Conditional Random Fields for Named Entity Recognition

COSC 6336: Natural Language Processing
Spring 2020

UNIVERSITY of HOUSTON

Some content on these slides was adapted from Greg Durrett, J&M, Wei Xu, Sutton & McCallum, and Laura Kallmeyer

Previous Lecture

- ★ HMMs
  - Three tasks
  - Forward algorithm
  - Viterbi algorithm
  - Forward-backward (Baum-Welch) algorithm

Today's lecture

- ★ Named Entity Recognition
- ★ Conditional Random Fields (CRFs)

Named Entity Recognition (NER)

- ★ Specific type of information extraction in which the goal is to extract proper names of particular types of entities such as people, places, organizations, etc.
- ★ Usually a preprocessing step for subsequent task-specific IE, or other tasks such as question answering.
- ★ NEs are application specific

NER Example

U.S. Supreme Court quashes 'illegal' Guantanamo trials

Military trials arranged by the Bush administration for detainees at Guantanamo Bay are illegal, the United States Supreme Court ruled Thursday. The court found that the trials — known as military commissions — for people detained on suspicion of terrorist activity abroad do not conform to any act of Congress. The justices also rejected the government's argument that the Geneva Conventions regarding prisoners of war do not apply to those held at Guantanamo Bay. Writing for the 5-3 majority, Justice Stephen Breyer said the White House had overstepped its powers under the U.S. Constitution. "Congress has not issued the executive a blank cheque," Breyer wrote.

President George W. Bush said he takes the ruling very seriously and would find a way to both respect the court's findings and protect the American people.

NER Example

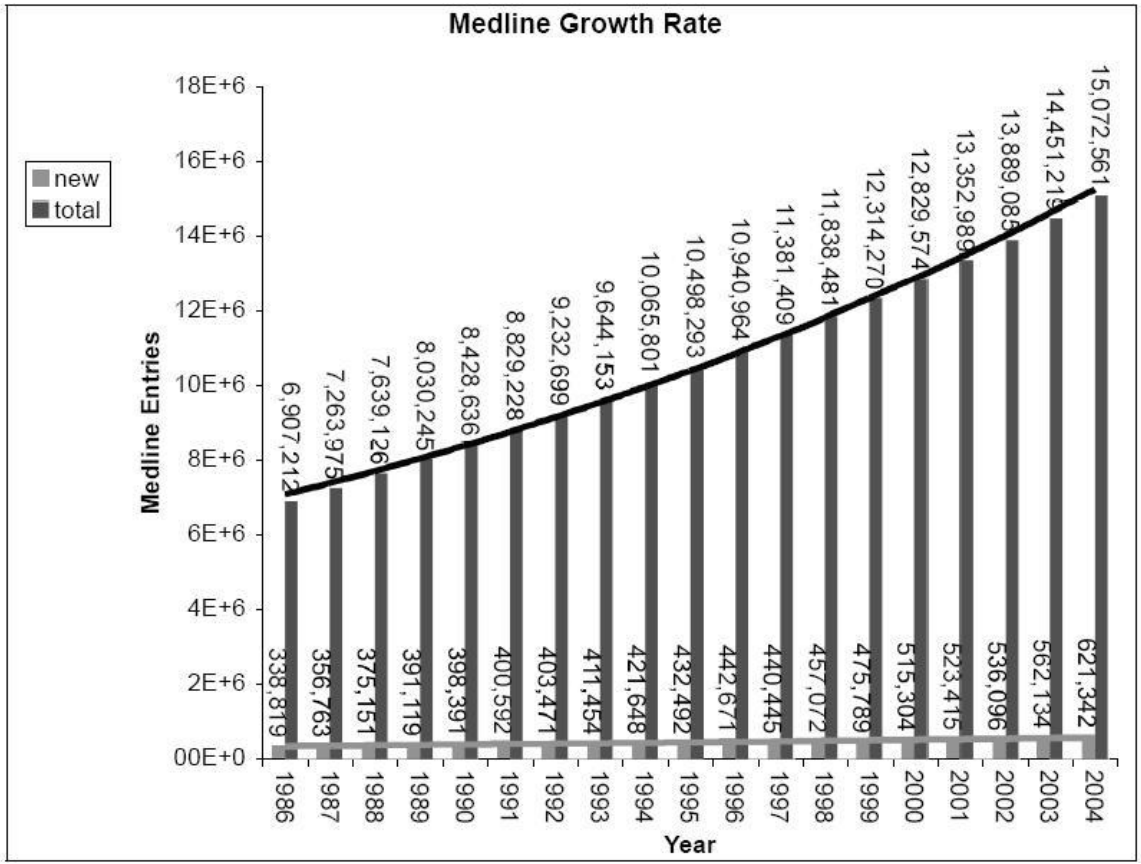
people places organizations

U.S. Supreme Court quashes 'illegal' Guantanamo trials

Military trials arranged by the Bush administration for detainees at Guantanamo Bay are illegal, the United States Supreme Court ruled Thursday. The court found that the trials — known as military commissions — for people detained on suspicion of terrorist activity abroad do not conform to any act of Congress. The justices also rejected the government's argument that the Geneva Conventions regarding prisoners of war do not apply to those held at Guantanamo Bay. Writing for the 5-3 majority, Justice Stephen Breyer said the White House had overstepped its powers under the U.S. Constitution. "Congress has not issued the executive a blank cheque," Breyer wrote.

President George W. Bush said he takes the ruling very seriously and would find a way to both respect the court's findings and protect the American people.

Biomedical Information Extraction



UNIVERSITY of HOUSTON

Medline Corpus

TI - Two potentially oncogenic cyclins, cyclin A and cyclin D1, share common properties of subunit configuration, tyrosine phosphorylation and physical association with the Rb protein

AB - Originally identified as a ‘mitotic cyclin’, cyclin A exhibits properties of growth factor sensitivity, susceptibility to viral subversion and association with a tumor-suppressor protein, properties which are indicative of an S-phase-promoting factor (SPF) as well as a candidate proto-oncogene …

Moreover, cyclin D1 was found to be phosphorylated on tyrosine residues in vivo and, like cyclin A, was readily phosphorylated by pp60c-src in vitro.

In synchronized human osteosarcoma cells, cyclin D1 is induced in early G1 and becomes associated with p9Ckshs1, a Cdk-binding subunit.

Immunoprecipitation experiments with human osteosarcoma cells and Ewing’s sarcoma cells demonstrated that cyclin D1 is associated with both p34cdc2 and p33cdk2, and that cyclin D1 immune complexes exhibit appreciable histone H1 kinase activity …

Medline Corpus: NER (Proteins)

TI - Two potentially oncogenic cyclins, cyclin A and cyclin D1, share common properties of subunit configuration, tyrosine phosphorylation and physical association with the Rb protein

AB - Originally identified as a 'mitotic cyclin', cyclin A exhibits properties of growth factor sensitivity, susceptibility to viral subversion and association with a tumor-suppressor protein, properties which are indicative of an S-phase-promoting factor (SPF) as well as a candidate proto-oncogene …

Moreover, cyclin D1 was found to be phosphorylated on tyrosine residues in vivo and, like cyclin A, was readily phosphorylated by pp60c-src in vitro.

In synchronized human osteosarcoma cells, cyclin D1 is induced in early G1 and becomes associated with p9Ckshs1, a Cdk-binding subunit.

Immunoprecipitation experiments with human osteosarcoma cells and Ewing's sarcoma cells demonstrated that cyclin D1 is associated with both p34cdc2 and p33cdk2, and that cyclin D1 immune complexes exhibit appreciable histone H1 kinase activity …

Short History of NER

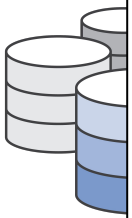
1996-1997 — Rule-based Approaches

2003 — Traditional Machine Learnin

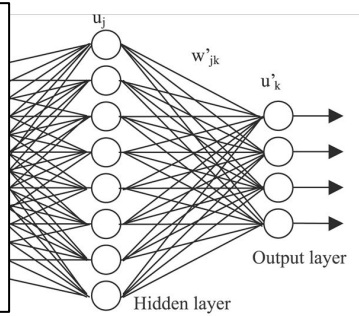
+ L

2017 — Embeddings + Neural Networks

- ★ Evaluation metrics
- ★ Text: Wall Street Journal
- ★ F-measure > 96%
- ★ Classes: PER, LOC, ORG
- ★ Grishman and Sundheim (1996)



Given this level of performance, there is probably little point in repeating this task with the same ground rules in a future MUC (although there might be interest in processing monocase text and in performing comparable tasks on a more varied corpus and for languages other than English).

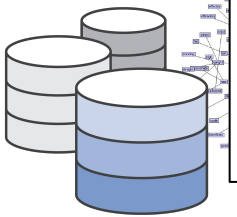


UNIVERSITY of HOUSTON

Short History of NER

1996-1997 — Rule-based Approaches

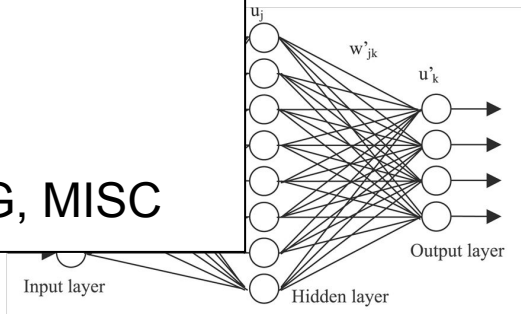
2003 — Traditional Machine Learning



Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition

Erik F. Tjong Kim Sang and **Fien De Meulder**
CNTS - Language Technology Group
University of Antwerp
{erikt,fien.demeulder}@uia.ua.ac.be

- ★ Text: Reuters
- ★ English and German
- ★ F-measure ~ 88%
- ★ Classes: PER, LOC, ORG, MISC



UNIVERSITY of HOUSTON

Short History of NER

1996-1997

Rule-based Approaches

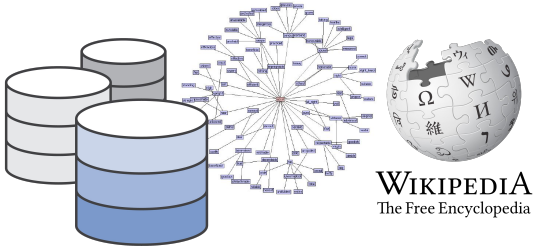
2003

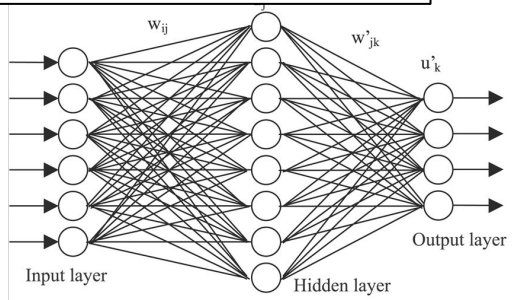
Traditional Machine Learning

2006

Exploiting Web-based Resources

- ★ ACE (Walker et al., 2006)
- ★ OntoNotes (Hovy et al., 2006) (English, Arabic Chinese)
- ★ Genres: newswire, speech, conversations, weblog, usenet groups
- ★ Use of web resources
- ★ State of the art: traditional ML





UNIVERSITY of HOUSTON

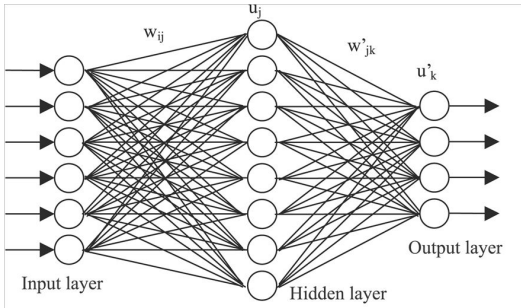
Short History of NER

- ★ UMBC (Finin et al., 2010)
- ★ Twitter Freebase (Ritter et al., 2011)
- ★ NE Classes: PERSON, GEO-LOCATION, COMPANY, PRODUCT, FACILITY, TV-SHOW, MOVIE, SPORTS TEAM, BAND, and OTHER

2010 — Social media data

2015 — Embeddings + Traditional ML

2016 — Embeddings + Neural Networks



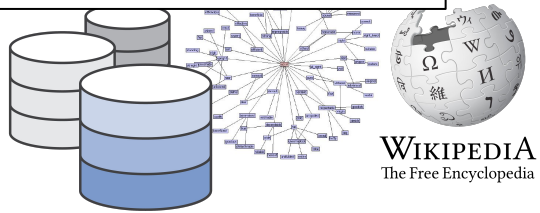
Short History of NER

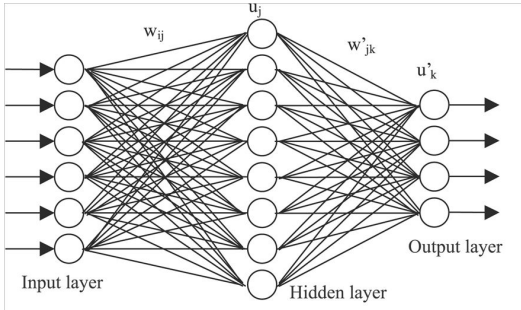
- ★ EMNLP W-NUT Shared Task on NER (Twitter)
- ★ Best system 57% f-measure!
- ★ Still traditional ML

2010 — Social media data

2015 — Embeddings + Traditional ML

2016 — Embeddings + Neural Networks





UNIVERSITY of HOUSTON

Short History of NER

1996-1997

Rule-based Approaches

M

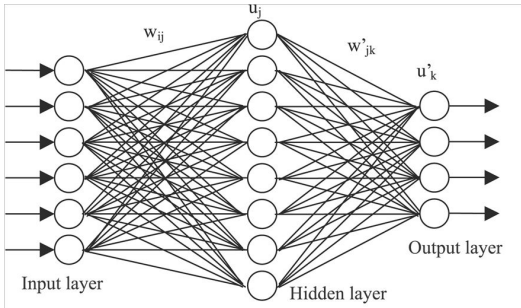
- ★ 2017 EMNLP W-NUT Shared Task on NER
- ★ Twitter, YouTube, Reddit, StackExchange
- ★ Best system 40% f-measure!
- ★ Deep Learning
- ★ Classes: PERSON, LOCATION, CORPORATION, PRODUCT, CREATIVE-WORK, GROUP

2015

Embeddings + Traditional ML

2017

Embeddings + Neural Networks



UNIVERSITY
HOUSTON

Short History of NER

1996-1997

Rule-based Approaches

M

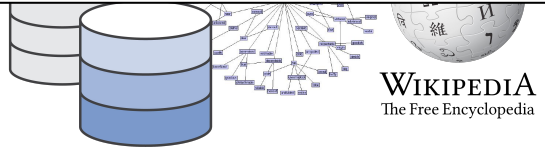
- ★ There is still a huge interest in solving NER
- ★ All major tech companies are working on solving this problem
  - Funding opportunities available

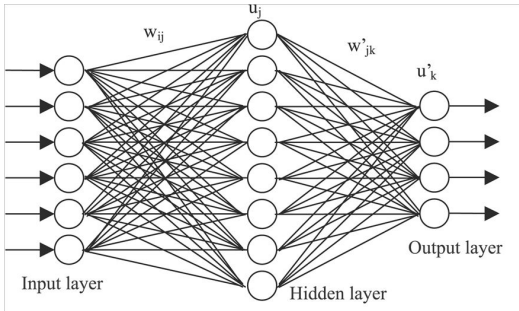
2015

Embeddings + Traditional ML

2017

Embeddings + Neural Networks





UNIVERSITY of HOUSTON

Short History of NER

★ PhD thesis: NER for low resourced languages

2006 2010 2015 2017

Rule-based Approaches

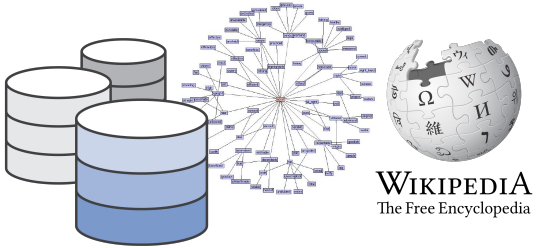
Traditional Machine Learning

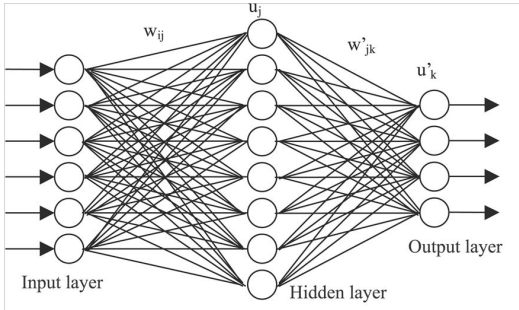
Exploiting Web-based Resources

Social media data

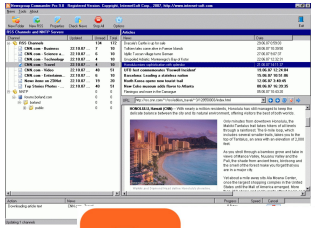
Embeddings + Traditional ML

Embeddings + Neural Networks





UNIVERSITY of HOUSTON



Blogger

~1996 — Rule-based Approaches

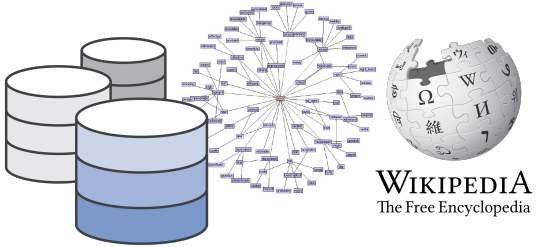
2003 — Traditional Machine Learning

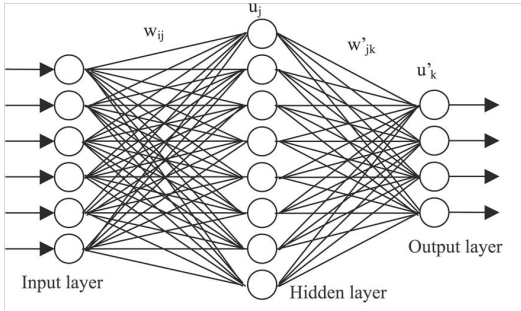
2006 — Exploiting Web-based Resources

2010 — Social media data

2015 — Embeddings + Traditional ML

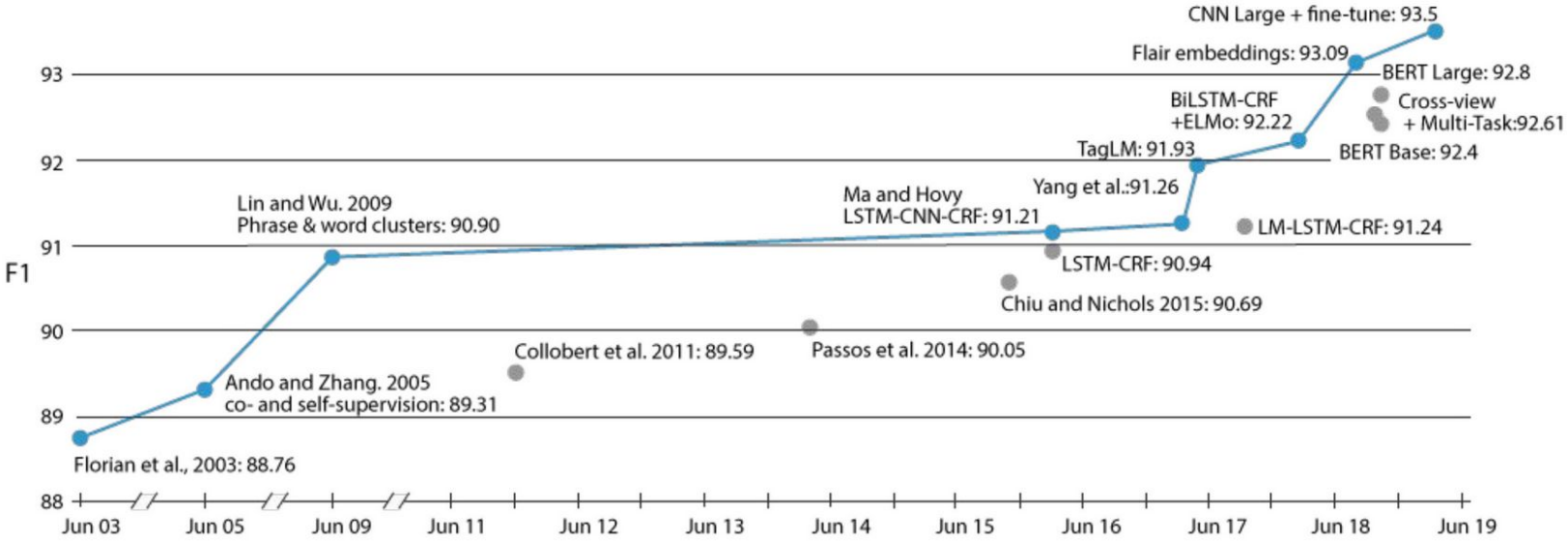
2017 — Embeddings + Neural Networks





UNIVERSITY of HOUSTON

Breakthroughs in Named Entity Recognition (NER)



Performance on Named Entity Recognition (NER) on CoNLL-2003 (English) over time.

UNIVERSITY of HOUSTON

Image borrowed from Sebastian Ruder (http://ruder.io/state-of-transfer-learning-in-nlp/)

SOTA Comparisons

Task	Language	Dataset	Flair	Previous best
Named Entity Recognition	English	Conll-03	**93.18** (F1)	*92.22 (Peters et al., 2018)*
Named Entity Recognition	English	Ontonotes	**89.3** (F1)	*86.28 (Chiu et al., 2016)*
Emerging Entity Detection	English	WNUT-17	**49.49** (F1)	*45.55 (Aguilar et al., 2018)*
Part-of-Speech tagging	English	WSJ	**97.85**	*97.64 (Choi, 2016)*
Chunking	English	Conll-2000	**96.72** (F1)	*96.36 (Peters et al., 2017)*
Named Entity Recognition	German	Conll-03	**88.27** (F1)	*78.76 (Lample et al., 2016)*
Named Entity Recognition	German	Germeval	**84.65** (F1)	*79.08 (Hänig et al, 2014)*
Named Entity Recognition	Dutch	Conll-03	**90.44** (F1)	*81.74 (Lample et al., 2016)*
Named Entity Recognition	Polish	PolEval-2018	**86.6** (F1) (Borchmann et al., 2018)	*85.1 (PolDeepNer)*

Other Applications

- Job postings
- Job resumes
- Seminar announcements
- Company information from the web
- Continuing education course info from the web
- University information from the web
- Apartment rental ads
- Molecular biology information from MEDLINE

How Difficult is NER?

Name	Possible Categories
Washington	Person, Location, Political Entity, Organization, Facility
Downing St.	Location, Organization
IRA	Person, Organization, Monetary Instrument
Louis Vuitton	Person, Organization, Commercial Product

[$_{PERS}$ Washington] was born into slavery on the farm of James Burroughs.
[$_{ORG}$ Washington] went up 2 games to 1 in the four-game series.
Blair arrived in [$_{LOC}$ Washington] for what may well be his last state visit.
In June, [$_{GPE}$ Washington] passed a primary seatbelt law.
The [$_{FAC}$ Washington] had proved to be a leaky ship, every passage I made...

IE as Sequence Labeling

- Can extract features describing each token in the text.
- Can apply a sliding window classifier using various classification algorithms.
- Can apply probabilistic sequence models:
  - HMM
  - CRF (Conditional Random Fields)

Sequence Labeling for NER

Words	Label
American	B_{ORG}
Airlines	I_{ORG}
,	O
a	O
unit	O
of	O
AMR	B_{ORG}
Corp.	I_{ORG}
,	O
immediately	O
matched	O
the	O
move	O
,	O
spokesman	O
Tim	B_{PERS}
Wagner	I_{PERS}
said	O
.	O

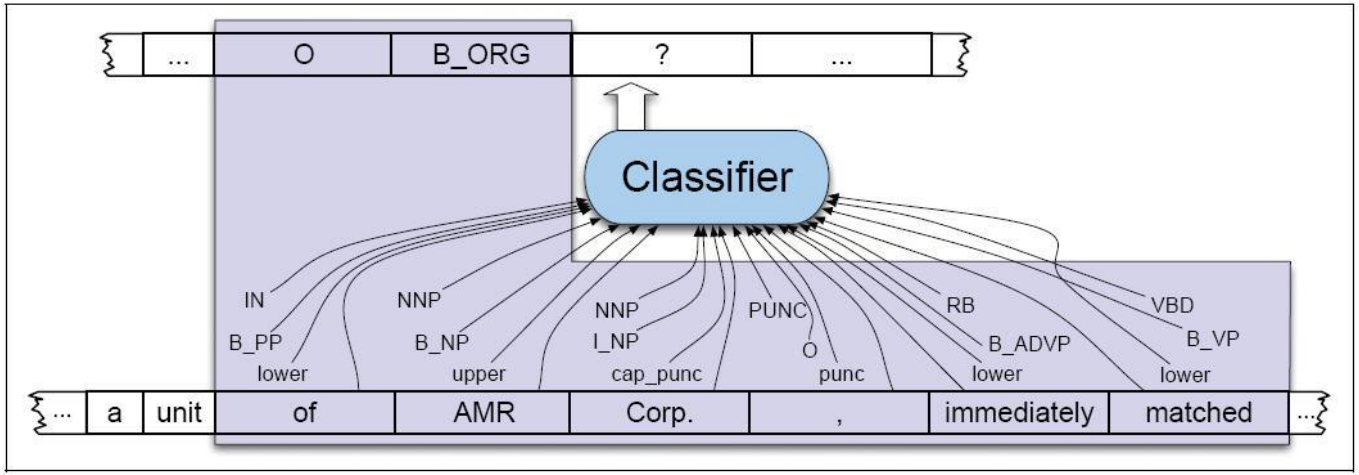
Typical features

- Lexical items
- Shape features
- Gazetteers
- Stemmed lexical item
- POS tags
- Trigger words

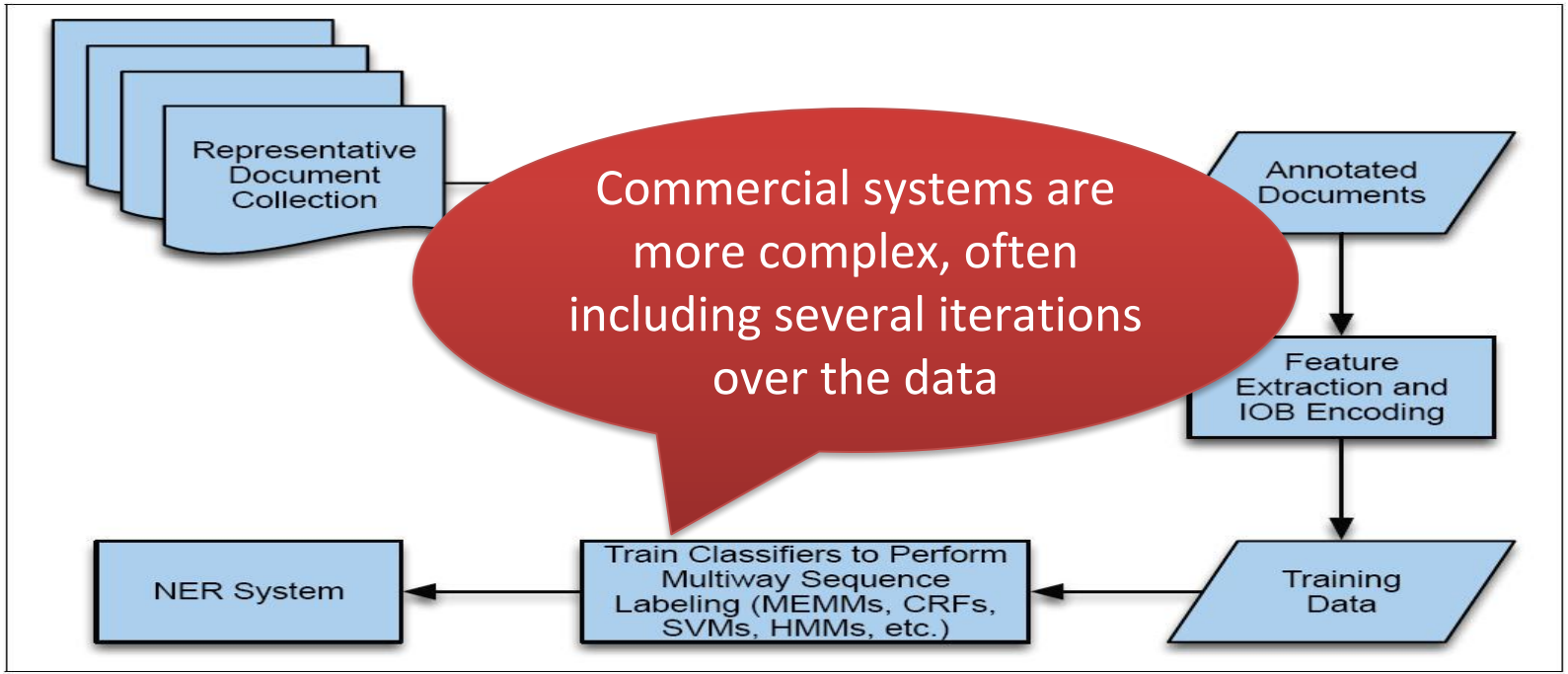
Sequence Labeling for NER

Features				Label
American	NNP	B_{NP}	cap	B_{ORG}
Airlines	NNPS	I_{NP}	cap	I_{ORG}
,	PUNC	O	punc	O
a	DT	B_{NP}	lower	O
unit	NN	I_{NP}	lower	O
of	IN	B_{PP}	lower	O
AMR	NNP	B_{NP}	upper	B_{ORG}
Corp.	NNP	I_{NP}	cap_punc	I_{ORG}
,	PUNC	O	punc	O
immediately	RB	B_{ADVP}	lower	O
matched	VBD	B_{VP}	lower	O
the	DT	B_{NP}	lower	O
move	NN	I_{NP}	lower	O
,	PUNC	O	punc	O
spokesman	NN	B_{NP}	lower	O
Tim	NNP	I_{NP}	cap	B_{PER}
Wagner	NNP	I_{NP}	cap	I_{PER}
said	VBD	B_{VP}	lower	O
.	PUNC	O	punc	O

Sequence Labeling for NER



Sequence Labeling for NER



Graphical Models in NLP

- ★ Often times the task itself suggests conditional independence assumptions
- ★ Graphical models (GMs) allow a factorization of the probability density to a specific set of conditional independence assumptions.
- ★ GMs model p(x,y) over inputs and outputs:
  - HMMs:

$$P(O,Q) = P(O|Q) \times P(Q) = \prod_{i=1}^{n} P(o_i|q_i) \times \prod_{i=1}^{n} P(q_i|q_{i-1})$$

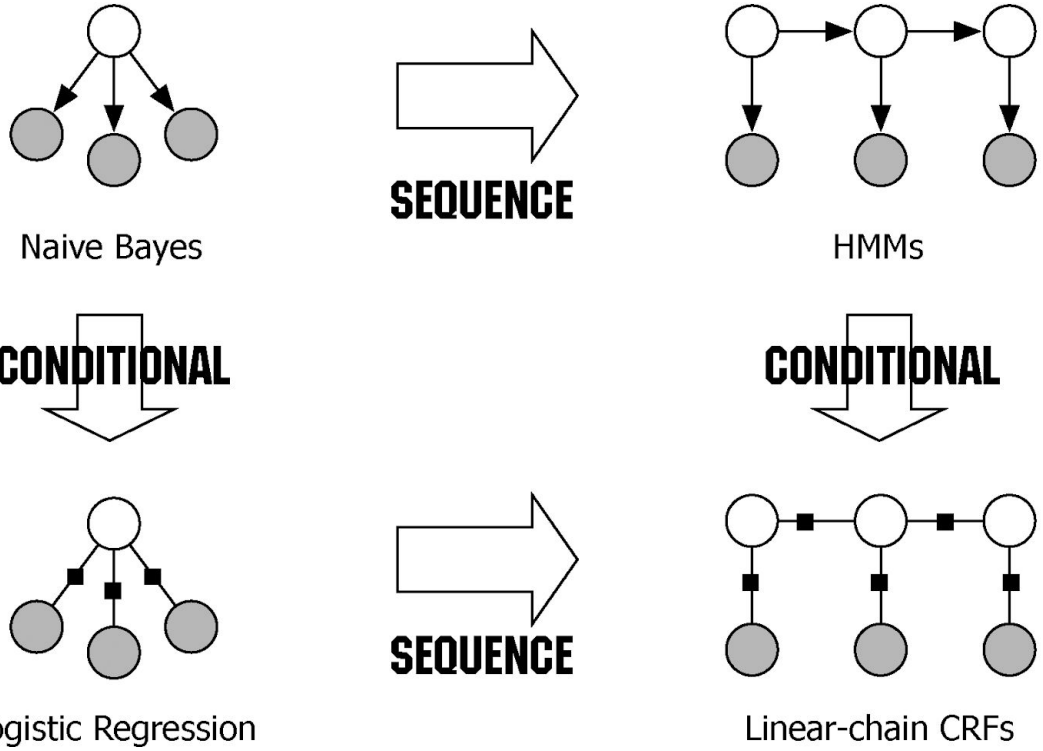
- ★ Discriminative models directly describe how to take a feature vector x and assign it a label y

Undirected Graphical Models

$$p(\mathbf{x}, \mathbf{y}) = \frac{1}{Z} \prod_{a=1}^{A} \Psi_a(\mathbf{x}_a, \mathbf{y}_a)$$

$$Z = \sum_{\mathbf{x},\mathbf{y}} \prod_{a \in F} \Psi_a(\mathbf{x}_a, \mathbf{y}_a)$$

NB, LR, HMMs, CRFs



Conditional Random Fields

Goal: determine the best sequence $y \in C^n$ of classes, given an input sequence x of length n:

$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y} \in C^n} P(\mathbf{y}|\mathbf{x})$$

- ★ This probability depends on potential functions (features)
- ★ Features are usually indicator functions that will be weighted

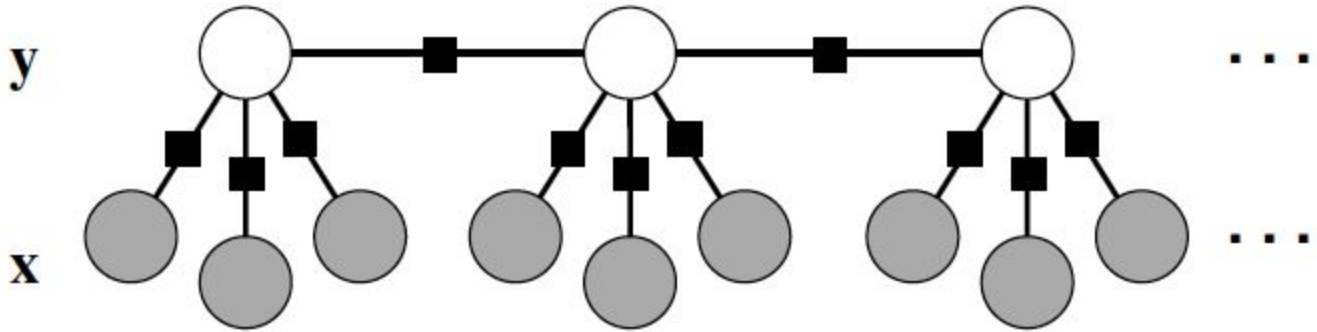
Sample features

$$t(\mathbf{y}_{i-1}, \mathbf{y}_i, \mathbf{x}, i) = \begin{cases} 1 \text{ if } \mathbf{x}_i = \text{“September”} \text{ and } \mathbf{y}_{i-1} = \text{IN} \text{ and } \mathbf{y}_i = \text{NNP} \\ 0 \text{ otherwise} \end{cases}$$

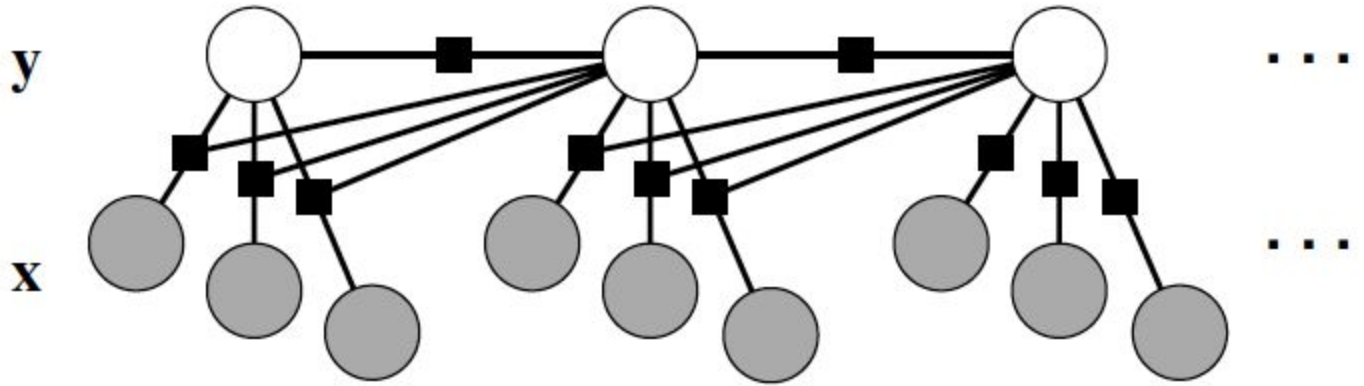
(taken from Wallach (2004))

$$s(\mathbf{y}_i, \mathbf{x}, i) = \begin{cases} 1 \text{ if } \mathbf{x}_i = \text{“to”} \text{ and } \mathbf{y}_i = \text{TO} \\ 0 \text{ otherwise} \end{cases}$$

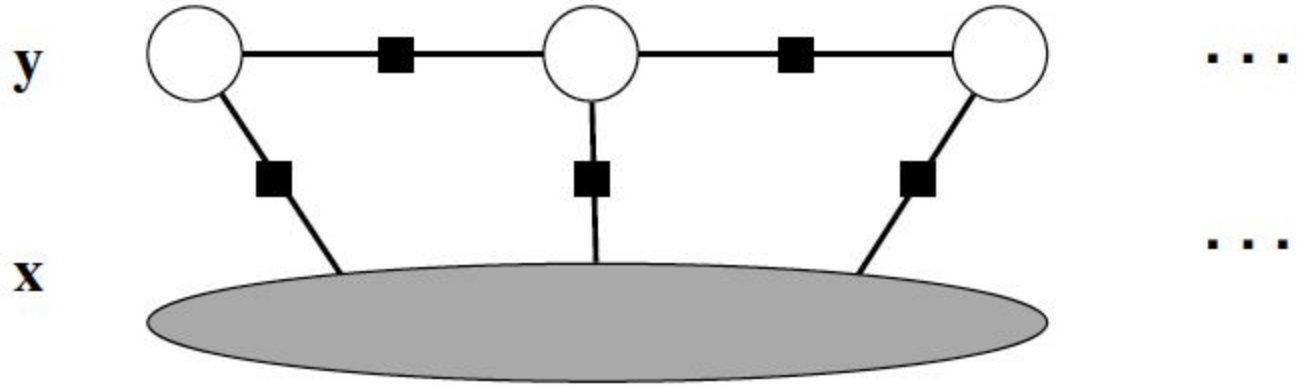
Linear-chain Conditional Random Fields



Linear-chain Conditional Random Fields

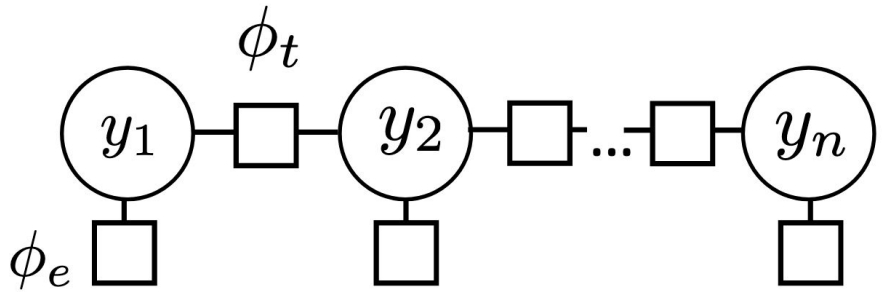


Linear-chain Conditional Random Fields



Sequential CRFs

$$P(\mathbf{y}|\mathbf{x}) = \frac{1}{Z} \prod_{i=2}^{n} \exp(\phi_t(y_{i-1}, y_i)) \prod_{i=1}^{n} \exp(\phi_e(y_i, i, \mathbf{x}))$$



Sequential CRFs

$$\phi_e(y_i, i, \mathbf{x}) = w^\top f_e(y_i, i, \mathbf{x}) \quad \phi_t(y_{i-1}, y_i) = w^\top f_t(y_{i-1}, y_i)$$

$$P(\mathbf{y}|\mathbf{x}) \propto \exp w^\top \left[\sum_{i=2}^{n} f_t(y_{i-1}, y_i) + \sum_{i=1}^{n} f_e(y_i, i, \mathbf{x}) \right]$$

Basic Features for NER

$$P(\mathbf{y}|\mathbf{x}) \propto \exp w^\top \left[\sum_{i=2}^{n} f_t(y_{i-1}, y_i) + \sum_{i=1}^{n} f_e(y_i, i, \mathbf{x}) \right]$$

O B-LOC

Barack Obama will travel to Hangzhou today for the G20 meeting .

Transitions: $f_t(y_{i-1}, y_i) = \text{Ind}[y_{i-1} \ \& \ y_i]$ = Ind[O — B-LOC]

Emissions: $f_e(y_6, 6, \mathbf{x}) =$ Ind[B-LOC & Current word = *Hangzhou*]

Ind[B-LOC & Prev word = *to*]

Basic Features for NER

$$P(\mathbf{y}|\mathbf{x}) \propto \exp w^\top \left[\sum_{i=2}^{n} f_t(y_{i-1}, y_i) + \sum_{i=1}^{n} f_e(y_i, i, \mathbf{x}) \right]$$

O B-LOC

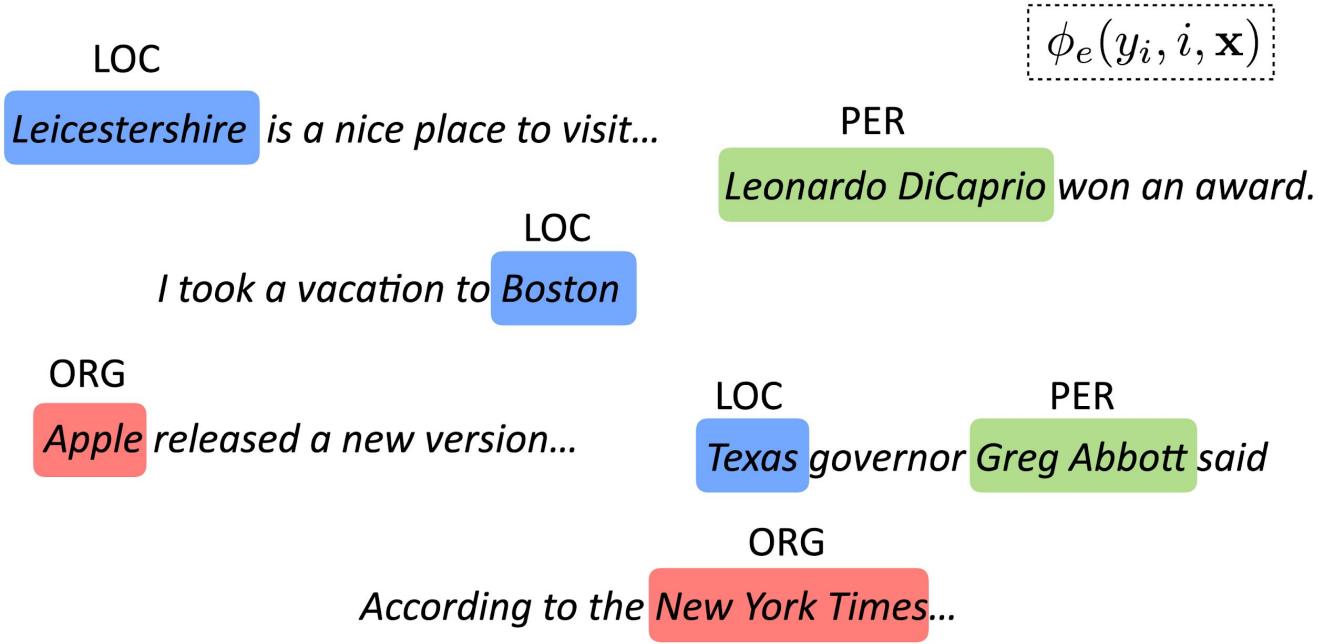
Barack Obama will travel to Hangzhou today for the G20 meeting .

Transitions: $f_t(y_{i-1}, y_i) = \text{Ind}[y_{i-1} \ \& \ y_i]$ = Ind[O — B-LOC]

Emissions: $f_e(y_6, 6, \mathbf{x}) =$ Ind[B-LOC & Current word = *Hangzhou*]

Ind[B-LOC & Prev word = *to*]

Emission Features



Emission Features

- Word features (can use in HMM)
- Capitalization
- Word shape
- Prefixes/suffixes
- Lexical indicators
- Context features (can’t use in HMM!)
  - Words before/after
  - Tags before/after
- Gazeteers
- Word clusters

So far...

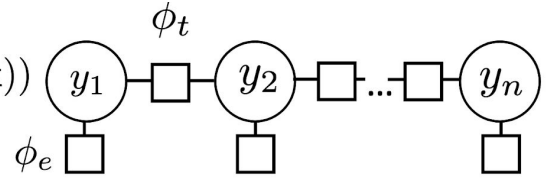
- ★ Motivation for GMs
- ★ NER problem definition
- ★ CRFs model definition

Now we need:

- ★ Inference
- ★ Training

Inference in CRFs

$$P(\mathbf{y}|\mathbf{x}) = \frac{1}{Z}\prod_{i=2}^{n}\exp(\phi_t(y_{i-1},y_i))\prod_{i=1}^{n}\exp(\phi_e(y_i,i,\mathbf{x}))$$



- $\text{argmax}_{\mathbf{y}} P(\mathbf{y}|\mathbf{x})$: can use Viterbi exactly as in HMM case

$$\max_{y_1,\ldots,y_n} e^{\phi_t(y_{n-1},y_n)}e^{\phi_e(y_n,n,\mathbf{x})}\ldots e^{\phi_e(y_2,2,\mathbf{x})}e^{\phi_t(y_1,y_2)}e^{\phi_e(y_1,1,\mathbf{x})}$$

$$= \max_{y_2,\ldots,y_n} e^{\phi_t(y_{n-1},y_n)}e^{\phi_e(y_n,n,\mathbf{x})}\ldots e^{\phi_e(y_2,2,\mathbf{x})}\max_{y_1} e^{\phi_t(y_1,y_2)}e^{\phi_e(y_1,1,\mathbf{x})}$$

$$= \max_{y_3,\ldots,y_n} e^{\phi_t(y_{n-1},y_n)}e^{\phi_e(y_n,n,\mathbf{x})}\ldots \max_{y_2} e^{\phi_t(y_2,y_3)}e^{\phi_e(y_2,2,\mathbf{x})}\max_{y_1} e^{\phi_t(y_1,y_2)}\text{score}_1(y_1)$$

- $\exp(\phi_t(y_{i-1},y_i))$ and $\exp(\phi_e(y_i,i,\mathbf{x}))$ play the role of the Ps now, same dynamic program

Parameter Estimation in CRFs

$$P(\mathbf{y}|\mathbf{x}) \propto \exp w^\top \left[\sum_{i=2}^{n} f_t(y_{i-1}, y_i) + \sum_{i=1}^{n} f_e(y_i, i, \mathbf{x}) \right]$$

- Logistic regression: $P(y|x) \propto \exp w^\top f(x, y)$
- Maximize $\mathcal{L}(\mathbf{y}^*, \mathbf{x}) = \log P(\mathbf{y}^*|\mathbf{x})$
- Gradient is completely analogous to logistic regression:

$$\frac{\partial}{\partial w} \mathcal{L}(\mathbf{y}^*, \mathbf{x}) = \sum_{i=2}^{n} f_t(y^*_{i-1}, y^*_i) + \sum_{i=1}^{n} f_e(y^*_i, i, \mathbf{x})$$

intractable! $-\mathbb{E}_{\mathbf{y}} \left[\sum_{i=2}^{n} f_t(y_{i-1}, y_i) + \sum_{i=1}^{n} f_e(y_i, i, \mathbf{x}) \right]$

Parameter Estimation in CRFs

$$\frac{\partial}{\partial w}\mathcal{L}(\mathbf{y}^*, \mathbf{x}) = \sum_{i=2}^{n} f_t(y^*_{i-1}, y^*_i) + \sum_{i=1}^{n} f_e(y^*_i, i, \mathbf{x}) - \mathbb{E}_{\mathbf{y}}\left[\sum_{i=2}^{n} f_t(y_{i-1}, y_i) + \sum_{i=1}^{n} f_e(y_i, i, \mathbf{x})\right]$$

- Let's focus on emission feature expectation

$$\mathbb{E}_{\mathbf{y}}\left[\sum_{i=1}^{n} f_e(y_i, i, \mathbf{x})\right] = \sum_{\mathbf{y}\in\mathcal{Y}} P(\mathbf{y}|\mathbf{x})\left[\sum_{i=1}^{n} f_e(y_i, i, \mathbf{x})\right] = \sum_{i=1}^{n}\sum_{\mathbf{y}\in\mathcal{Y}} P(\mathbf{y}|\mathbf{x}) f_e(y_i, i, \mathbf{x})$$

$$= \sum_{i=1}^{n}\sum_{s} P(y_i = s|\mathbf{x}) f_e(s, i, \mathbf{x})$$

Parameter Estimation in CRFs

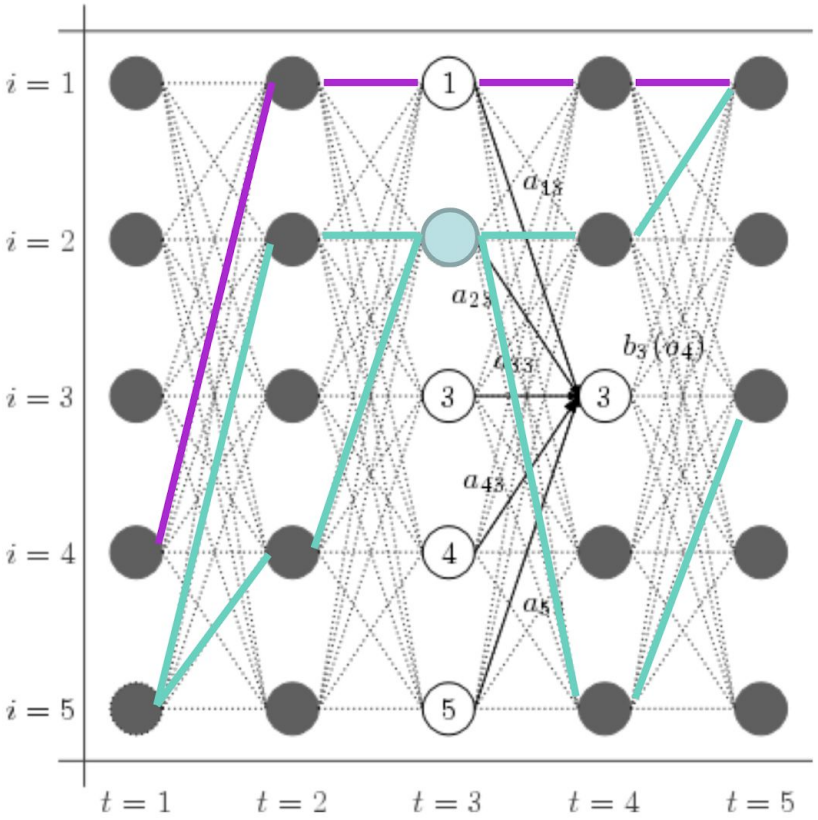
How do we compute these marginals $P(y_i = s|\mathbf{x})$?

$$P(y_i = s|\mathbf{x}) = \sum_{y_1,\cdots,y_{i-1},y_{i+1},\cdots,y_n} P(\mathbf{y}|\mathbf{x})$$

What did Viterbi compute? $P(\mathbf{y}_{\max}|\mathbf{x}) = \max_{y_1,\cdots,y_n} P(\mathbf{y}|\mathbf{x})$

Can compute marginals with dynamic programming as well using forward-backward

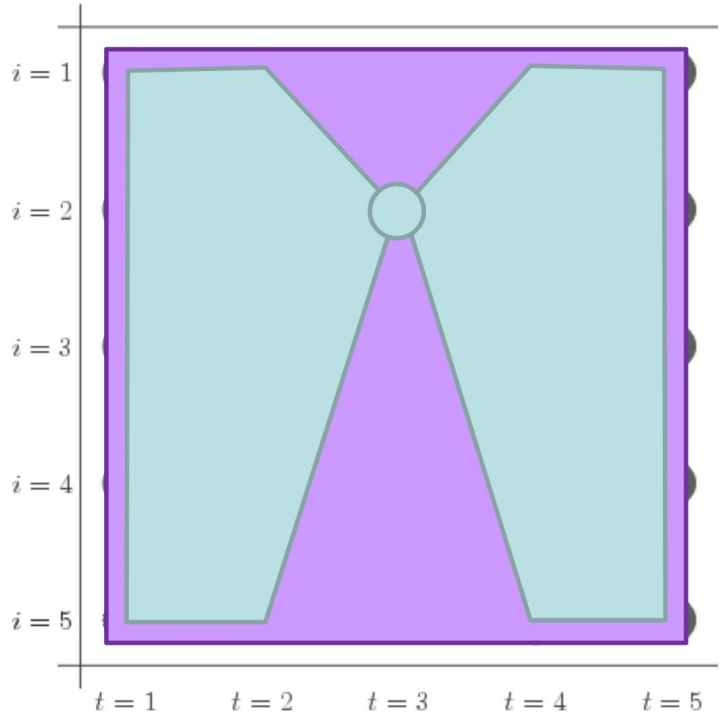
Forward-Backward Algorithm



$$P(y_3 = 2|\mathbf{x}) =$$

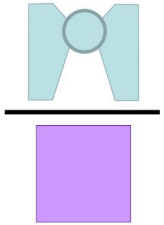
$$\frac{\text{sum of all paths through state 2 at time 3}}{\text{sum of all paths}}$$

Forward-Backward Algorithm



$$P(y_3 = 2|\mathbf{x}) =$$

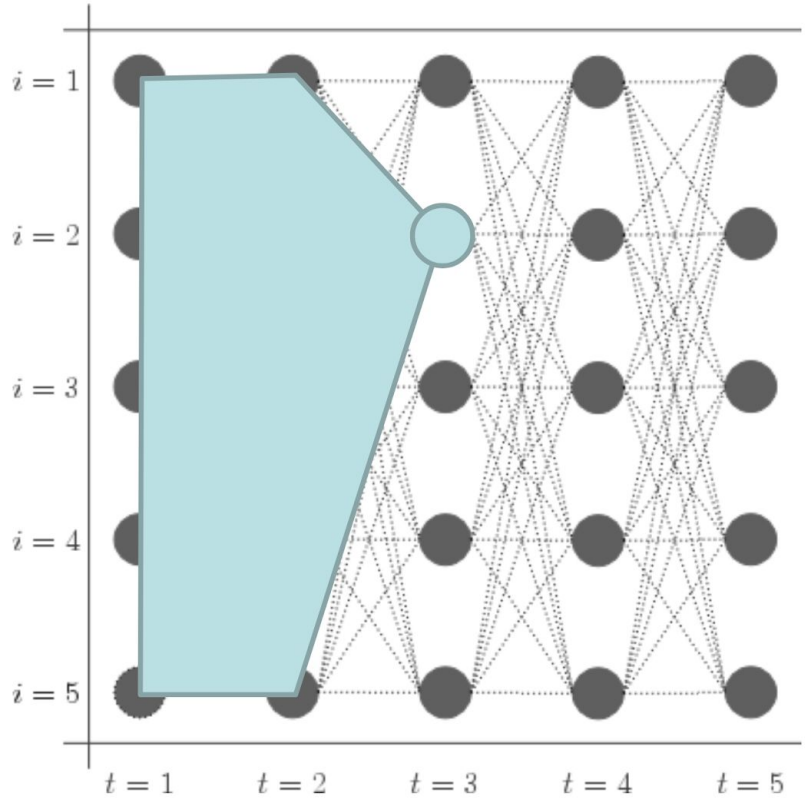
$$\frac{\text{sum of all paths through state 2 at time 3}}{\text{sum of all paths}}$$

$=$ 

- Easiest and most flexible to do one pass to compute  and one to compute 

slide credit: Dan Klein

Forward-Backward Algorithm



- Initial:

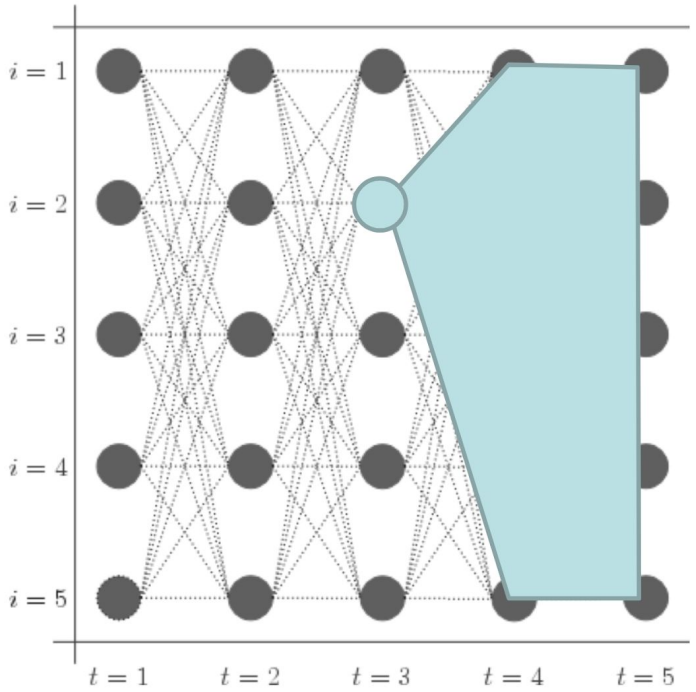
$$\alpha_1(s) = \exp(\phi_e(s, 1, \mathbf{x}))$$

- Recurrence:

$$\alpha_t(s_t) = \sum_{s_{t-1}} \alpha_{t-1}(s_{t-1}) \exp(\phi_e(s_t, t, \mathbf{x})) \exp(\phi_t(s_{t-1}, s_t))$$

- Same as Viterbi but summing instead of maxing!
- These quantities get very small! Store everything as log probabilities

Forward-Backward Algorithm



- Initial:

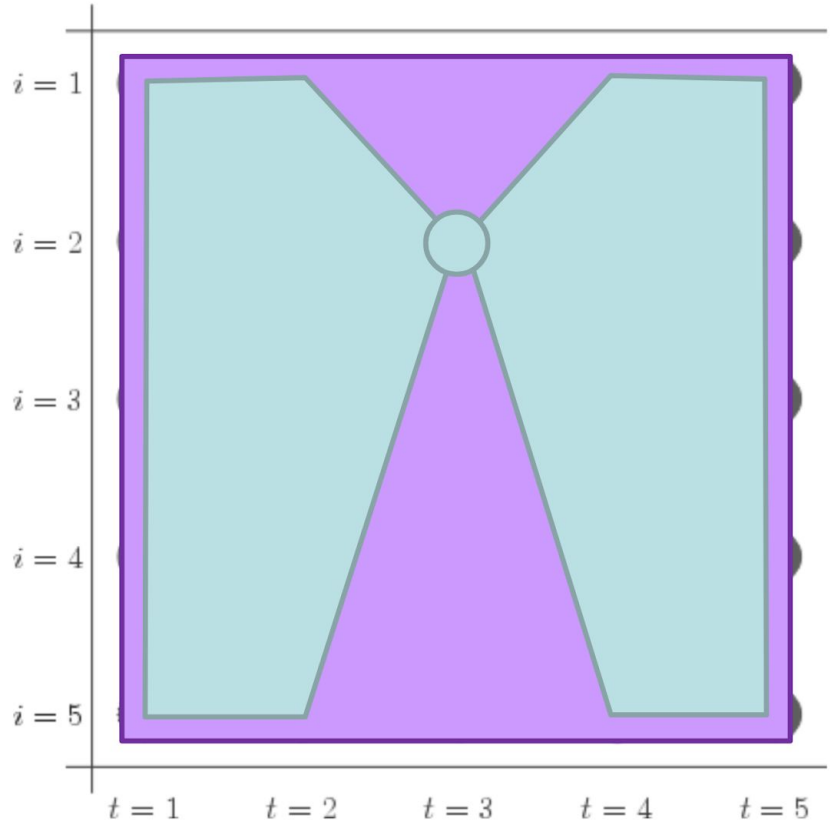
$$\beta_n(s) = 1$$

- Recurrence:

$$\beta_t(s_t) = \sum_{s_{t+1}} \beta_{t+1}(s_{t+1}) \exp(\phi_e(s_{t+1}, t+1, \mathbf{x})) \exp(\phi_t(s_t, s_{t+1}))$$

- Big differences: count emission for the *next* timestep (not current one)

Forward-Backward Algorithm



$$\alpha_1(s) = \exp(\phi_e(s, 1, \mathbf{x}))$$

$$\alpha_t(s_t) = \sum_{s_{t-1}} \alpha_{t-1}(s_{t-1}) \exp(\phi_e(s_t, t, \mathbf{x})) \exp(\phi_t(s_{t-1}, s_t))$$

$$\beta_n(s) = 1$$

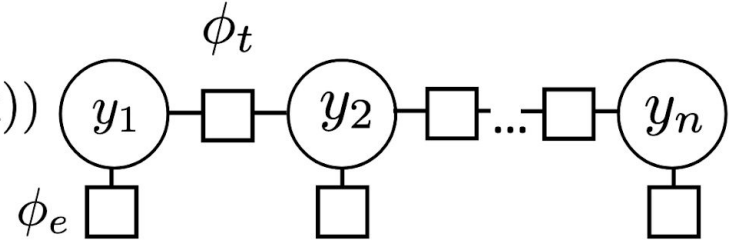
$$\beta_t(s_t) = \sum_{s_{t+1}} \beta_{t+1}(s_{t+1}) \exp(\phi_e(s_{t+1}, t+1, \mathbf{x})) \exp(\phi_t(s_t, s_{t+1}))$$

$$P(s_3 = 2|\mathbf{x}) = \frac{\alpha_3(2)\beta_3(2)}{\sum_i \alpha_3(i)\beta_3(i)}$$

- Does this explain why beta is what it is?

Computing Marginals

$$P(\mathbf{y}|\mathbf{x}) = \frac{1}{Z} \prod_{i=2}^{n} \exp(\phi_t(y_{i-1}, y_i)) \prod_{i=1}^{n} \exp(\phi_e(y_i, i, \mathbf{x}))$$



- Normalizing constant $Z = \sum_{\mathbf{y}} \prod_{i=2}^{n} \exp(\phi_t(y_{i-1}, y_i)) \prod_{i=1}^{n} \exp(\phi_e(y_i, i, \mathbf{x}))$
- Analogous to P(**x**) for HMMs
- For both HMMs and CRFs:

$$P(y_i = s|\mathbf{x}) = \frac{\text{forward}_i(s)\text{backward}_i(s)}{\sum_{s'} \text{forward}_i(s')\text{backward}_i(s')}$$

Z for CRFs, P(**x**) for HMMs

Parameter Estimation in CRFs

For emission features:

$$\frac{\partial}{\partial w}\mathcal{L}(\mathbf{y}^*, \mathbf{x}) = \sum_{i=1}^{n} f_e(y_i^*, i, \mathbf{x}) - \sum_{i=1}^{n}\sum_{s} P(y_i = s|\mathbf{x}) f_e(s, i, \mathbf{x})$$

gold features — expected features under model

Transition features: need to compute $P(y_i = s_1, y_{i+1} = s_2|\mathbf{x})$ using forward-backward as well

Pseudocode

```
for each epoch
    for each example
        extract features on each emission and transition (look up in cache)
        compute potentials phi based on features + weights
        compute marginal probabilities with forward-backward
        accumulate gradient over all emissions and transitions
```

Evaluating NER Accuracy

- ★ Always evaluate performance on independent, manually-annotated test data not used during system development.
- ★ Compute metrics reviewed earlier:
  - ○ Recall
  - ○ Precision
  - ○ F-Measure = Harmonic mean of recall and precision

    2 x Precision x Recall / (Precision + Recall)

In Summary

- ★ NER task
- ★ Conditional Random Fields:
  - Indicator Functions
  - Features for NER
  - Inference
  - Training