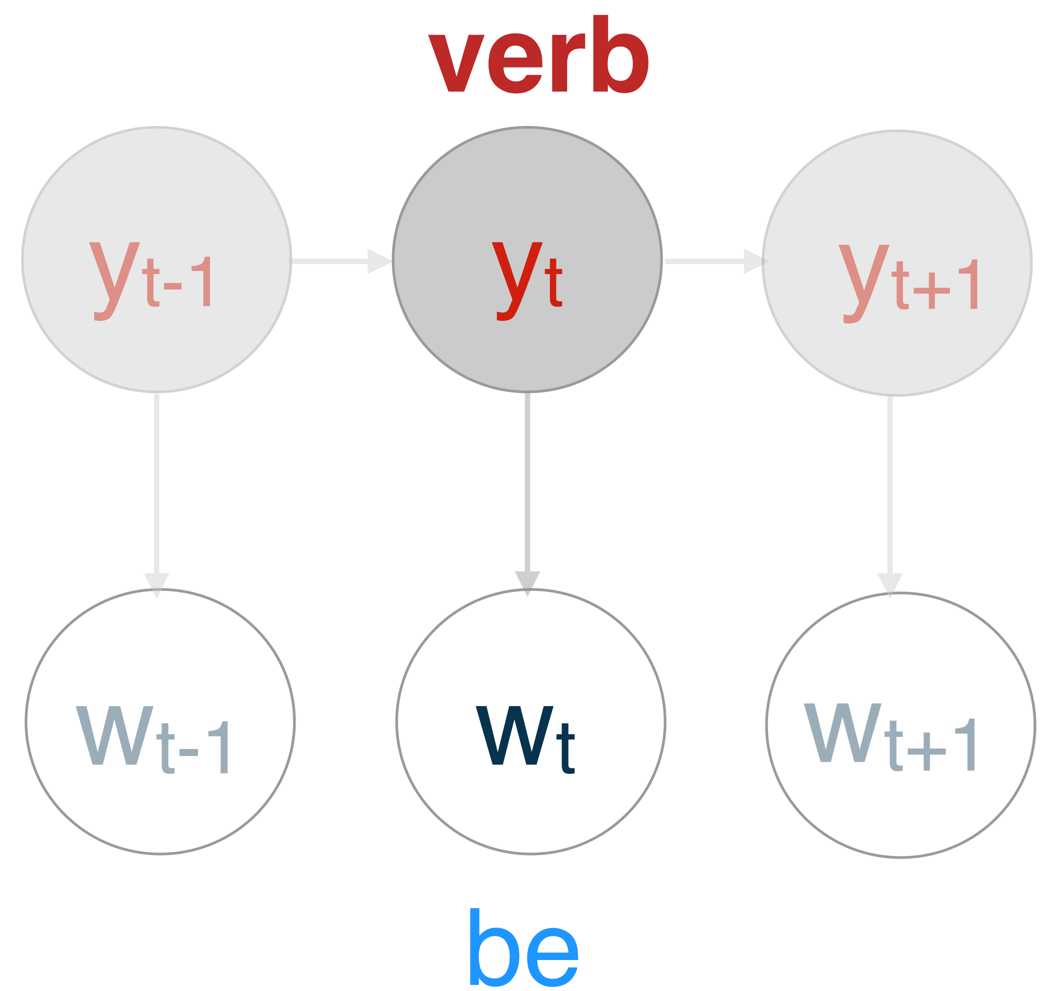
word
context

2. Anchor Trick

all instances of **be** = verb

$p($ verb | be $) = 1$

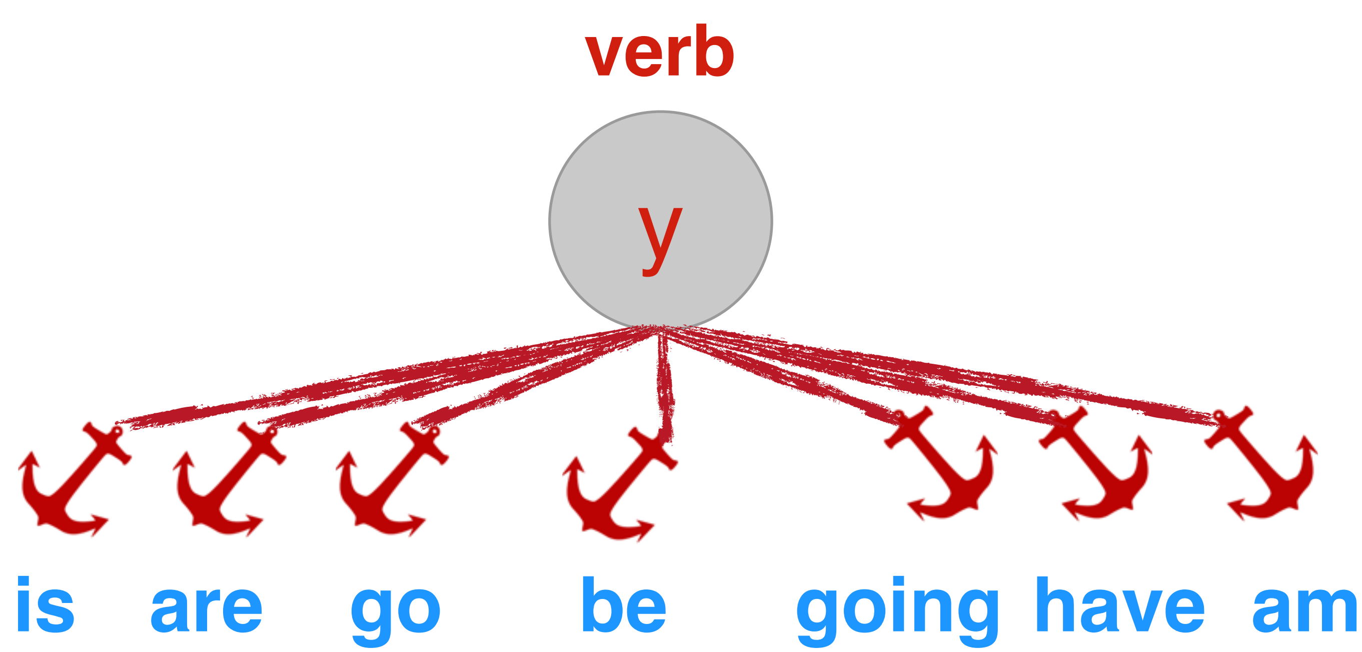
$p($ label $\neq$ verb | be $) = 0$



Arora et al., *A Practical Algorithm for Topic Modeling with Provable Guarantees,* ICML 2013

More anchors per label

verb = **b, be, are, is, am, have, going**



more than 1 anchor word → less biased context estimates

How to find **anchors**?

- small labeled corpus
- small lexicon

Austin airport playground	**noun**
am,be,is,are go, make,made become	**verb**
he,it,she	**pron**
so,on,of	**adp**

Method of moments

unlabeled

co-occurrences in data

w_{t-1} w_t w_{t+1} w_{t+2}

context

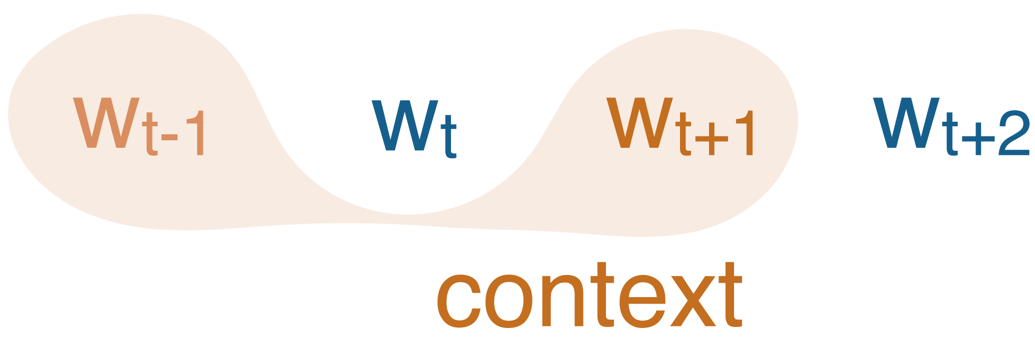
Andrew fights **like** Jet Li.

Ann sings **like** me.

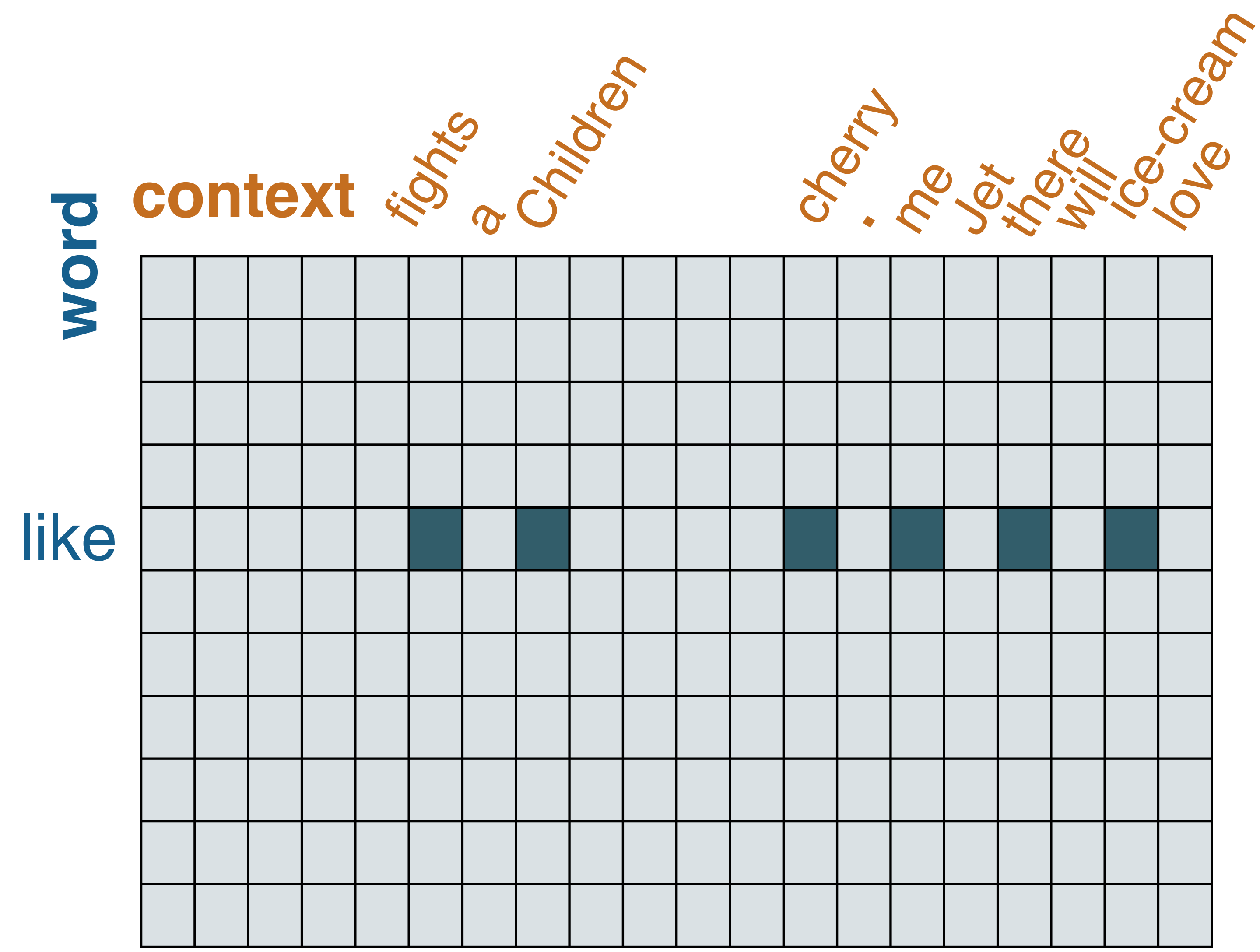
eat Fruit **like** cherry.

Children **like** ice-cream.

Method of moments



Q *p*(context | word)



Andrew fights **like** Jet Li.

Ann sings **like** me.

eat Fruit **like** cherry.

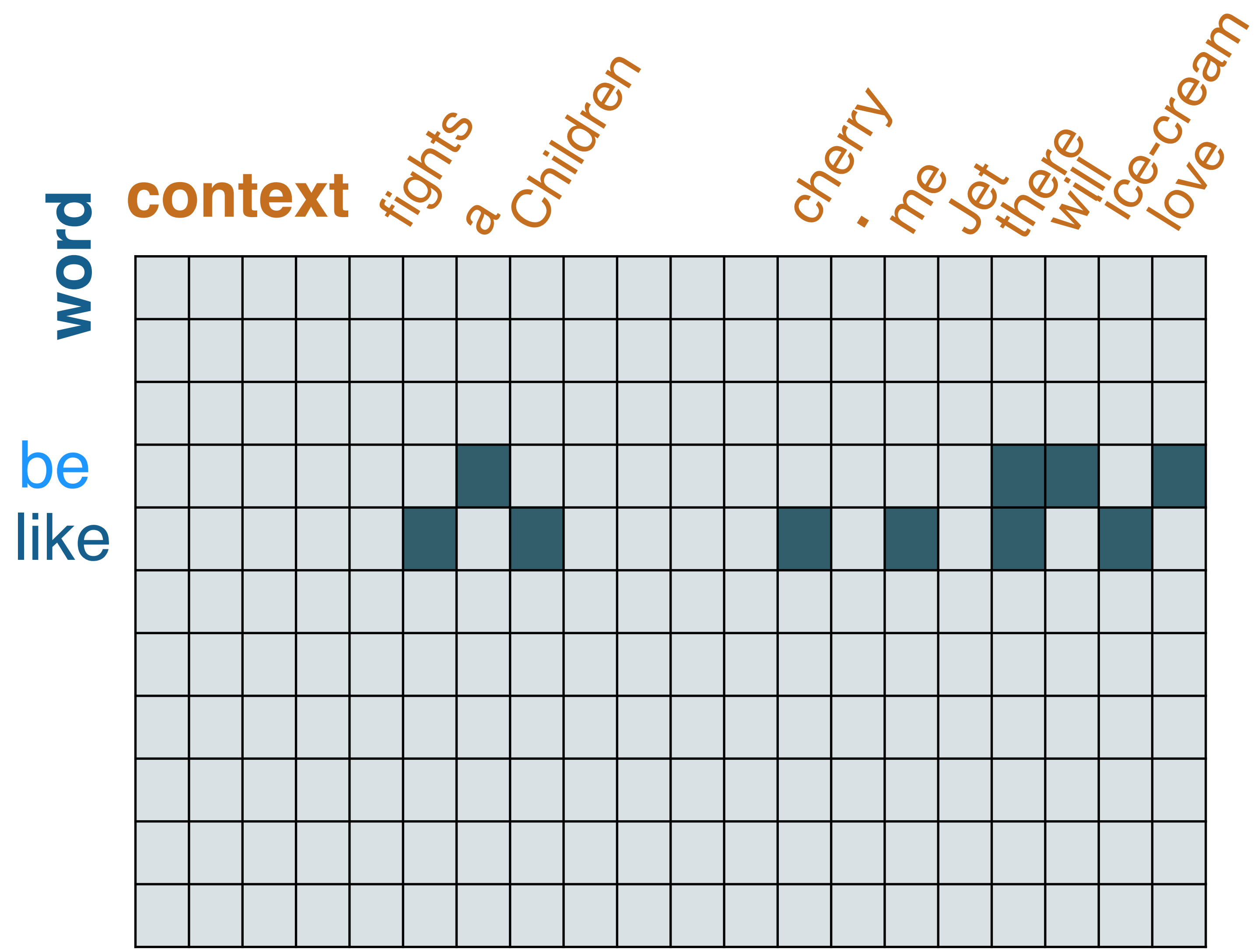
Children **like** ice-cream.

Method of moments

Q $p(\text{context} \mid \text{word})$

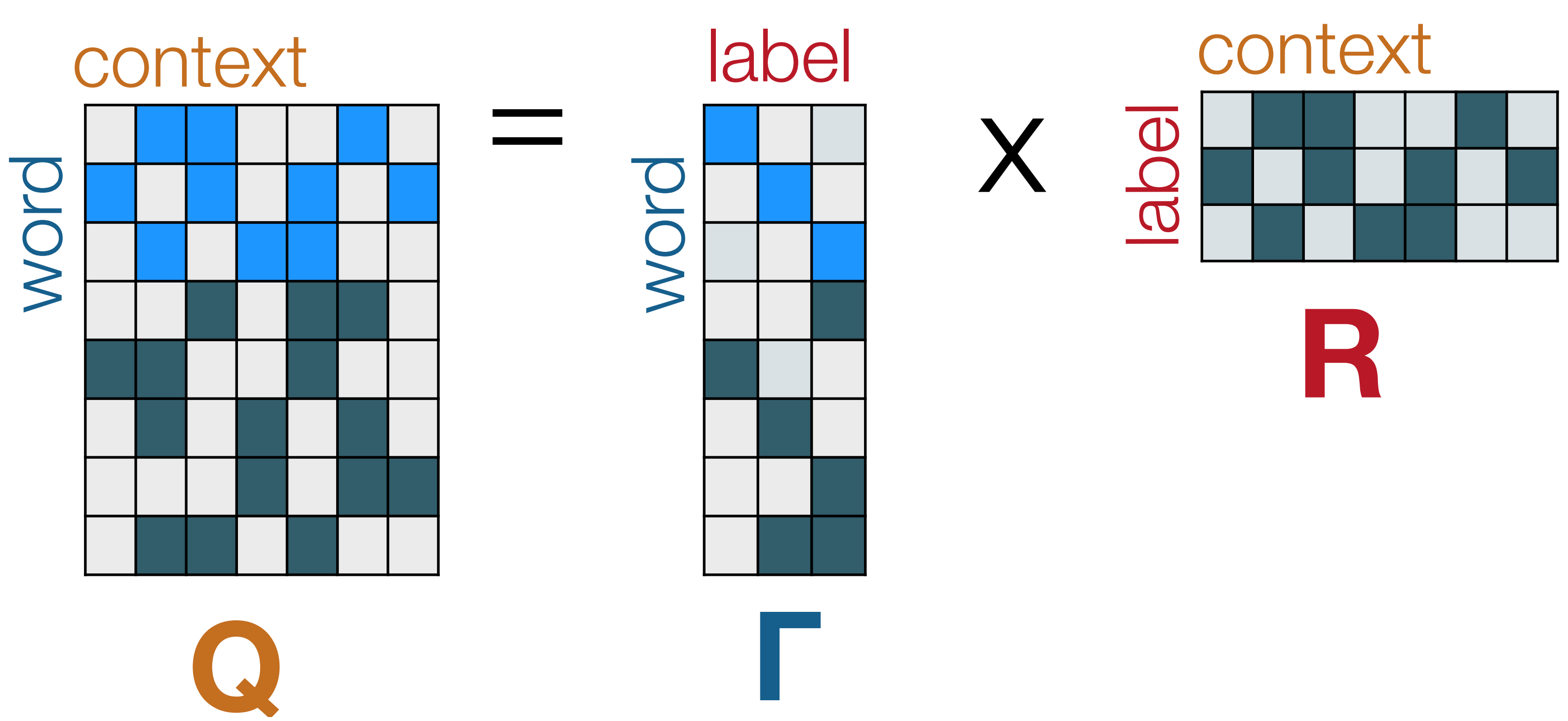
Let there **be** love.

Bill will **be** a ninja.



1. Conditional Independence word $\perp$ context | label

$$p(\text{context} \mid \text{word}) = \sum_{\text{labels}} p(\text{label} \mid \text{word}) \; p(\text{context} \mid \text{label})$$



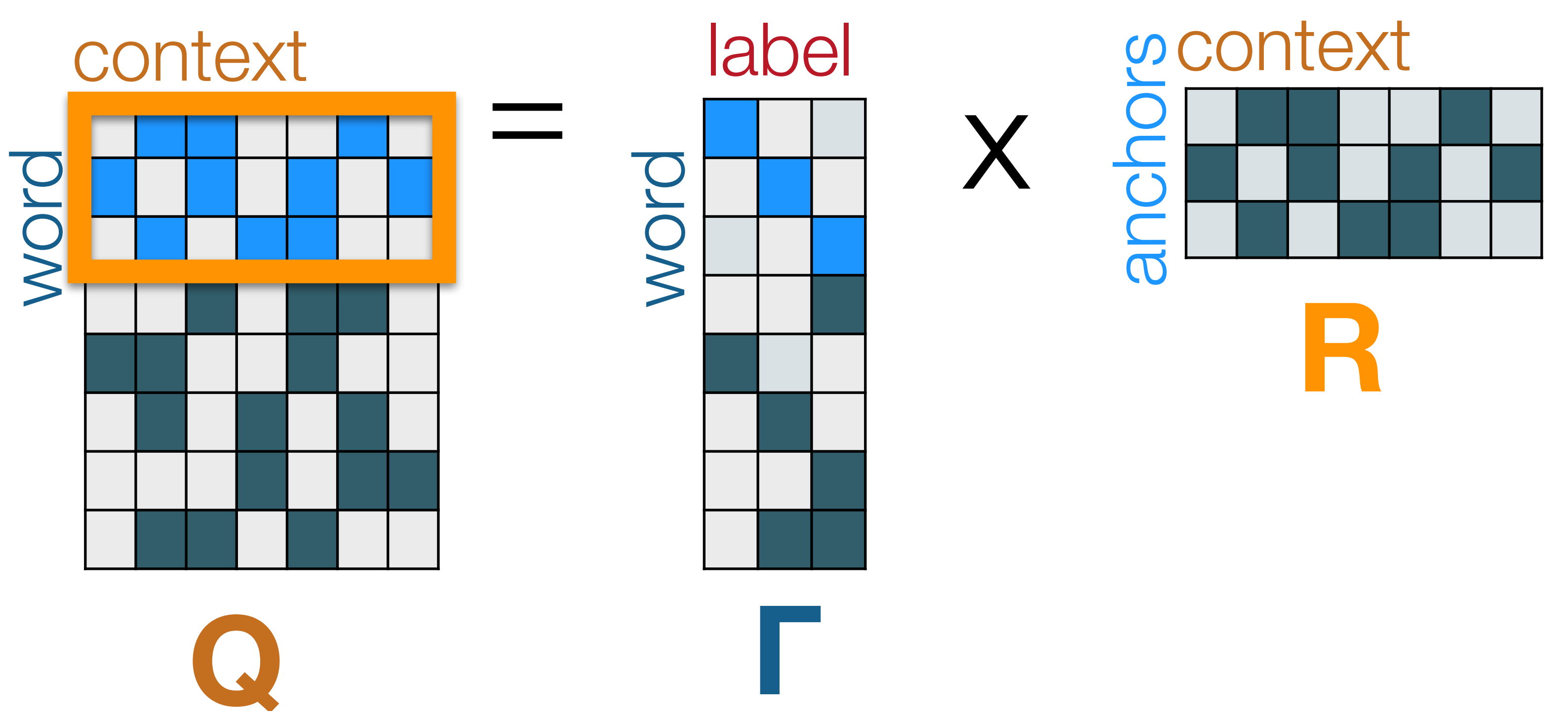
Method of moments

1. Conditional Independence   word $\perp$ context | label

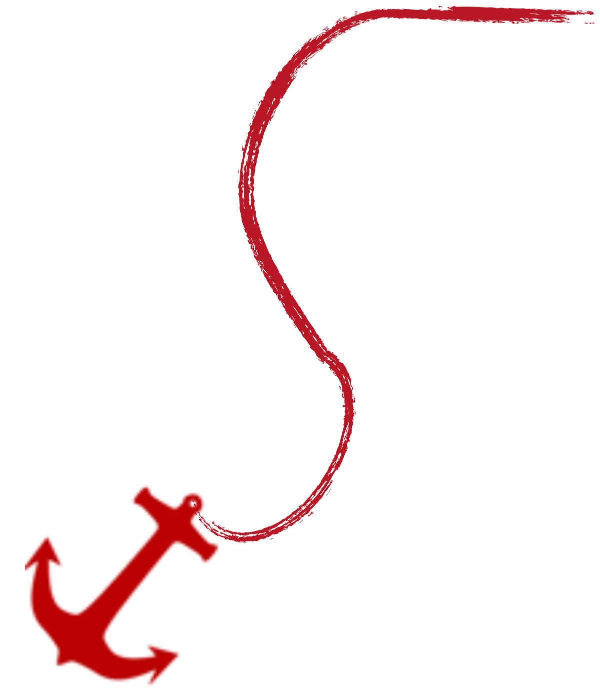
$$p(\text{context} \mid \text{word}) = \sum_{\text{labels}} p(\text{label} \mid \text{word}) \; p(\text{context} \mid \text{label})$$

2. Anchor Trick

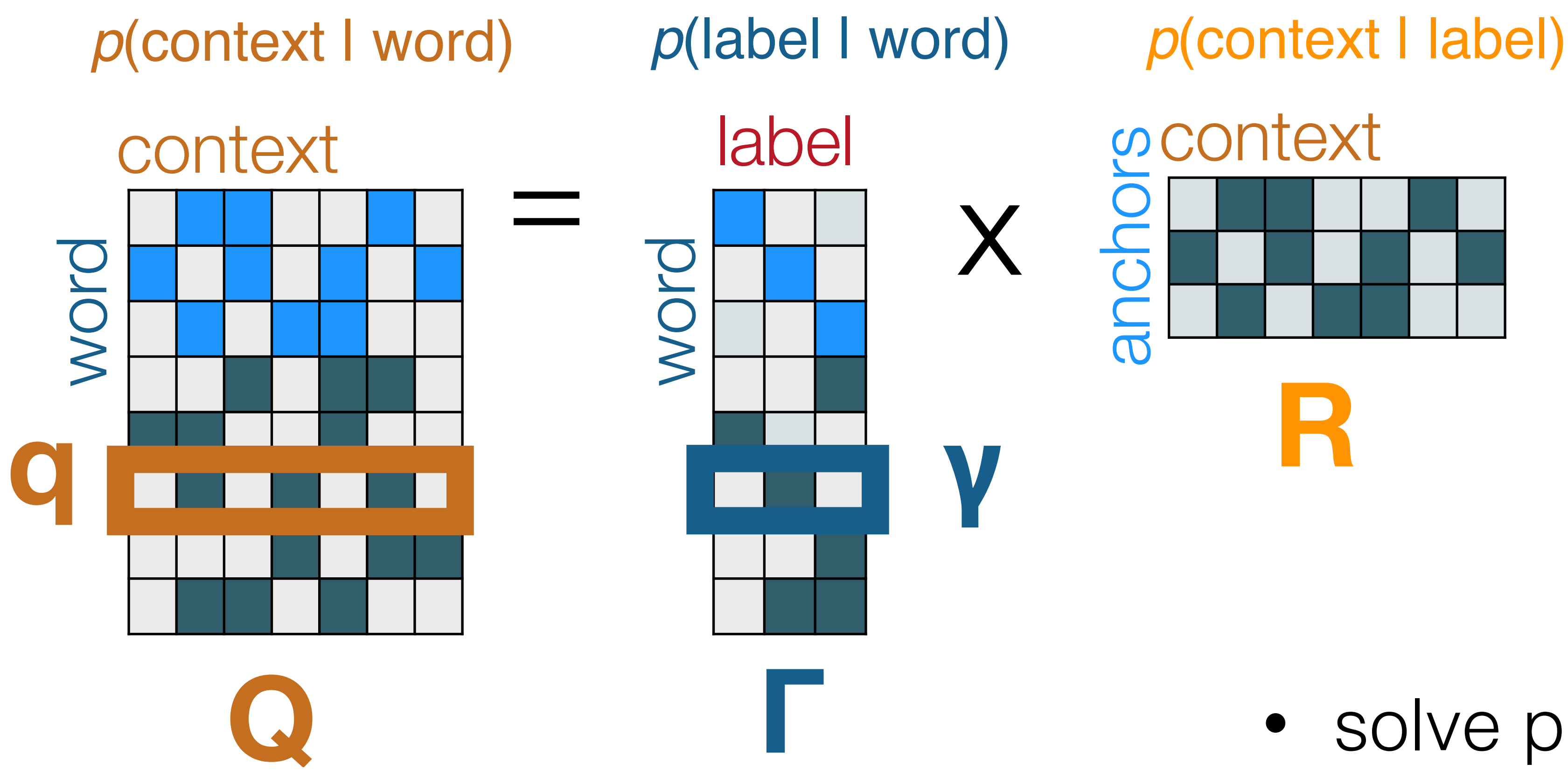
$$p(\text{context} \mid \text{word}) = \sum_{\text{labels}} p(\text{label} \mid \text{word}) \; p(\text{context} \mid \text{anchors})$$



Outline



1. Learn HMM models via MoM
2. Solve a QP
3. Extend to feature-based model
4. Experiments



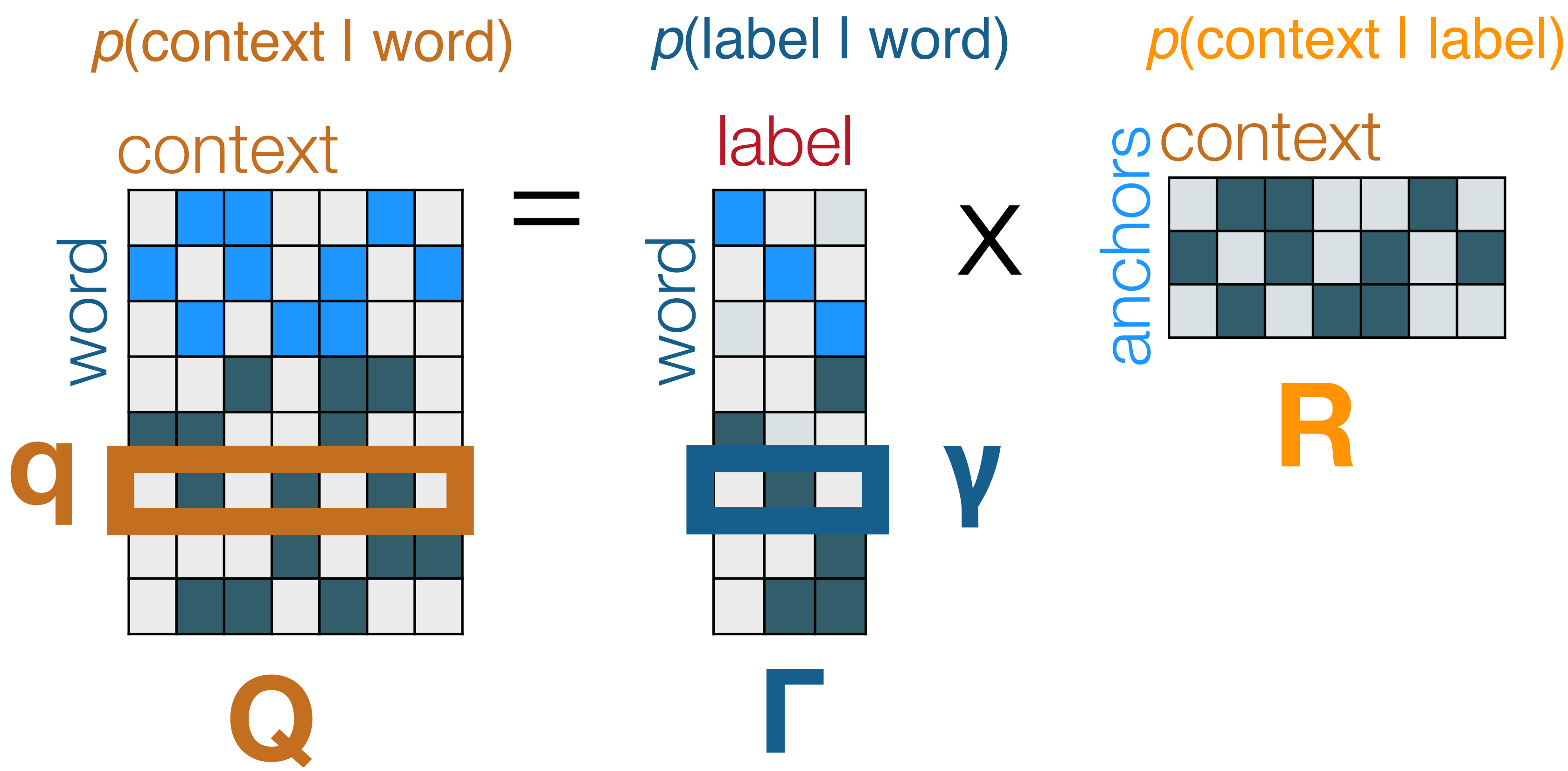
- solve per word type ~(ms)

$$\boldsymbol{\gamma} = \mathrm{argmin} \; || \, \mathbf{q} - \mathbf{R} \, \boldsymbol{\gamma} \, ||^2$$

$$0 \leq \boldsymbol{\gamma} \leq 1$$

$$\sum_{\text{labels}} \boldsymbol{\gamma} = 1$$

Method of Moments

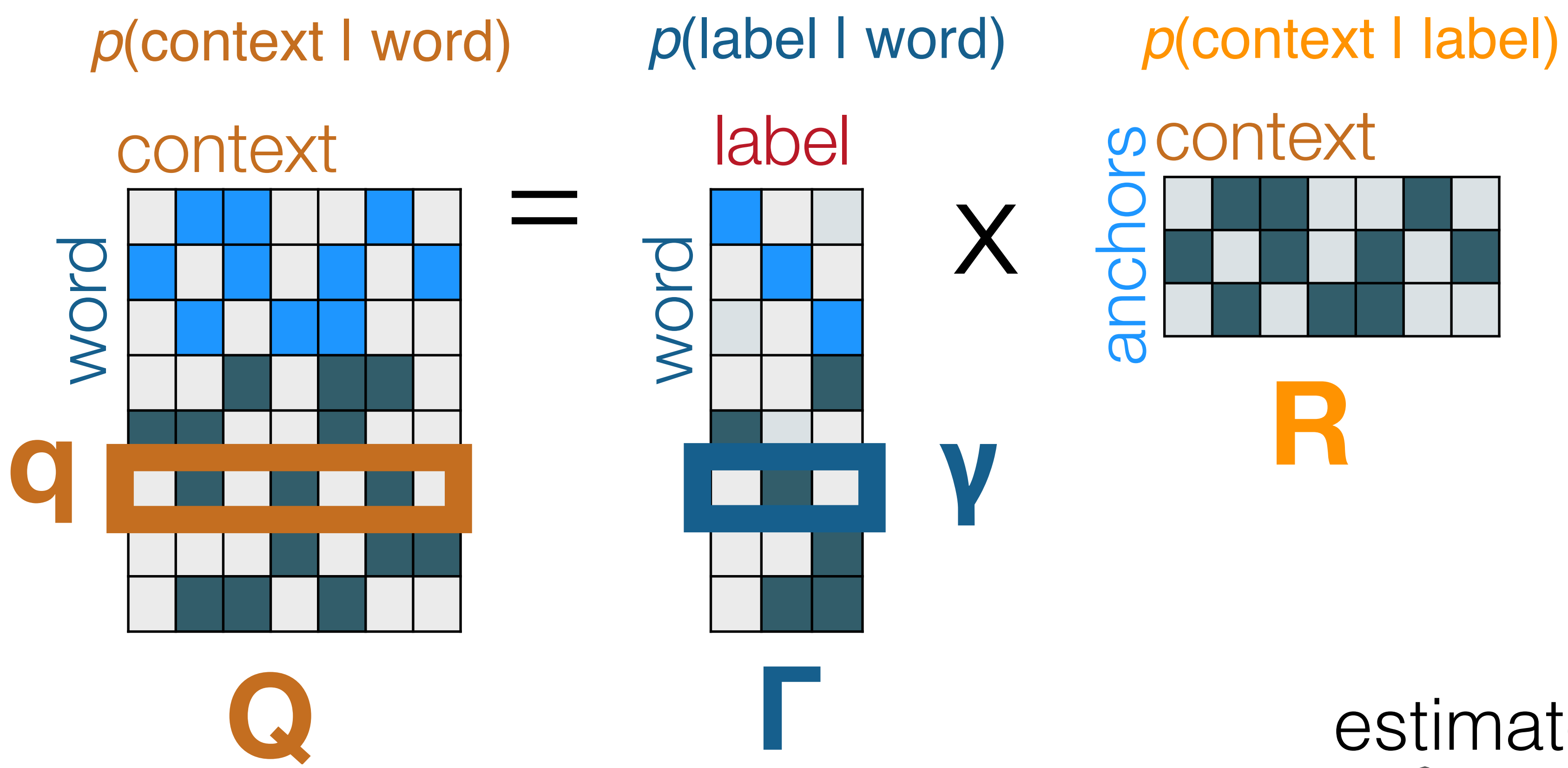


$$\boldsymbol{\gamma} = \text{argmin} \; \| \mathbf{q} - \mathbf{R}\boldsymbol{\gamma} \|^2 \quad + \quad \lambda \, \| \boldsymbol{\gamma}_{\mathbf{sup}} - \boldsymbol{\gamma} \|^2$$

$$0 \leq \boldsymbol{\gamma} \leq 1$$

$$\sum_{\text{labels}} \boldsymbol{\gamma} = 1$$

Method of Moments



estimated from labeled data

$$\gamma = \operatorname{argmin} \| q - R\gamma \|^2 \quad + \quad \lambda \| \gamma_{sup} - \gamma \|^2$$

estimated from unlabeled data

$$0 \leq \gamma \leq 1$$

$$\sum_{\text{labels}} \gamma = 1$$

Learn parameters ?

p(label | word)

γ

coefficients

↓

Observation Matrix

Bayes' Rule

$$p(\textbf{word} \mid \textbf{label}) = \gamma \frac{p(\text{word})}{p(\text{label})}$$

$$p(\text{label}) = \sum_{\text{words}} \gamma \; p(\text{word})$$

Learn parameters ?

Observation Matrix

Bayes' Rule

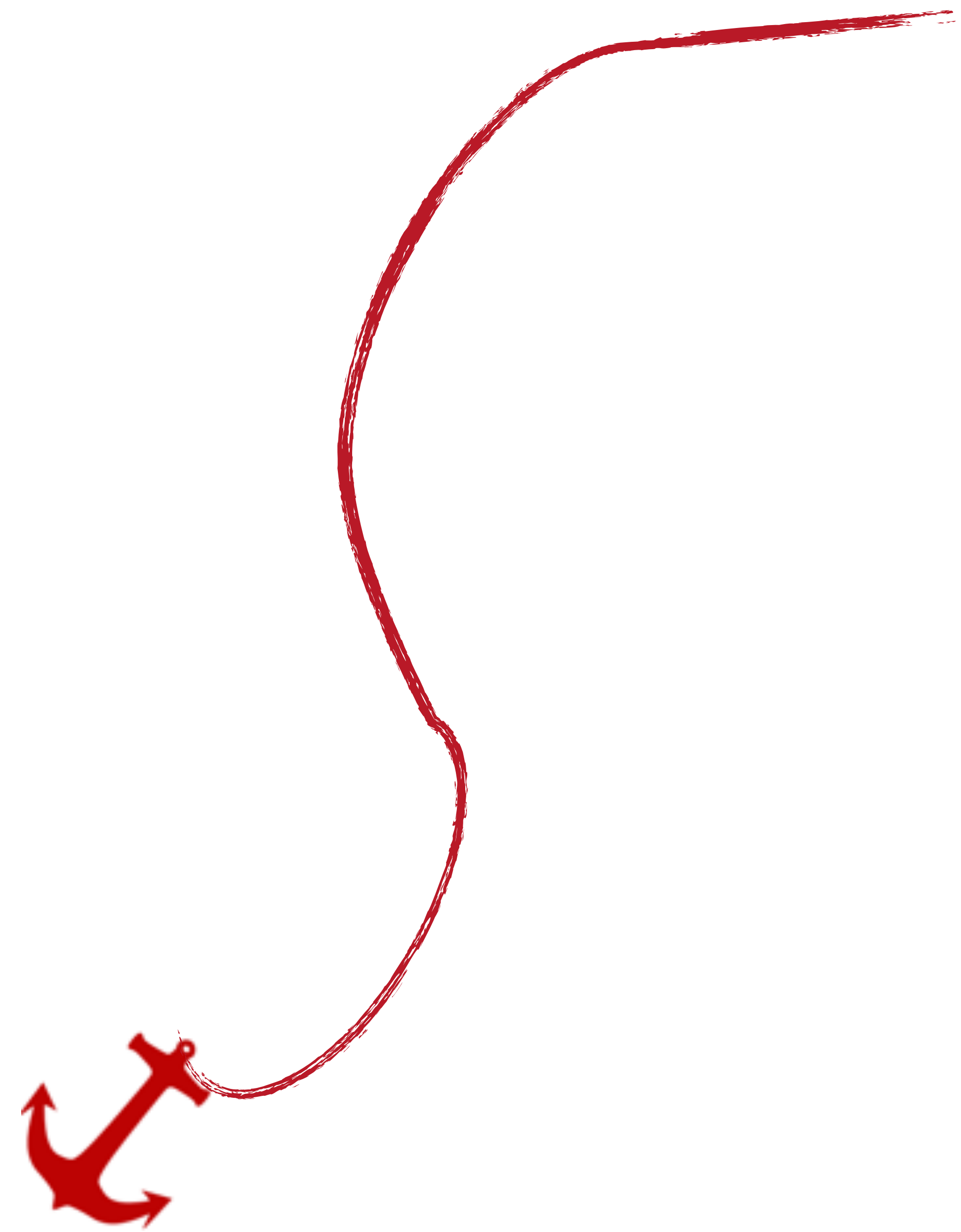
$$p(\textbf{word} \mid \textbf{label}) = \boldsymbol{\gamma} \frac{p(\text{word})}{p(\text{label})}$$

Transition Matrix

- estimate from labeled data only

Outline

1. Learn HMM models via MoM
2. Relax the notion of *anchors*
3. Solve a QP
4. Experiments



Semi-supervised Twitter POS tagging

Twitter dataset

2.7 M unlabeled tweets
1000-100 labeled tweets

$\approx$ 200k words

12 Universal POS

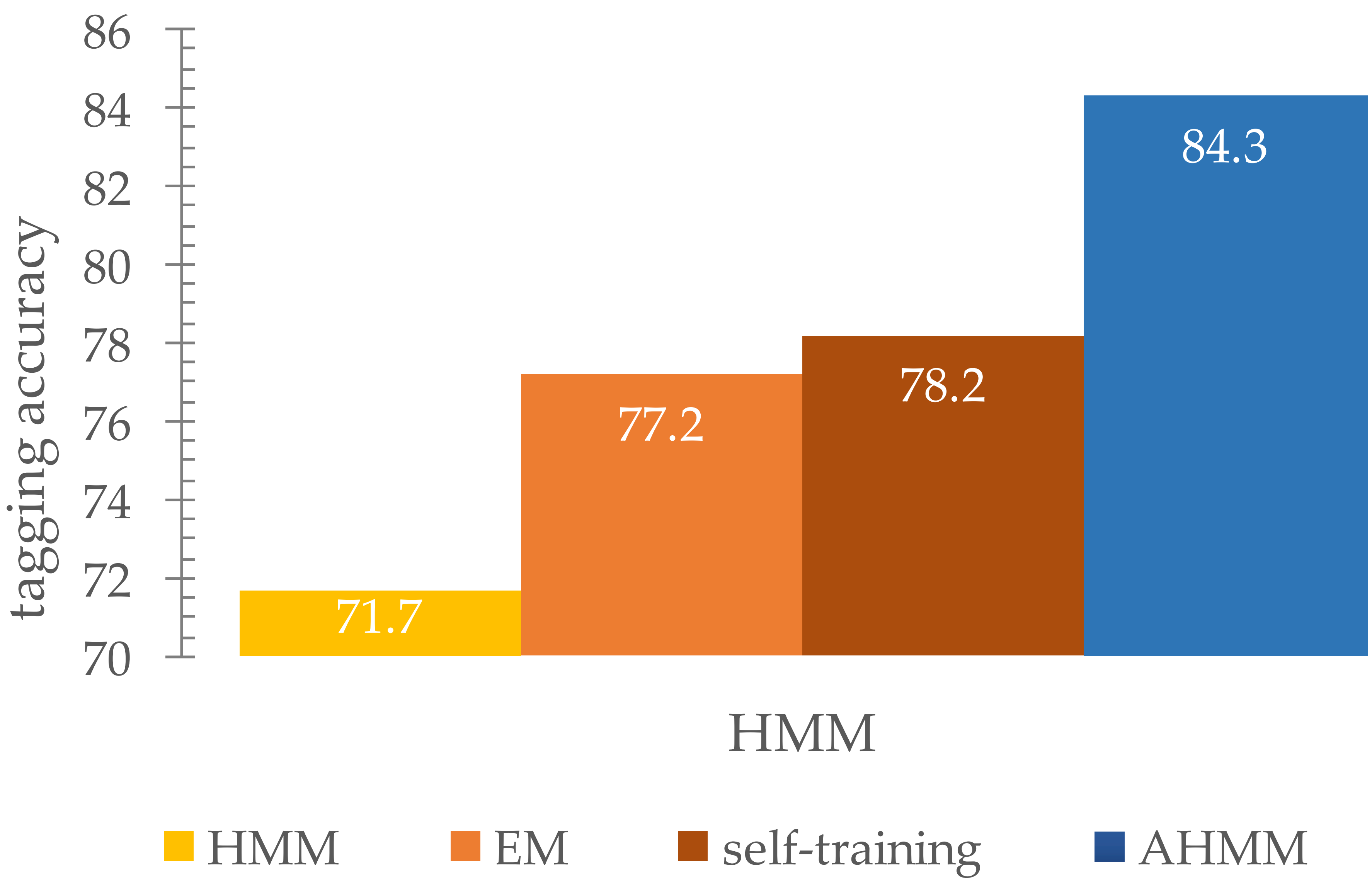
x	prt	verb	verb	det	adj	noun
hehe	its	gonna	b	a	good	day

Slav Petrov et al., *A Universal Part-of-Speech Tagset,* 2011

Owoputi et al., Improved part-of-speech tagging for online conversational text with word clusters. 2013

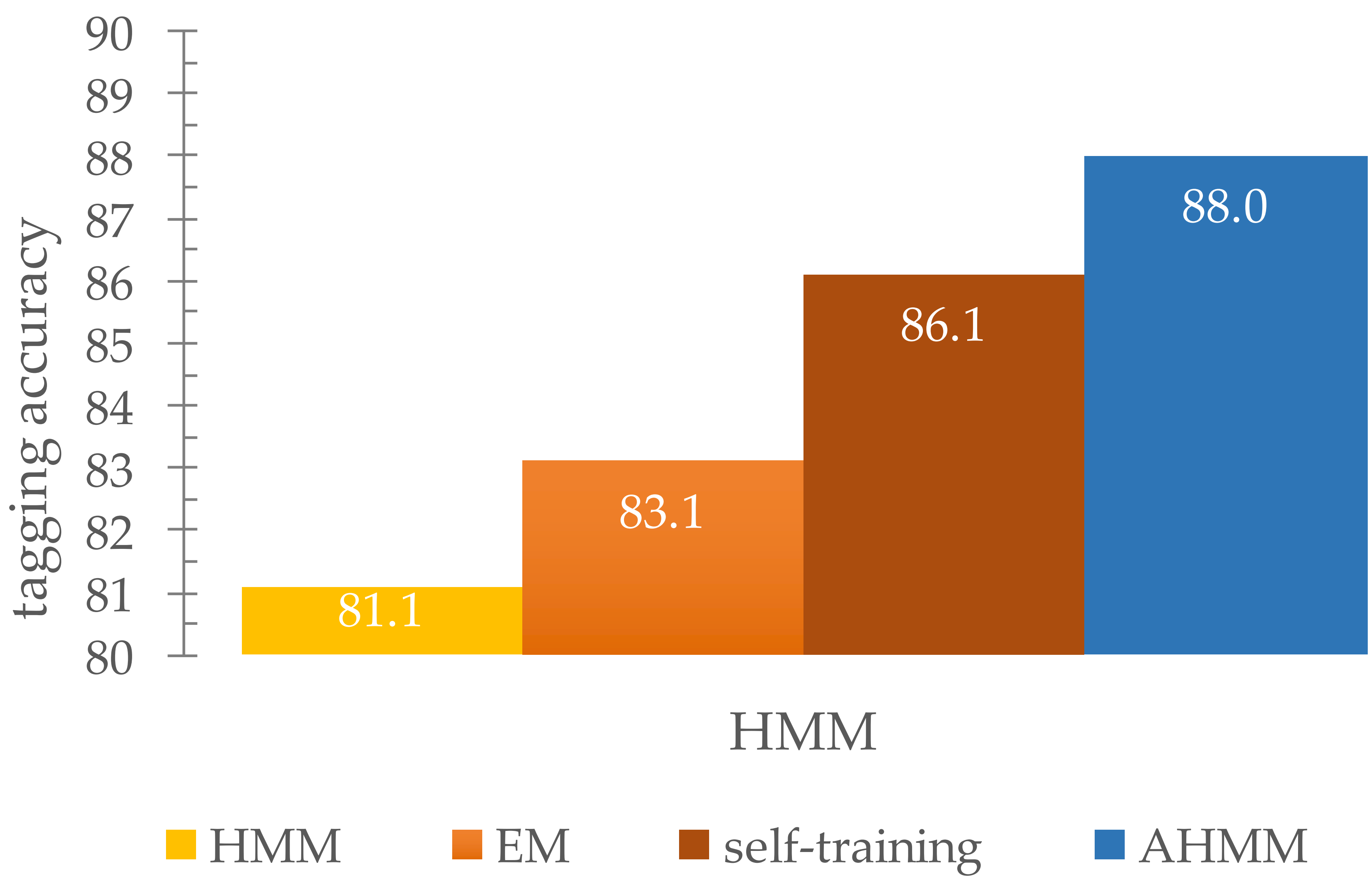
Twitter POS tagging

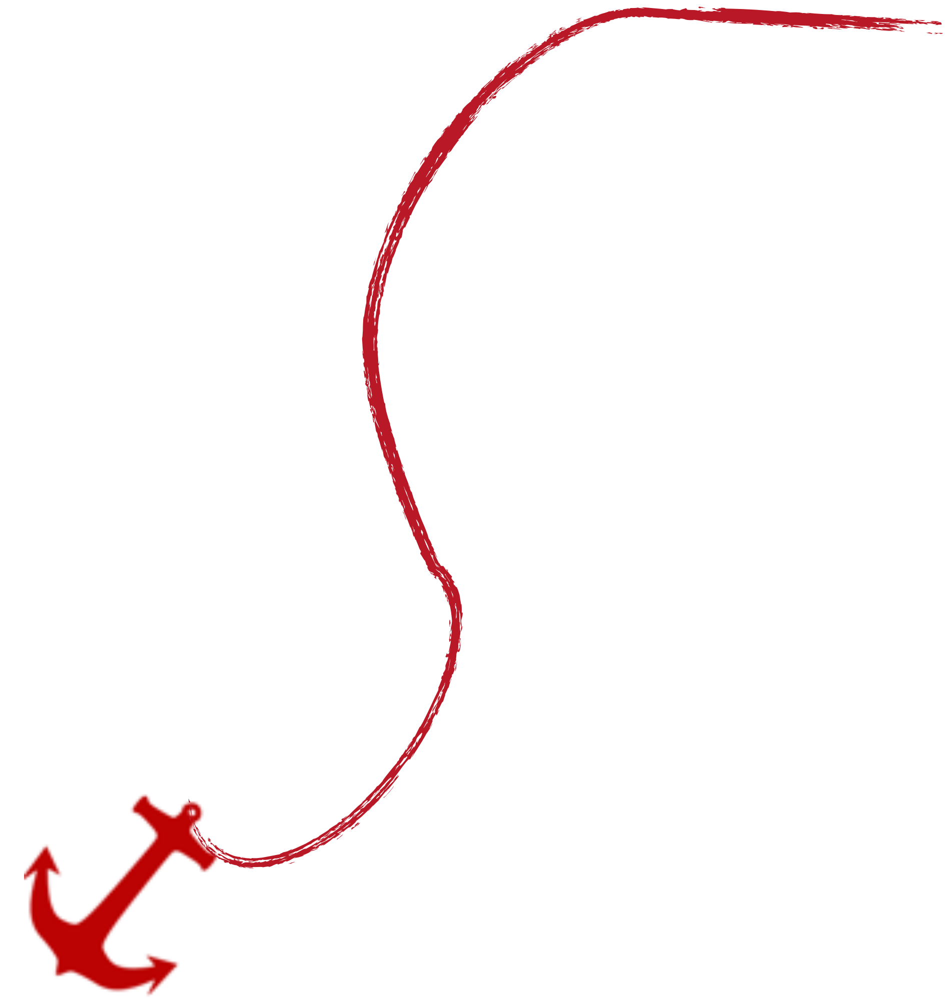
150 training labeled sequences



Twitter POS tagging

1000 training labeled sequences





Outline

1. Learn HMM models via MoM

2. Relax the notion of *anchors*

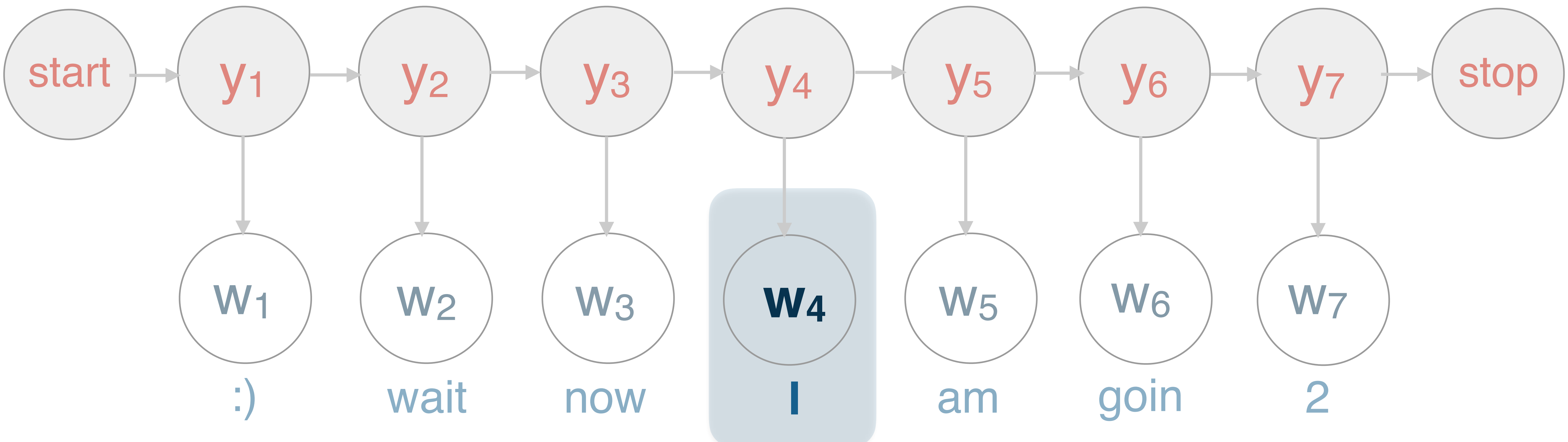
3. Extend to feature HMM

4. Experiments

Log-linear model

ϕ (word) →

- is upper
- is title
- is digit
- is url
- starts #
- is emoticon

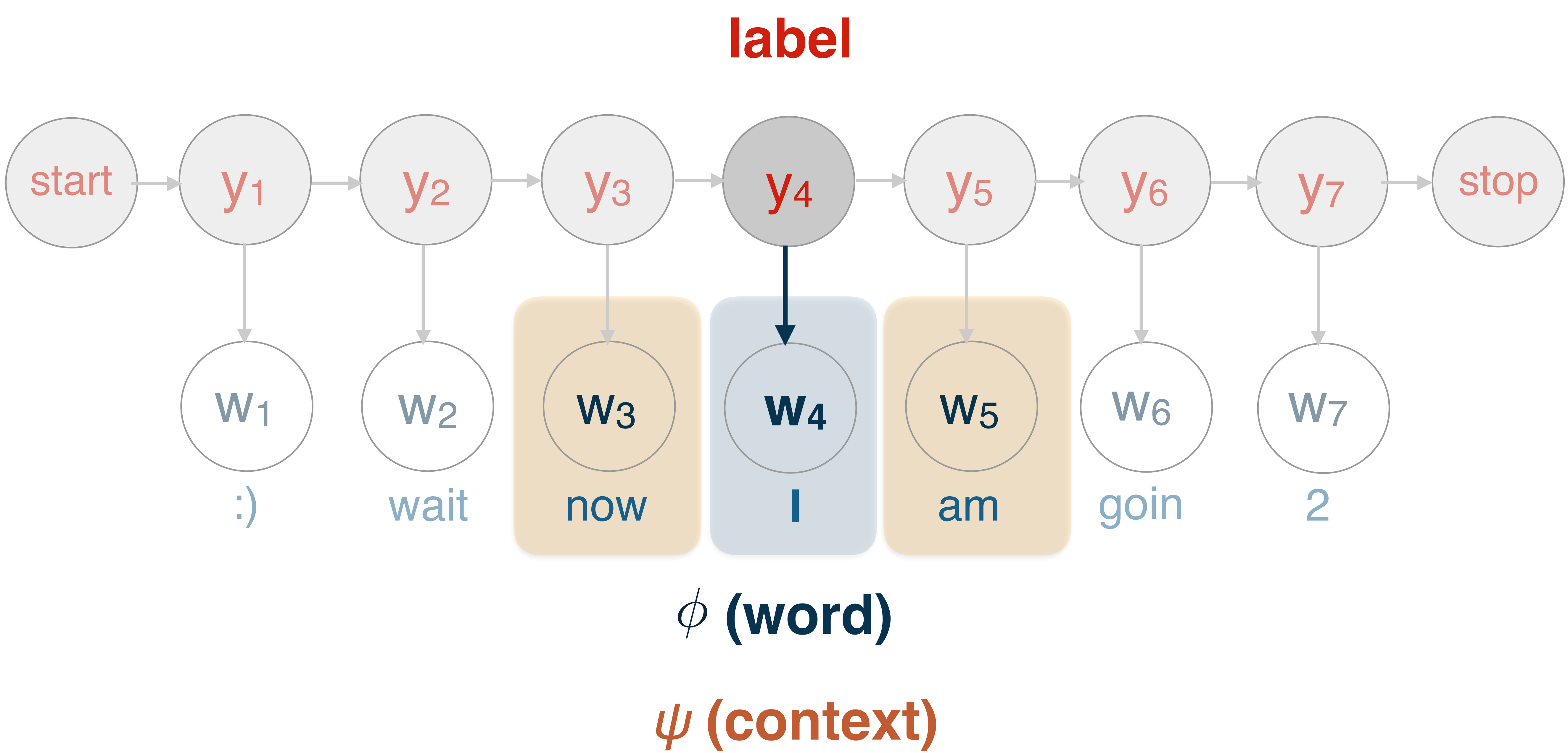


ϕ (word)

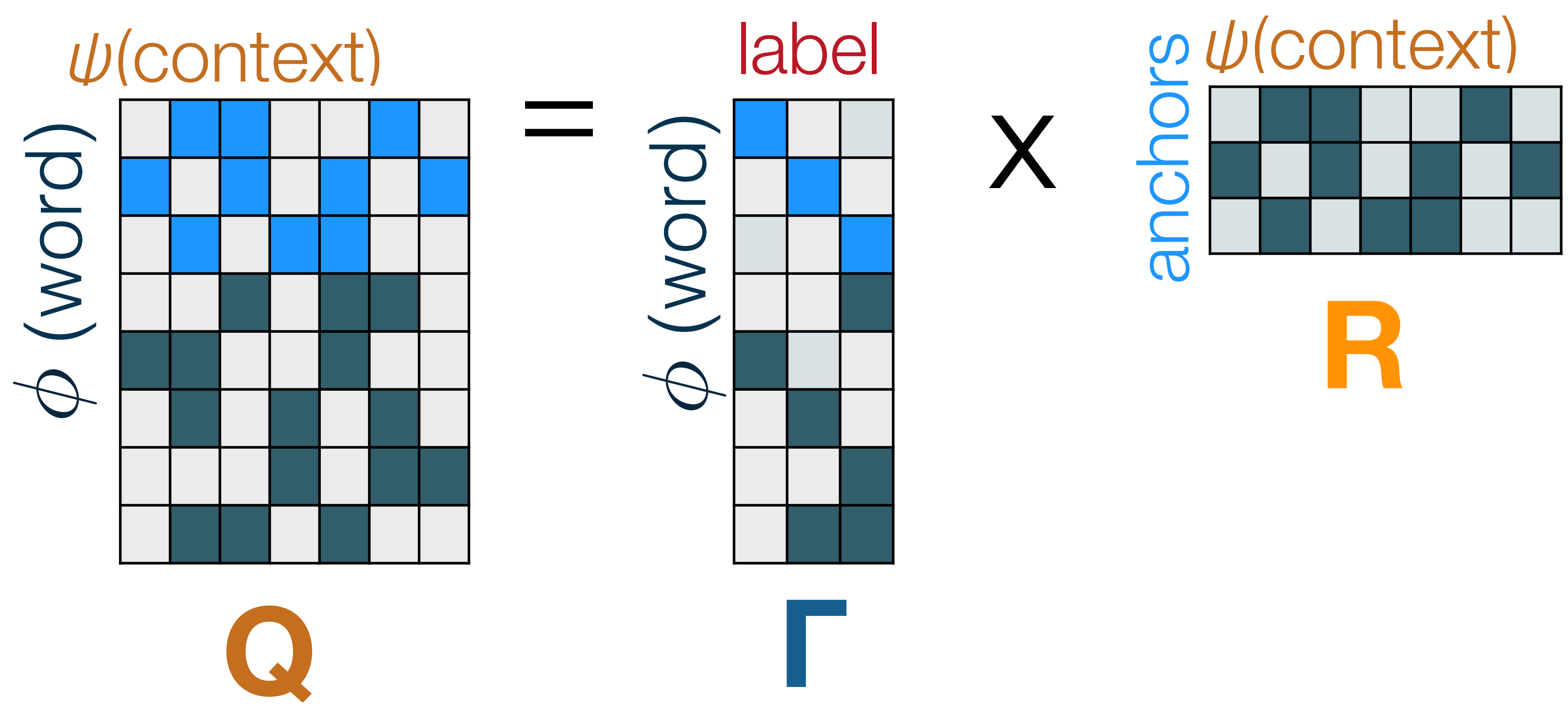
T. Berg-Kirkpatrick, *Painless unsupervised learning with features*, ACL 2010.

1. Conditional Independence

word $\perp$ context | label



Log-linear model

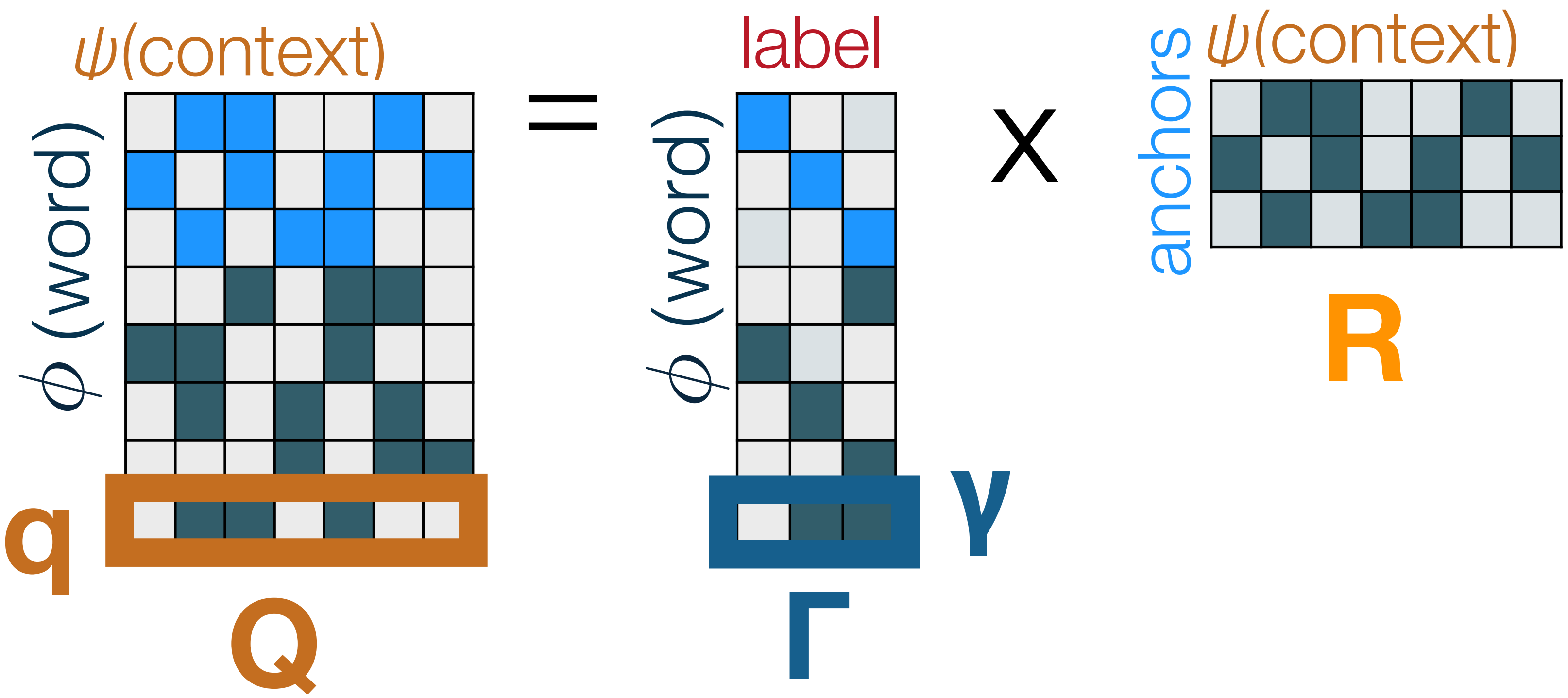


R p (context | label) $\longrightarrow$ $E[\psi(\text{context}) \mid \text{label}]$

Q p (context | word) $\longrightarrow$ $\dfrac{E[\,\psi(\text{context}) \times \Phi(\text{word})\,]}{E[\,\Phi(\text{word})\,]}$

Γ p (label | word) $\longrightarrow$ $\dfrac{E[\,\Phi(\text{word}) \mid \text{label}\,]\, p(\text{label})}{E[\,\Phi(\text{word})\,]}$

Log-linear model



- solve per feature dimension Φ_j

$$\gamma = \operatorname{argmin} \| q - R\gamma \|^2 + \lambda \| \gamma_{sup} - \gamma \|^2$$

$$\sum_{\text{labels}} \gamma = 1$$

Learn parameters ?

$$\gamma = \frac{E\,[\,\Phi(\text{word}) \mid \text{label}\,]\; p(\,\text{label}\,)}{E\,[\,\Phi(\text{word})\,]}$$

mean parameters

$$\mu = E[\,\Phi(\textbf{word}) \mid \textbf{label}\,] = \gamma\,\frac{E[\Phi(\text{word})]}{p(\text{label})}$$

Learn parameters ?

mean parameters   canonical parameters

$$\mu \longrightarrow \boldsymbol{\theta_y}$$

Fenchel-Legendre Duality

$$\boldsymbol{\theta}_{\boldsymbol{y}}^{*} = \underset{\theta_y}{\operatorname{argmax}}\ {\theta_y}^{\top} \boldsymbol{\mu_y} - \log \mathcal{Z}_y$$

partition function

$$\mathcal{Z}_y = \sum_{w} \exp\left(\theta_y^{\top} t_w\right)$$

Algorithm

compute moments — **Q** — ~ 10s min

↓

find anchors — **R**  — ~ 10s sec

↓

solve QP — **Γ** — ~ 2-3h

↓

mean parameters μ — ~ secs

↓

solve maxent problem — canonical parameters θ_y — ~ 10s min

Algorithm

compute moments $\mathbf{Q}$

find anchors $\mathbf{R}$

supervision

solve QP $\mathbf{\Gamma}$

mean parameters μ

solve maxent problem

canonical parameters $\boldsymbol{\theta_y}$

Learning sequence models via MoM

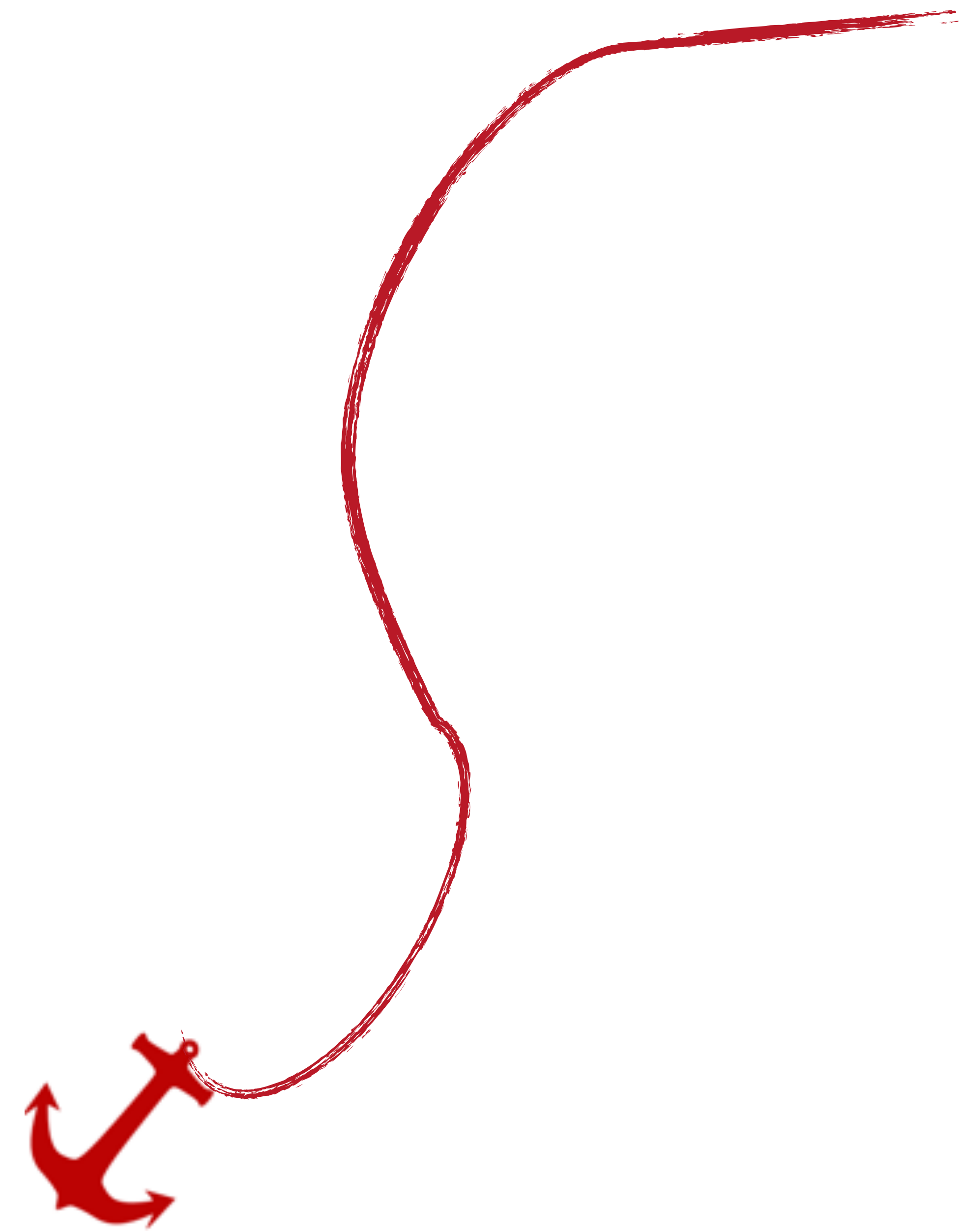
Outline

1. Learn HMM models via MoM

2. Relax the notion of *anchors*

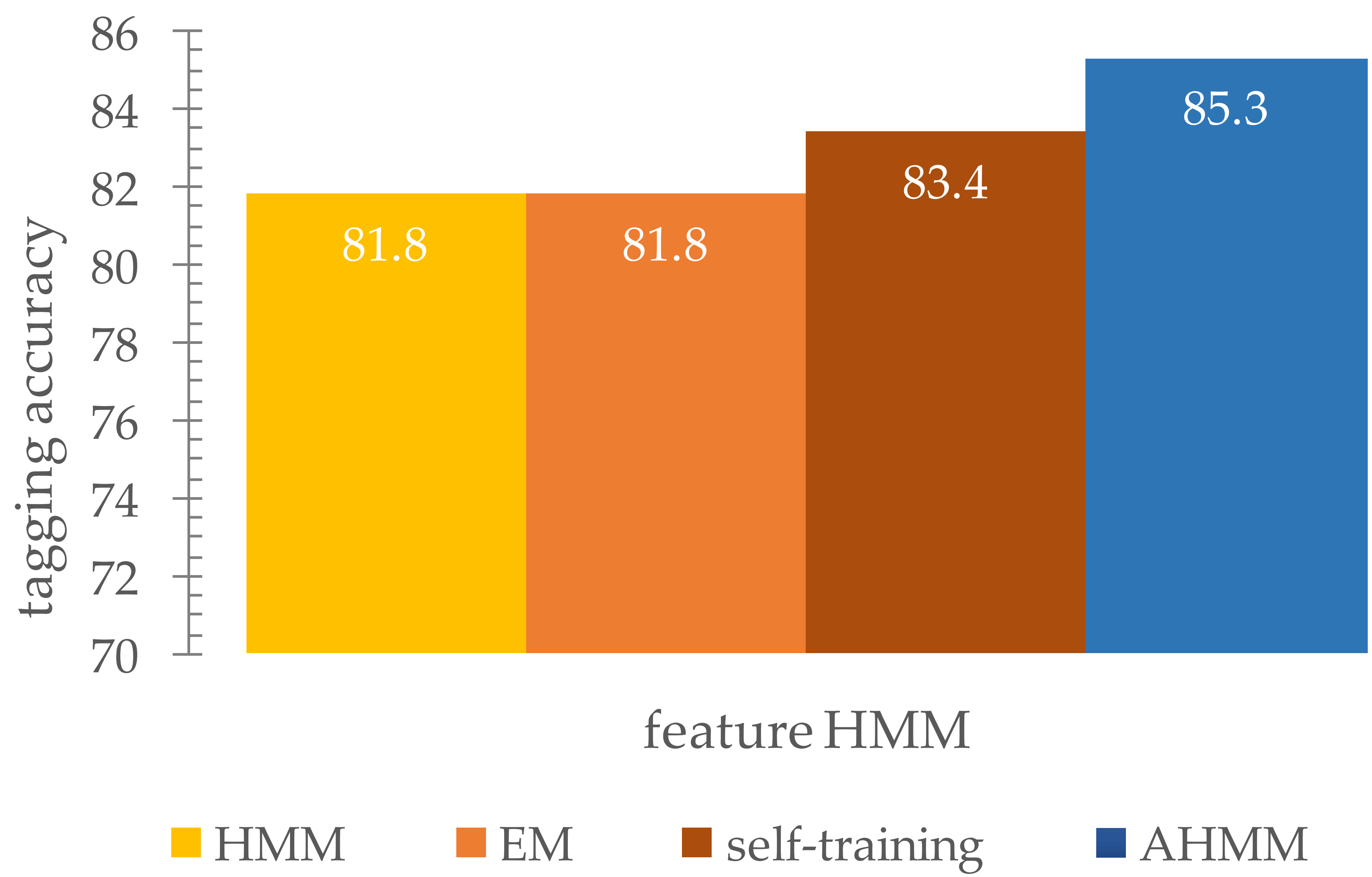
3. Solve a QP

4. Experiments



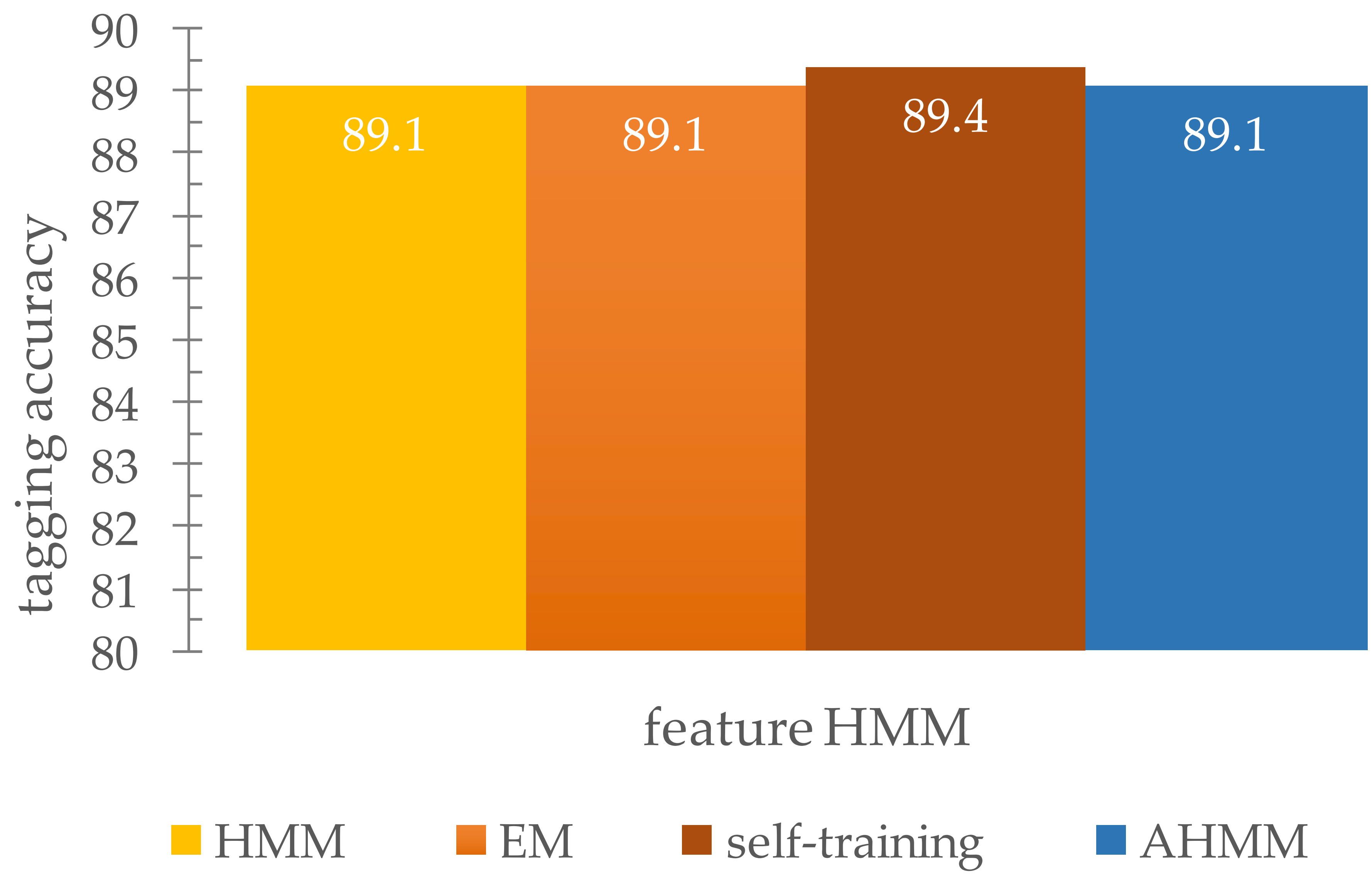
Twitter POS tagging

150 training labeled sequences

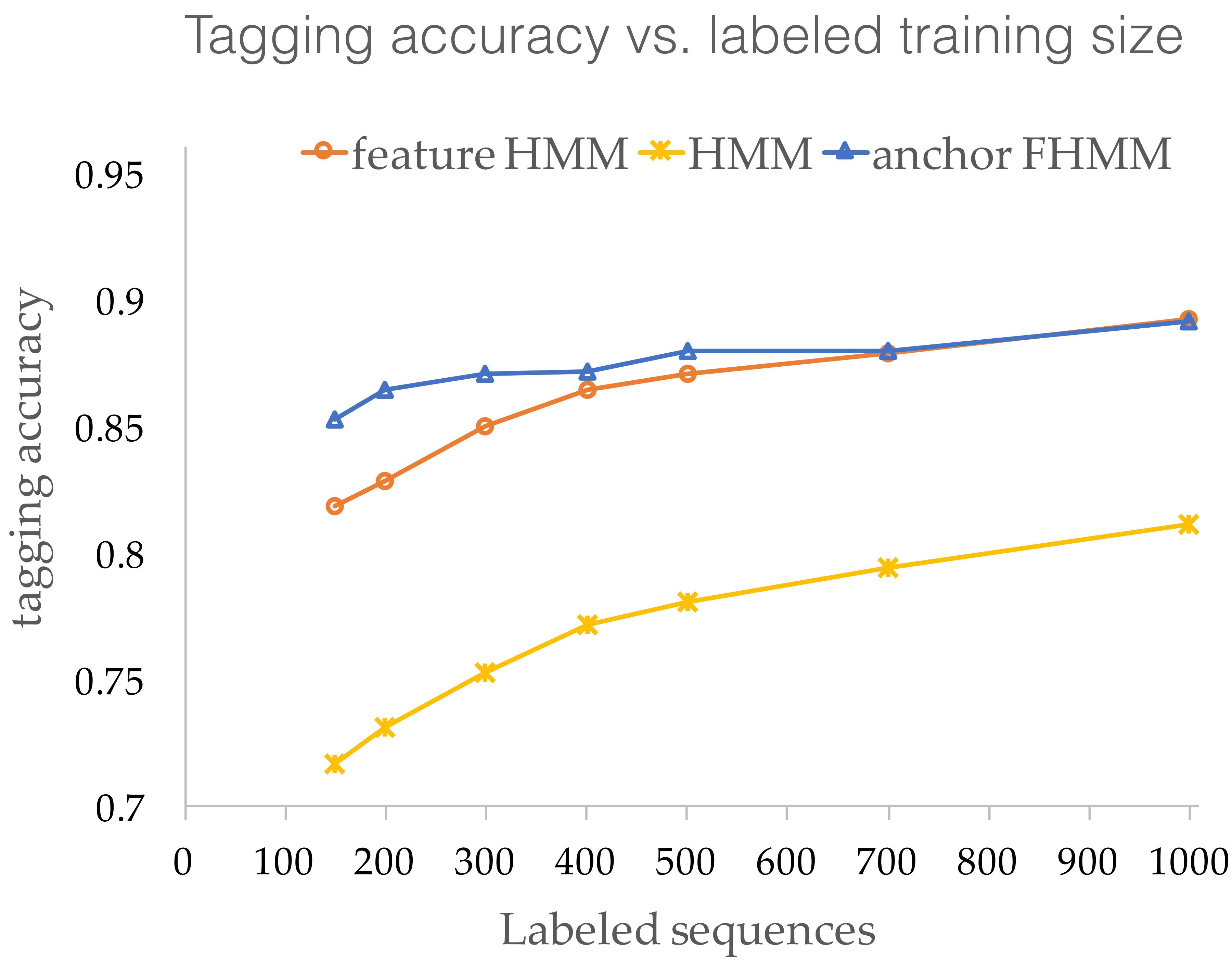


Twitter POS tagging

1000 training labeled sequences

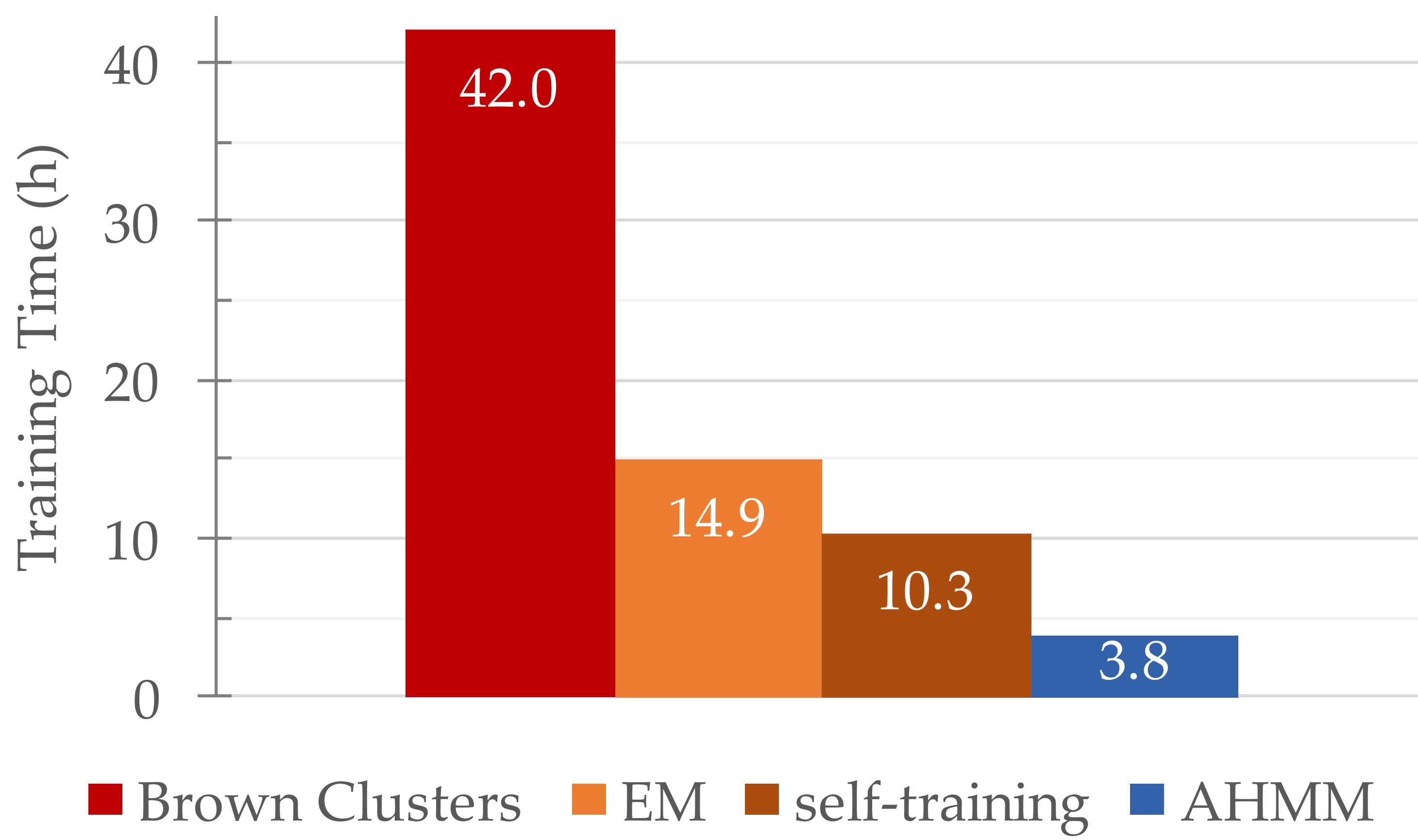


Twitter POS tagging



Twitter POS tagging

1000 training sequences



Conclusions

- MoM algorithm for semi-supervised learning
- flexible method
  (easy to add supervision)
- fast to train
  (only one pass over the data)
- particularly good with little supervision

Thank you !

zmarinho@cmu.edu

Support for this research was provided by the Portuguese Science and Technology Foundation (FCT) and CMU Portugal Program, grant SFRH/BD/52015/2012. This work has also been partially supported by the European Union under H2020 project SUMMA, grant 688139, and by FCT, through contracts UID/EEA/50008/2013, through the LearnBig project (PTDC/EEISII/7092/2014), and the GoLocal project (grant CMUPERI/TIC/0046/2014).