L101: Machine Learning for Language Processing

Lecture 3

Guy Emerson

Today's Lecture

- Discriminative Models
  - Logistic Regression
  - Maximum Entropy Markov Model
  - Conditional Random Field
- Named Entity Recognition

Recap – Models

- Generative – $P(x, y)$
- Discriminative – $P(y|x)$

Recap – Naive Bayes

$$\operatorname*{argmax}_y P(y|x) = \operatorname*{argmax}_y P(y)\,P(x|y)$$

$$\approx \operatorname*{argmax}_y P(y)\prod_i P(x_i|y)$$

Recap – Naive Bayes

$$\begin{aligned}\operatorname*{argmax}_y P(y|x) &= \operatorname*{argmax}_y P(y)\,P(x|y) \\ &\approx \operatorname*{argmax}_y P(y)\prod_i P(x_i|y)\end{aligned}$$

Recap – Naive Bayes

$$\operatorname*{argmax}_y P(y|x) = \operatorname*{argmax}_y P(y)\,P(x|y)$$

$$\approx \operatorname*{argmax}_y P(y)\prod_i P(x_i|y)$$

Discriminative – approximate $P(y|x)$?

Logistic Regression

$$P(y|x) \approx \frac{1}{Z} \exp\left(\sum_i \theta_{y,i} x_i\right)$$

Logistic Regression

$$P(y|x) \approx \frac{1}{Z} \exp\left(\sum_i \theta_{y,i} x_i\right)$$

$$= \frac{\exp\left(\sum_i \theta_{y,i} x_i\right)}{\sum_{y'} \exp\left(\sum_i \theta_{y',i} x_i\right)}$$

Logistic Regression

$$P(y|x) \approx \frac{1}{Z} \exp\left(\theta_y + \sum_i \theta_{y,i} x_i\right)$$

$$= \frac{\exp\left(\theta_y + \sum_i \theta_{y,i} x_i\right)}{\sum_{y'} \exp\left(\theta_{y'} + \sum_i \theta_{y',i} x_i\right)}$$

Logistic Regression

$$P(y|x) \approx \frac{1}{Z} \exp\left(\theta_y + \sum_i \theta_{y,i} x_i\right)$$

$$= \frac{\exp\left(\theta_y + \sum_i (\theta_{y,i} + k) x_i\right)}{\sum_{y'} \exp\left(\theta_{y'} + \sum_i (\theta_{y',i} + k) x_i\right)}$$

Logistic Regression

$$P(y|x) \approx \frac{1}{Z} \exp\left(\theta_y + \sum_i \theta_{y,i} x_i\right)$$

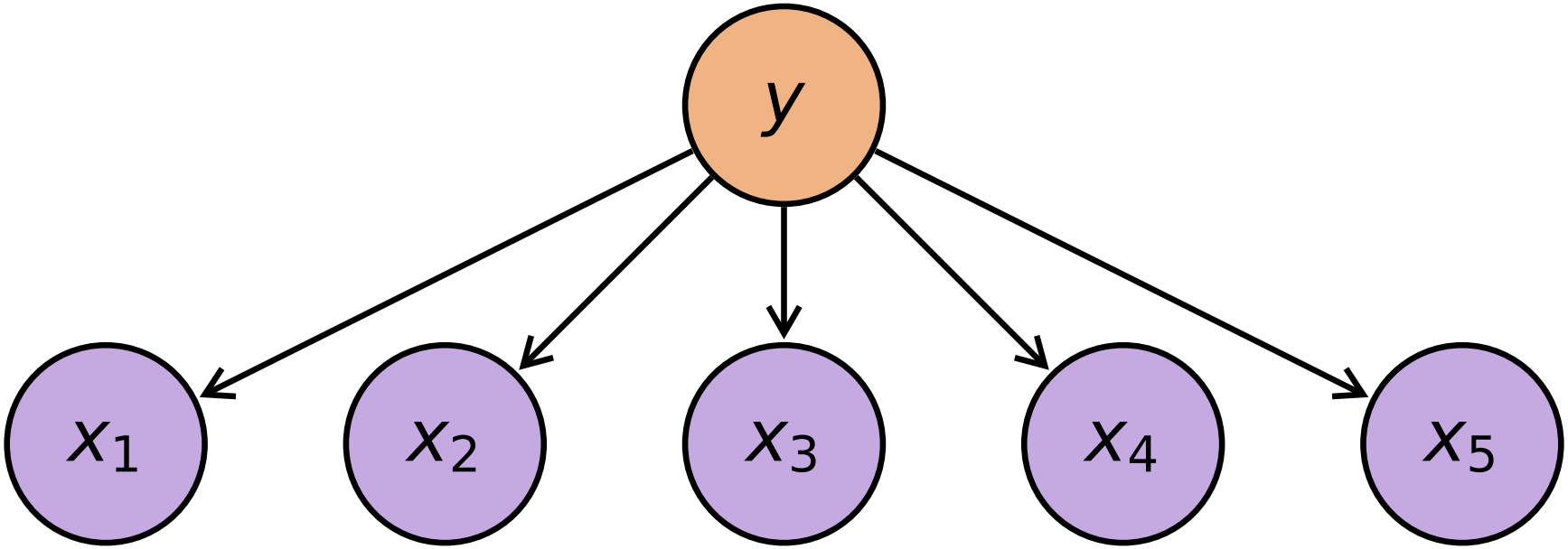
$$= \frac{e^{kx_i} \exp\left(\theta_y + \sum_i \theta_{y,i} x_i\right)}{e^{kx_i} \sum_{y'} \exp\left(\theta_{y'} + \sum_i \theta_{y',i} x_i\right)}$$

Logistic Regression

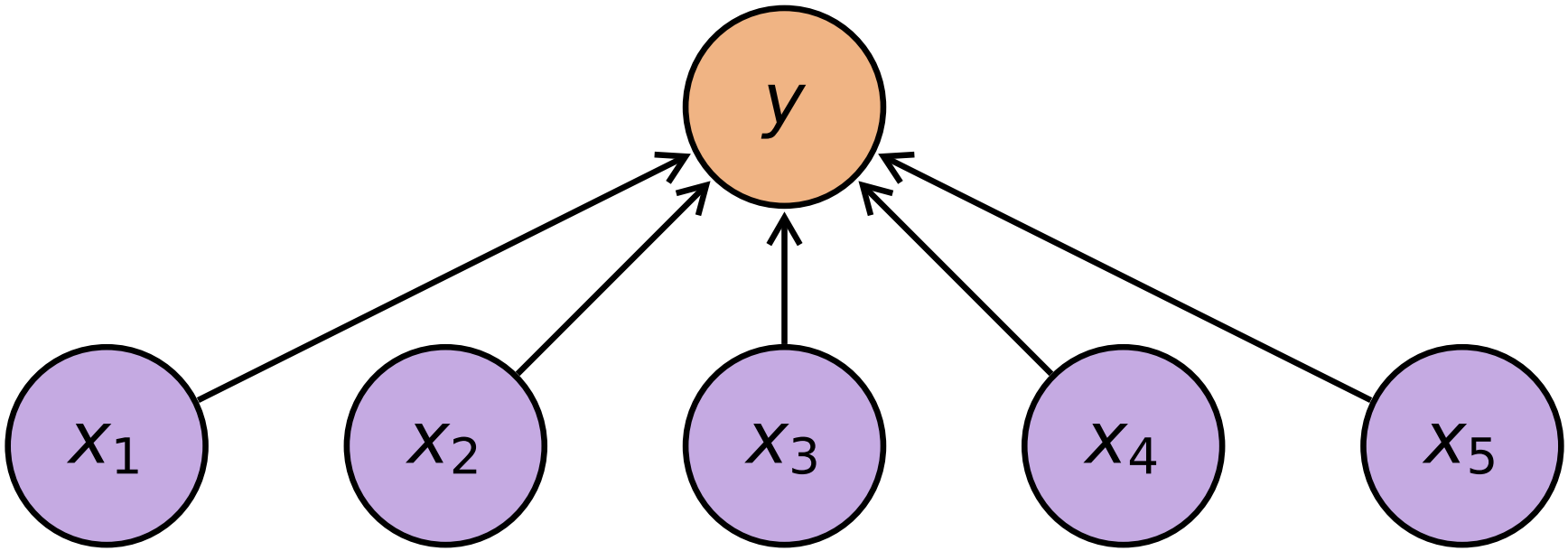
$$P(y|x) \approx \frac{1}{Z} \exp\left(\theta_y + \sum_i \theta_{y,i} x_i\right)$$

$$= \frac{\exp\left(\theta_y + \sum_i \theta_{y,i} x_i\right)}{\sum_{y'} \exp\left(\theta_{y'} + \sum_i \theta_{y',i} x_i\right)}$$

Naive Bayes



Logistic Regression



Logistic Regression

- Parameters: θ_y, $\theta_{y,i}$

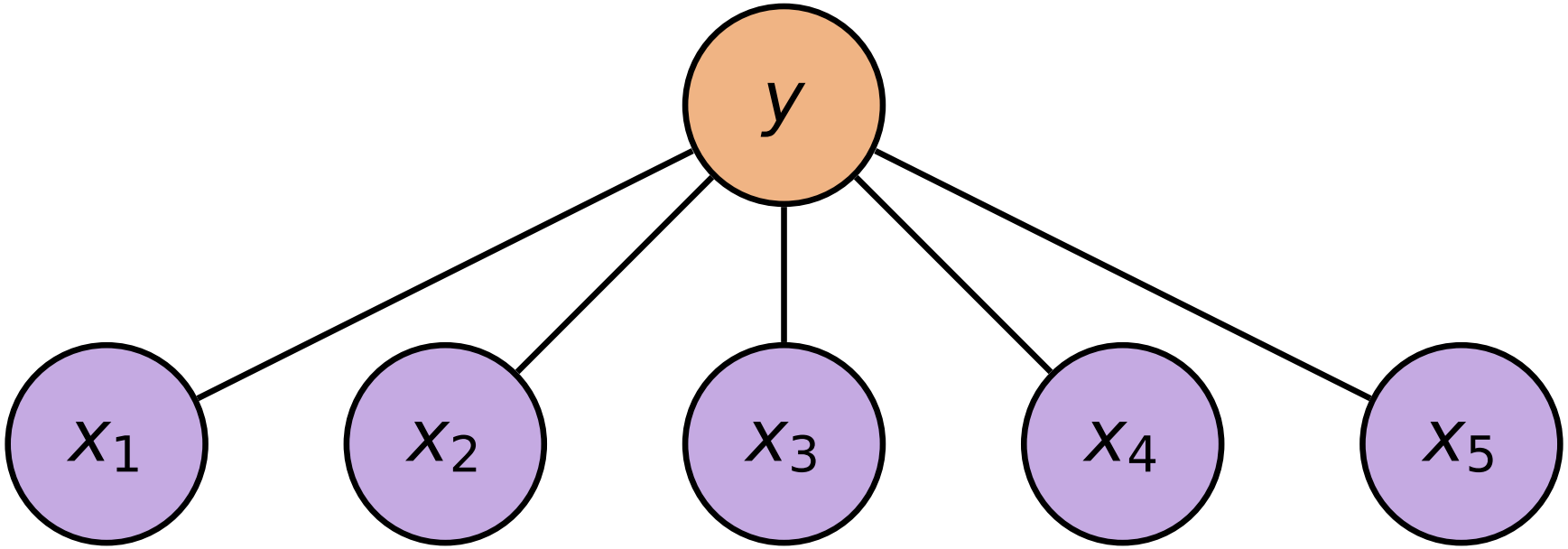
Logistic Regression

- Parameters: θ_y, $\theta_{y,i}$
- Optimise for: $\sum_{(x,y)\in D} \log P(y|x)$

Logistic Regression

- Parameters: θ_y, $\theta_{y,i}$
- Optimise for: $\sum_{(x,y)\in D} \log P(y|x)$
- No closed form formula!

Independence of Features



Independence of Features

- Hong Kong vs. HongKong

Independence of Features

- Hong Kong vs. HongKong
- Naive Bayes:
  - $P(x_i|y)$ same
  - $P(y|x)$ over-estimated

Independence of Features

- Hong Kong vs. HongKong
- Naive Bayes:
  - $P(x_i|y)$ same
  - $P(y|x)$ over-estimated
- Logistic Regression:
  - $P(y|x)$ same
  - $P(x_i|y)$ never used!

Logistic Regression

$$P(y|x) \approx \frac{1}{Z} \exp\left(\sum_i \theta_{y,i} x_i\right)$$

Why Log-Linear?

- Consider all distributions $P(y|x)$

Why Log-Linear?

- Consider all distributions $P(y|x)$
- Under constraints:
  - $P(y|x_i)$ matches observed data

Why Log-Linear?

- Consider all distributions $P(y|x)$
- Under constraints:
  - $P(y|x_i)$ matches observed data
- Maximise conditional entropy $H(Y|X)$ on observed data

Why Log-Linear?

- Consider all distributions $P(y|x)$
- Under constraints:
  - $P(y|x_i)$ matches observed data
- Maximise conditional entropy $H(Y|X)$ on observed data

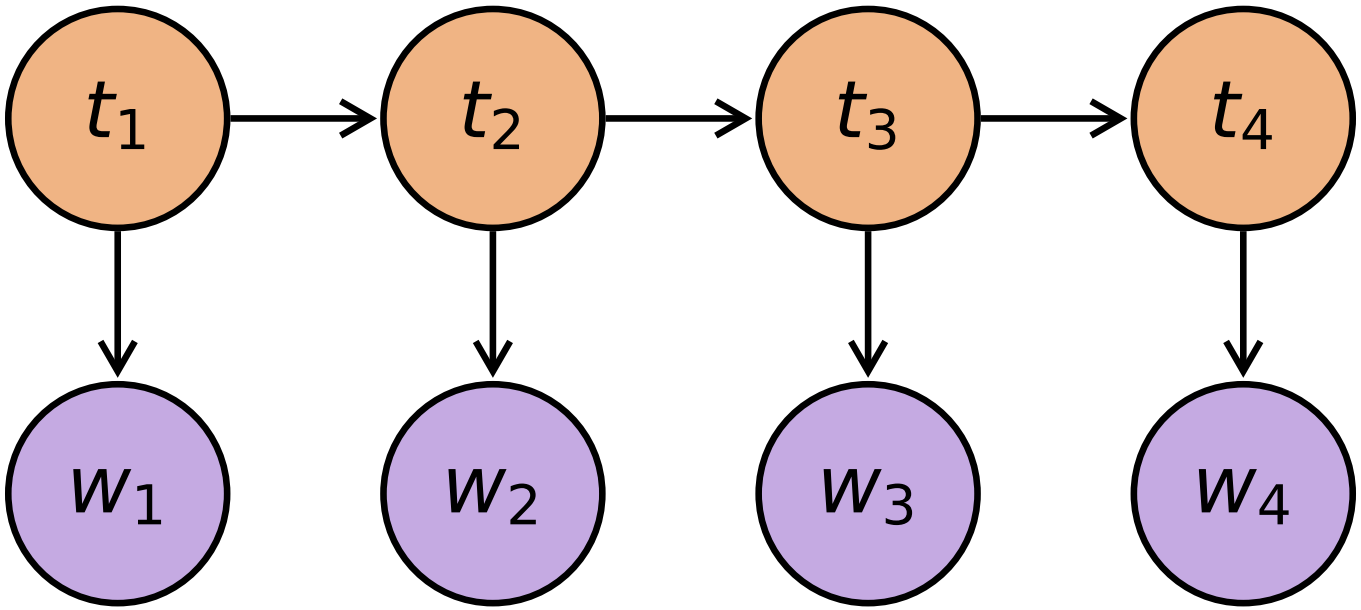
→ Logistic regression

Regularisation

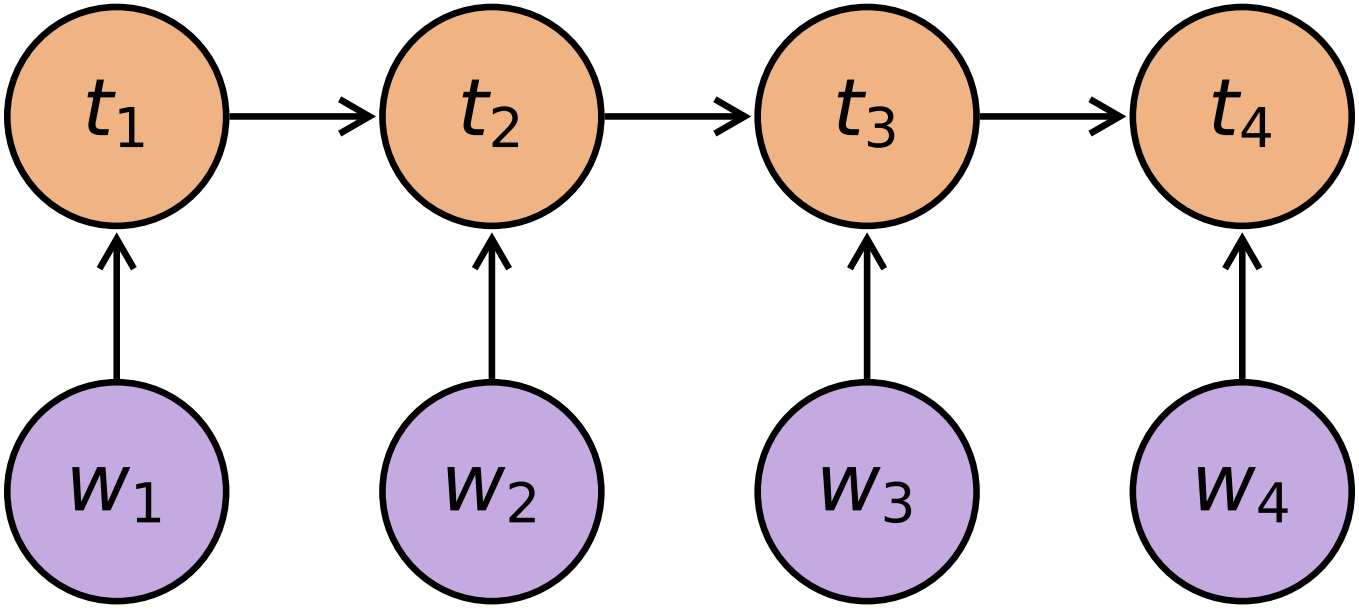
- Equivalent of smoothing
- Optimise objective function:

$$\mathcal{L} = \log P(y|x) - \lambda|\theta|$$

Recap: Hidden Markov Model



MaxEnt Markov Model



MaxEnt Markov Model

- MaxEnt: logistic regression
- Markov: limited context

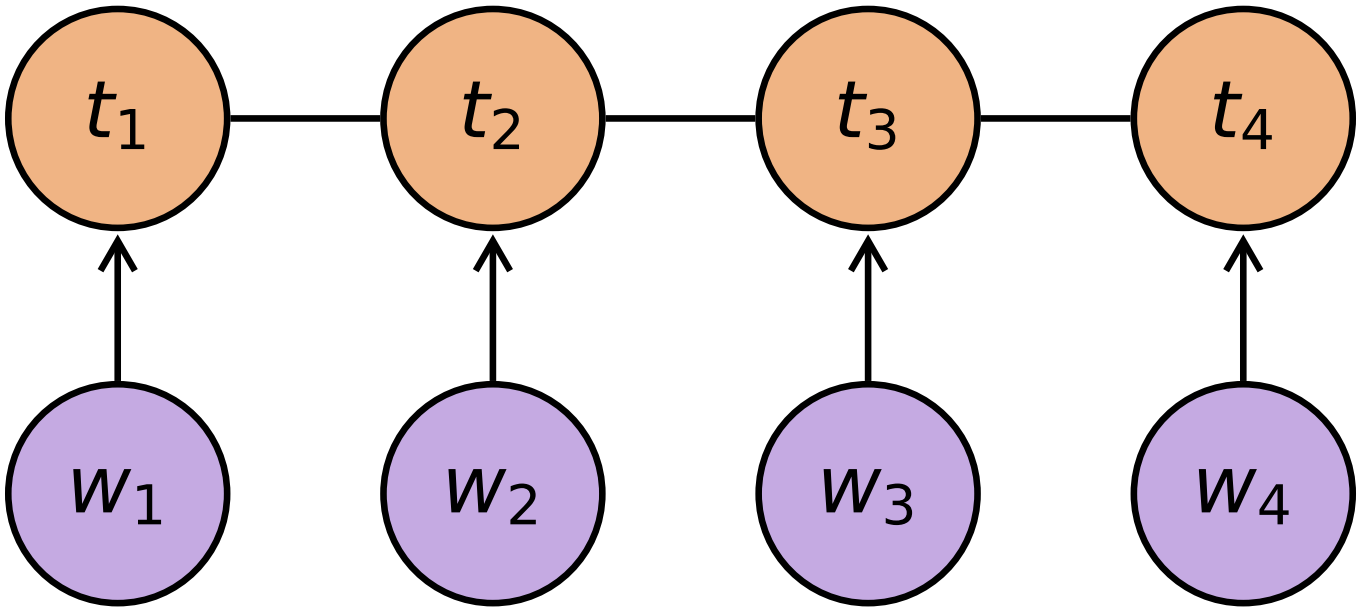
MaxEnt Markov Model

- MaxEnt: logistic regression
- Markov: limited context
- Locally normalised: token by token

MaxEnt Markov Model

- MaxEnt: logistic regression
- Markov: limited context
- Locally normalised: token by token
- Dynamic programming for inference

Conditional Random Field



Conditional Random Field

- Conditional: discriminative
- Random field: undirected

Conditional Random Field

- Conditional: discriminative
- Random field: undirected
- Globally normalised: all at once

Conditional Random Field

- Conditional: discriminative
- Random field: undirected
- Globally normalised: all at once
- Dynamic programming or beam search for inference

Named Entity Recognition

Bill Gates says mosquitoes scare him more than sharks.

Named Entity Recognition

Bill Gates says mosquitoes scare him more than sharks.

Named Entity Recognition

Bill Gates says mosquitoes scare him more than sharks.

The reaction will produce 2,4- and 2,6-dinitrotoluene.

Named Entity Recognition

Bill Gates says mosquitoes scare him more than sharks.

The reaction will produce 2,4- and 2,6-dinitrotoluene.

Named Entity Recognition

- Sequence labelling task
- Usually into classes: PER, LOC, etc.

BIO scheme

Bill Gates says mosquitoes

scare him more than sharks

B beginning

I inside

O outside

BIO scheme

Bill Gates says mosquitoes
B

scare him more than sharks

B beginning

I inside

O outside

BIO scheme

Bill	Gates	says	mosquitoes
B	I		

scare him more than sharks

B beginning

I inside

O outside

BIO scheme

Bill	Gates	says	mosquitoes
B	I	O	

scare him more than sharks

B beginning

I inside

O outside

BIO scheme

Bill	Gates	says	mosquitoes
B	I	O	O

scare him more than sharks

B beginning

I inside

O outside

BIO scheme

Bill	Gates	says	mosquitoes
B	I	O	O

scare	him	more	than	sharks
O	O	O	O	O

B beginning

I inside

O outside

BIO scheme

Bill	Gates	says	mosquitoes
B-PER	I-PER	O	O

scare	him	more	than	sharks
O	O	O	O	O

B	beginning
I	inside
O	outside

Defining the Task

The New York Stock Exchange fell today.

Defining the Task

The New York Stock Exchange fell today.

Defining the Task

The New York Stock Exchange fell today.

Defining the Task

The New York Stock Exchange fell today.

Defining the Task

The New York and Chicago Stock Exchanges fell today.

Defining the Task

Queen Elizabeth
the Queen
the Queen of England
the queen of England
a queen of England
the queen of France

Features for Named Entity Recognition

- Gazeteers (lists of names)
- Capitalisation
- Digits
- Punctuation
- Specific words preceding/following (Prof., Inc.)