Information Extraction

Lecture 5 – Named Entity Recognition III

CIS, LMU München
Winter Semester 2015-2016

Dr. Alexander Fraser, CIS

Administravia

- Seminar on Thursday (ONLY!) is cancelled this week
  - Both Wednesday and Thursday seminars will meet next week!
- See Seminar web page for updated schedule

Outline

- IE end-to-end
- Introduction: named entity detection as a classification problem

CMU Seminars task

- Given an email about a seminar
- Annotate
  - Speaker
  - Start time
  - End time
  - Location

CMU Seminars - Example

<0.24.4.93.20.59.10.jgc+@NL.CS.CMU.EDU (Jaime Carbonell).0>

Type: cmu.cs.proj.mt

Topic: <speaker>Nagao</speaker> Talk

Dates: 26-Apr-93

Time: <stime>10:00</stime> - <etime>11:00 AM</etime>

PostedBy: jgc+ on 24-Apr-93 at 20:59 from NL.CS.CMU.EDU (Jaime Carbonell)

Abstract:

<paragraph><sentence>This Monday, 4/26, <speaker>Prof. Makoto Nagao</speaker> will give a seminar in the <location>CMT red conference room</location> <stime>10</stime>-<etime>11am</etime> on recent MT research results</sentence>.</paragraph>

IE Template

Slot Name	Value
Speaker	Prof. Makoto Nagao
Start time	1993-04-26 10:00
End time	1993-04-26 11:00
Location	CMT red conference room
Message Identifier (Filename)	0.24.4.93.20.59.10.jgc+@NL.CS.CMU.EDU (Jaime Carbonell).0

- Template contains *canonical* version of information
  - There are several "mentions" of speaker, start time and end-time (see previous slide)
  - Only one value for each slot
  - Location could probably also be canonicalized
  - Important: also keep link back to original text

How many database entries?

- In the CMU seminars task, one message generally results in one database entry
  - Or no database entry if you process an email that is not about a seminar
- In other IE tasks, can get multiple database entries from a single document or web page
  - A page of concert listings -> database entries
  - Entries in timeline -> database entries

Summary

- IR: end-user
  - Start with information need
  - Gets relevant documents, hopefully information need is solved
  - Important difference: Traditional IR vs. Web R
- IE: analyst (you)
  - Start with template design and corpus
  - Get database of filled out templates
    - Followed by subsequent processing (e.g., data mining, or user browsing, etc.)

IE: what we've seen so far

So far we have looked at:

- Source issues (selection, tokenization, etc)
- Extracting regular entities
- Rule-based extraction of named entities
- Learning rules for rule-based extraction of named entities
- We also jumped ahead and looked briefly at end-to-end IE for the CMU Seminars task

Information Extraction

Information Extraction (IE) is the process of extracting structured information from unstructured machine-readable documents

and beyond

Ontological Information Extraction

Fact Extraction

Instance Extraction

Elvis Presley	singer
Angela Merkel	politician

Named Entity Recognition

...married <u>Elvis</u> on 1967-05-01

✓

Tokenization& Normalization

05/01/67 → 1967-05-01

✓

Source Selection

?

Where we are going

- We will stay with the named entity recognition (NER) topic for a while
  - How to formulate this as a machine learning problem (later in these slides)
  - Next time: brief introduction to machine learning

Named Entity Recognition

Named Entity Recognition (NER) is the process of finding entities (people, cities, organizations, dates, ...) in a text.

Elvis Presley was born in 1935 in East Tupelo, Mississippi.

Extracting Named Entities

Person: Mr. Hubert J. Smith, Adm. McInnes, Grace Chan

Title: Chairman, Vice President of Technology, Secretary of State

Country: USSR, France, Haiti, Haitian Republic

City: New York, Rome, Paris, Birmingham, Seneca Falls

Province: Kansas, Yorkshire, Uttar Pradesh

Business: GTE Corporation, FreeMarkets Inc., Acme

University: Bryn Mawr College, University of Iowa

Organization: Red Cross, Boys and Girls Club

More Named Entities

Currency: 400 yen, $100, DM 450,000

Linear: 10 feet, 100 miles, 15 centimeters

Area: a square foot, 15 acres

Volume: 6 cubic feet, 100 gallons

Weight: 10 pounds, half a ton, 100 kilos

Duration: 10 day, five minutes, 3 years, a millennium

Frequency: daily, biannually, 5 times, 3 times a day

Speed: 6 miles per hour, 15 feet per second, 5 kph

Age: 3 weeks old, 10-year-old, 50 years of age

Slide from J. Lin

Information extraction approaches

For years, Microsoft Corporation CEO Bill Gates was against open source. But today he appears to have changed his mind. "We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access.“

Richard Stallman, founder of the Free Software Foundation, countered saying…

Name	Title	Organization
Bill Gates	CEO	Microsoft
Bill Veghte	VP	Microsoft
Richard Stallman	Founder	Free Soft..

Slide from Kauchak

IE Posed as a Machine Learning Task

- Training data: documents marked up with ground truth
- Extract features around words/information
- Pose as a classification problem

$$\ldots\ \underbrace{\textbf{00 : pm Place :}}_{\text{prefix}}\ \underbrace{\textbf{Wean Hall Rm 5409}}_{\text{contents}}\ \underbrace{\textbf{Speaker : Sebastian Thrun}}_{\text{suffix}}\ \ldots$$

Slide from Kauchak

Sliding Windows

Information Extraction: Tuesday 10:00 am, Rm 407b

For each position, ask: Is the current window a named entity?

Window size = 1

Slide from Suchanek

Sliding Windows

Information Extraction: Tuesday 10:00 am, Rm 407b

For each position, ask: Is the current window a named entity?

Window size = 2

Slide from Suchanek

Features

Information Extraction: Tuesday 10:00 am, Rm 407b

Prefix window   Content window   Postfix window

Choose certain **features** (properties) of windows that could be important:

- window contains colon, comma, or digits
- window contains week day, or certain other words
- window starts with lowercase letter
- window contains only lowercase letters
- ...

Feature Vectors

Information Extraction: Tuesday 10:00 am, Rm 407b

Prefix window | Content window | Postfix window

Features	Feature Vector
Prefix colon	1
Prefix comma	0
...	...
Content colon	1
Content comma	0
...	...
Postfix colon	0
Postfix comma	1

The **feature vector** represents the presence or absence of features of one content window (and its prefix window and postfix window)

Slide from Suchanek

Sliding Windows Corpus

Now, we need a **corpus** (set of documents) in which the entities of interest have been manually labeled.

time   location

NLP class: Wednesday, <u>7:30am</u> and Thursday all day, <u>rm 667</u>

From this corpus, compute the feature vectors with labels:

$$\begin{bmatrix}1\\0\\0\\0\\1\end{bmatrix} \cdots \begin{bmatrix}1\\1\\0\\0\\0\end{bmatrix} \cdots \begin{bmatrix}1\\0\\1\\1\\1\end{bmatrix} \cdots \begin{bmatrix}1\\0\\0\\0\\1\end{bmatrix} \cdots \begin{bmatrix}1\\0\\1\\0\\1\end{bmatrix}$$

Nothing Nothing Time Nothing Location

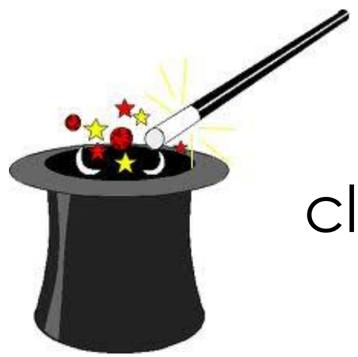
Slide from Suchanek

Machine Learning

Information Extraction: Tuesday $\underbrace{\text{10:00 am,}}$ Rm 407b

Use the labeled feature vectors as training data for Machine Learning

$\begin{bmatrix}1\\0\\0\\1\\1\\1\\1\end{bmatrix}$ $\begin{bmatrix}1\\1\\0\\0\\1\\0\end{bmatrix}$ $\Rightarrow$



Machine Learning

classify

$\begin{bmatrix}1\\1\\0\\0\\0\\0\end{bmatrix} \xrightarrow{\text{Result}}$ Time

Nothing Time

Slide from Suchanek

Sliding Windows Exercise

What features would you use to recognize person names?

$\underbrace{\text{Elvis Presley}}$ married $\underbrace{\text{Ms. Priscilla}}$ at the $\underbrace{\text{Aladin Hotel}}$.

$$
\begin{array}{l}
\text{UpperCase} \\
\text{hasDigit} \\
\ldots \\
\\
\\
\\
\end{array}
\begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \end{bmatrix}
\qquad
\begin{bmatrix} 1 \\ 0 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}
\qquad \ldots \qquad
\begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \\ 1 \\ 0 \end{bmatrix}
$$

Slide from Suchanek

Good Features for Information Extraction

begins-with-number
begins-with-ordinal
begins-with-punctuation
begins-with-question-word
begins-with-subject
blank
contains-alphanum
contains-bracketed-number
contains-http
contains-non-space
contains-number
contains-pipe

Example word features:
- identity of word
- is in all caps
- ends in "-ski"
- is part of a noun phrase
- is in a list of city names
- is under node X in WordNet or Cyc
- is in bold font
- is in hyperlink anchor
- *features of past & future*
- last person name was female
- next two words are "and Associates"

contains-question-mark
contains-question-word
ends-with-question-mark
first-alpha-is-capitalized
indented
indented-1-to-4
indented-5-to-10
more-than-one-third-space
only-punctuation
prev-is-blank
prev-begins-with-ordinal
shorter-than-30

Slide from Kauchak

Good Features for Information Extraction

- Is Capitalized
- Is Mixed Caps
- Is All Caps
- Initial Cap
- Contains Digit
- All lowercase
- Is Initial
- Punctuation
- Period
- Comma
- Apostrophe
- Dash
- Preceded by HTML tag

- Character n-gram classifier says string is a person name (80% accurate)
- In stopword list (the, of, their, etc)
- In honorific list (Mr, Mrs, Dr, Sen, etc)
- In person suffix list (Jr, Sr, PhD, etc)
- In name particle list (de, la, van, der, etc)
- In Census lastname list; segmented by P(name)
- In Census firstname list; segmented by P(name)
- In locations lists (states, cities, countries)
- In company name list ("J. C. Penny")
- In list of company suffixes (Inc, & Associates, Foundation)

Word Features
- lists of job titles,
- Lists of prefixes
- Lists of suffixes
- 350 informative phrases

HTML/Formatting Features
- {begin, end, in} x {, <i>, <a>, <hN>} x {lengths 1, 2, 3, 4, or longer}
- {begin, end} of line

Slide from Kauchak

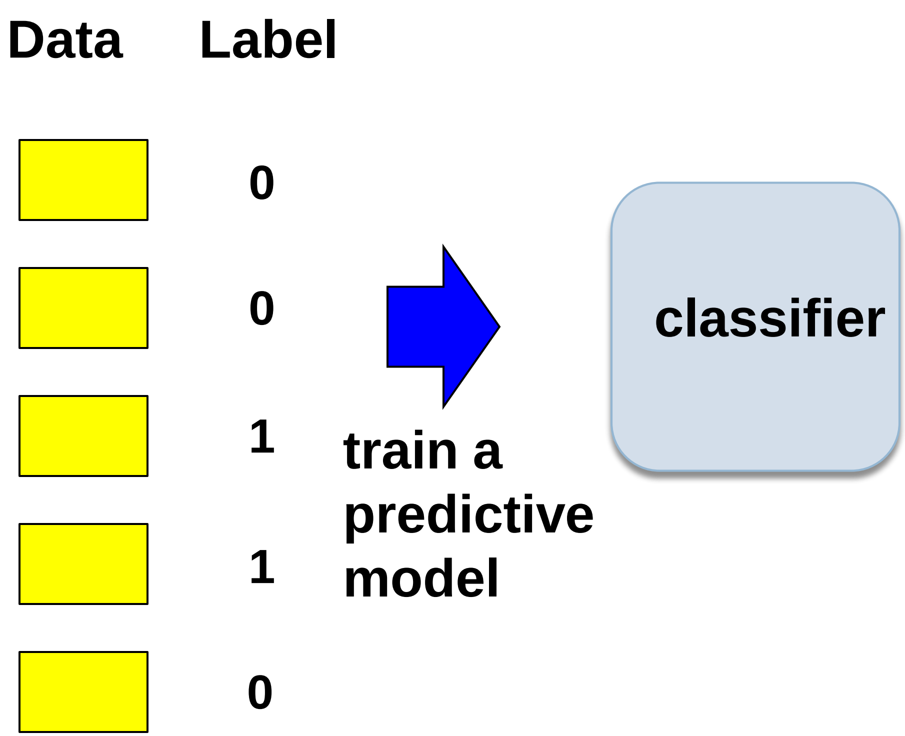
NER Classification in more detail

- In the previous slides, we covered a basic idea of how NER classification works
- In the next slides, I will go into more detail
  - I will compare sliding window with boundary detection
- Machine learning itself will be presented in more detail in the next lecture

How can we pose this as a classification (or learning) problem?

… 00 : pm Place : Wean Hall Rm 5409 Speaker : Sebastian Thrun …

prefix (00 : pm Place :) — contents (Wean Hall Rm 5409) — suffix (Speaker : Sebastian Thrun)

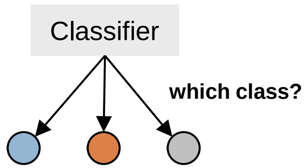


Slide from Kauchak

Lots of possible techniques

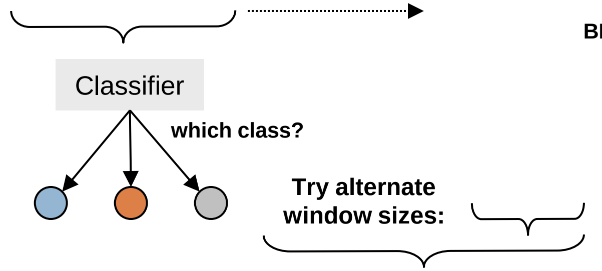
Classify Candidates

Abraham Lincoln was born in Kentucky.

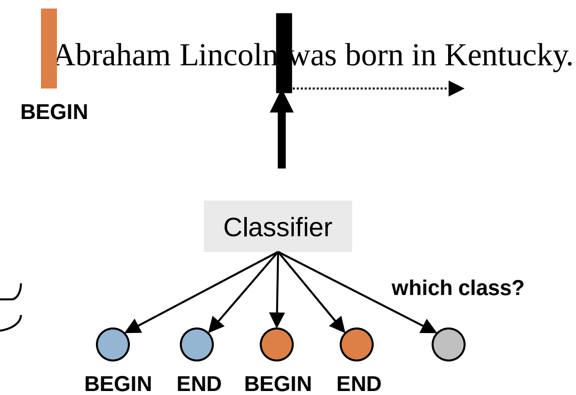


Sliding Window

Abraham Lincoln was born in Kentucky.

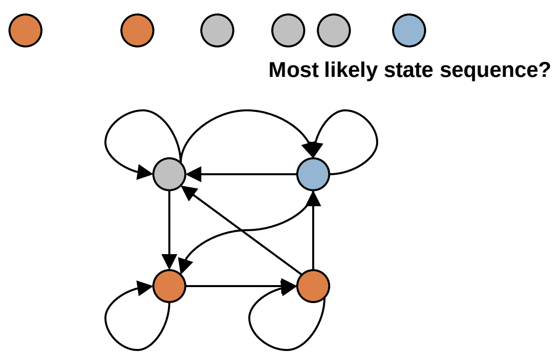


Boundary Models



Finite State Machines

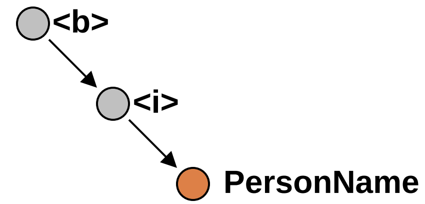
Abraham Lincoln was born in Kentucky.



Wrapper Induction

<i>***Abraham Lincoln***</i> was born in Kentucky.

Learn and apply pattern for a website



Any of these models can be used to capture words, formatting or both.

Slide from Kauchak

Information Extraction by Sliding Window

E.g.
Looking for seminar location

```
        GRAND CHALLENGES FOR MACHINE LEARNING

               Jaime Carbonell
           School of Computer Science
          Carnegie Mellon University

                  3:30 pm
               7500 Wean Hall

Machine learning has evolved from obscurity
in the 1970s into a vibrant and popular
discipline in artificial intelligence
during the 1980s and 1990s.   As a result
of its success and growth, machine learning
is evolving into a collection of related
disciplines: inductive concept acquisition,
analytic learning in problem solving (e.g.
analogy, explanation-based learning),
learning theory (e.g. PAC learning),
genetic algorithms, connectionist learning,
hybrid systems, and so on.
```

CMU UseNet Seminar Announcement

Slide from Kauchak

Information Extraction by Sliding Window

E.g.
Looking for seminar location

```
        GRAND CHALLENGES FOR MACHINE LEARNING

               Jaime Carbonell
           School of Computer Science
          Carnegie Mellon University

                   3:30 pm
               7500 Wean Hall

Machine learning has evolved from obscurity
in the 1970s into a vibrant and popular
discipline in artificial intelligence
during the 1980s and 1990s.   As a result
of its success and growth, machine learning
is evolving into a collection of related
disciplines: inductive concept acquisition,
analytic learning in problem solving (e.g.
analogy, explanation-based learning),
learning theory (e.g. PAC learning),
genetic algorithms, connectionist learning,
hybrid systems, and so on.
```

CMU UseNet Seminar Announcement

Slide from Kauchak

Information Extraction by Sliding Window

E.g.
Looking for seminar location

```
        GRAND CHALLENGES FOR MACHINE LEARNING

                 Jaime Carbonell
           School of Computer Science
          Carnegie Mellon University

                     3:30 pm
                  7500 Wean Hall

Machine learning has evolved from obscurity
in the 1970s into a vibrant and popular
discipline in artificial intelligence
during the 1980s and 1990s.   As a result
of its success and growth, machine learning
is evolving into a collection of related
disciplines: inductive concept acquisition,
analytic learning in problem solving (e.g.
analogy, explanation-based learning),
learning theory (e.g. PAC learning),
genetic algorithms, connectionist learning,
hybrid systems, and so on.
```

CMU UseNet Seminar Announcement

Information Extraction by Sliding Window

E.g.
Looking for seminar location

```
     GRAND CHALLENGES FOR MACHINE LEARNING

            Jaime Carbonell
        School of Computer Science
       Carnegie Mellon University

                3:30 pm
             7500 Wean Hall

Machine learning has evolved from obscurity
in the 1970s into a vibrant and popular
discipline in artificial intelligence
during the 1980s and 1990s.   As a result
of its success and growth, machine learning
is evolving into a collection of related
disciplines: inductive concept acquisition,
analytic learning in problem solving (e.g.
analogy, explanation-based learning),
learning theory (e.g. PAC learning),
genetic algorithms, connectionist learning,
hybrid systems, and so on.
```

CMU UseNet Seminar Announcement

Information Extraction by Sliding Window

E.g. Looking for seminar location

```
        GRAND CHALLENGES FOR MACHINE LEARNING

                 Jaime Carbonell
             School of Computer Science
            Carnegie Mellon University

                     3:30 pm
                  7500 Wean Hall

Machine learning has evolved from obscurity
in the 1970s into a vibrant and popular
discipline in artificial intelligence
during the 1980s and 1990s.   As a result
of its success and growth, machine learning
is evolving into a collection of related
disciplines: inductive concept acquisition,
analytic learning in problem solving (e.g.
analogy, explanation-based learning),
learning theory (e.g. PAC learning),
genetic algorithms, connectionist learning,
hybrid systems, and so on.
```

CMU UseNet Seminar Announcement

Slide from Kauchak

Information Extraction by Sliding Window

… 00 : pm Place : Wean Hall Rm 5409 Speaker : Sebastian Thrun …

$\underbrace{w_{t-m} \quad \ldots \quad w_{t-1}}_{\text{prefix}} \quad \underbrace{w_{t} \quad \ldots \quad w_{t+n}}_{\text{contents}} \quad \underbrace{w_{t+n+1} \quad \ldots \quad w_{t+n+m}}_{\text{suffix}}$

- Standard supervised learning setting
  - Positive instances?
  - Negative instances?

Information Extraction by Sliding Window

… 00 : pm Place : Wean Hall Rm 5409 Speaker : Sebastian Thrun …

$$\underbrace{w_{t-m} \quad \cdots \quad w_{t-1}}_{\text{prefix}} \quad \underbrace{w_{t} \quad \cdots \quad w_{t+n}}_{\text{contents}} \quad \underbrace{w_{t+n+1} \quad \cdots \quad w_{t+n+m}}_{\text{suffix}}$$

- Standard supervised learning setting
  - Positive instances: Windows with real label
  - Negative instances: All other windows
  - Features based on candidate, prefix and suffix

IE by Boundary Detection

E.g. Looking for seminar location

```
      GRAND CHALLENGES FOR MACHINE LEARNING

              Jaime Carbonell
          School of Computer Science
         Carnegie Mellon University

                  3:30 pm
               7500 Wean Hall

Machine learning has evolved from obscurity
in the 1970s into a vibrant and popular
discipline in artificial intelligence
during the 1980s and 1990s.   As a result
of its success and growth, machine learning
is evolving into a collection of related
disciplines: inductive concept acquisition,
analytic learning in problem solving (e.g.
analogy, explanation-based learning),
learning theory (e.g. PAC learning),
genetic algorithms, connectionist learning,
hybrid systems, and so on.
```

CMU UseNet Seminar Announcement

IE by Boundary Detection

**E.g.
Looking for
seminar
location**

```
     GRAND CHALLENGES FOR MACHINE LEARNING

            Jaime Carbonell
        School of Computer Science
       Carnegie Mellon University

                3:30 pm
             7500 Wean Hall

Machine learning has evolved from obscurity
in the 1970s into a vibrant and popular
discipline in artificial intelligence
during the 1980s and 1990s.   As a result
of its success and growth, machine learning
is evolving into a collection of related
disciplines: inductive concept acquisition,
analytic learning in problem solving (e.g.
analogy, explanation-based learning),
learning theory (e.g. PAC learning),
genetic algorithms, connectionist learning,
hybrid systems, and so on.
```

CMU UseNet Seminar Announcement

IE by Boundary Detection

E.g. Looking for seminar location

```
    GRAND CHALLENGES FOR MACHINE LEARNING

           Jaime Carbonell
       School of Computer Science
      Carnegie Mellon University

              3:30 pm
           7500 Wean Hall

Machine learning has evolved from obscurity
in the 1970s into a vibrant and popular
discipline in artificial intelligence
during the 1980s and 1990s.   As a result
of its success and growth, machine learning
is evolving into a collection of related
disciplines: inductive concept acquisition,
analytic learning in problem solving (e.g.
analogy, explanation-based learning),
learning theory (e.g. PAC learning),
genetic algorithms, connectionist learning,
hybrid systems, and so on.
```

CMU UseNet Seminar Announcement

Slide from Kauchak

IE by Boundary Detection

**E.g.
Looking for
seminar
location**

```
          GRAND CHALLENGES FOR MACHINE LEARNING

                 Jaime Carbonell
             School of Computer Science
            Carnegie Mellon University

                      3:30 pm
                  7500 Wean Hall

Machine learning has evolved from obscurity
in the 1970s into a vibrant and popular
discipline in artificial intelligence
during the 1980s and 1990s.   As a result
of its success and growth, machine learning
is evolving into a collection of related
disciplines: inductive concept acquisition,
analytic learning in problem solving (e.g.
analogy, explanation-based learning),
learning theory (e.g. PAC learning),
genetic algorithms, connectionist learning,
hybrid systems, and so on.
```

CMU UseNet Seminar Announcement

IE by Boundary Detection

E.g. Looking for seminar location

```
    GRAND CHALLENGES FOR MACHINE LEARNING

           Jaime Carbonell
       School of Computer Science
      Carnegie Mellon University

               3:30 pm
            7500 Wean Hall

Machine learning has evolved from obscurity
in the 1970s into a vibrant and popular
discipline in artificial intelligence
during the 1980s and 1990s.   As a result
of its success and growth, machine learning
is evolving into a collection of related
disciplines: inductive concept acquisition,
analytic learning in problem solving (e.g.
analogy, explanation-based learning),
learning theory (e.g. PAC learning),
genetic algorithms, connectionist learning,
hybrid systems, and so on.
```

CMU UseNet Seminar Announcement

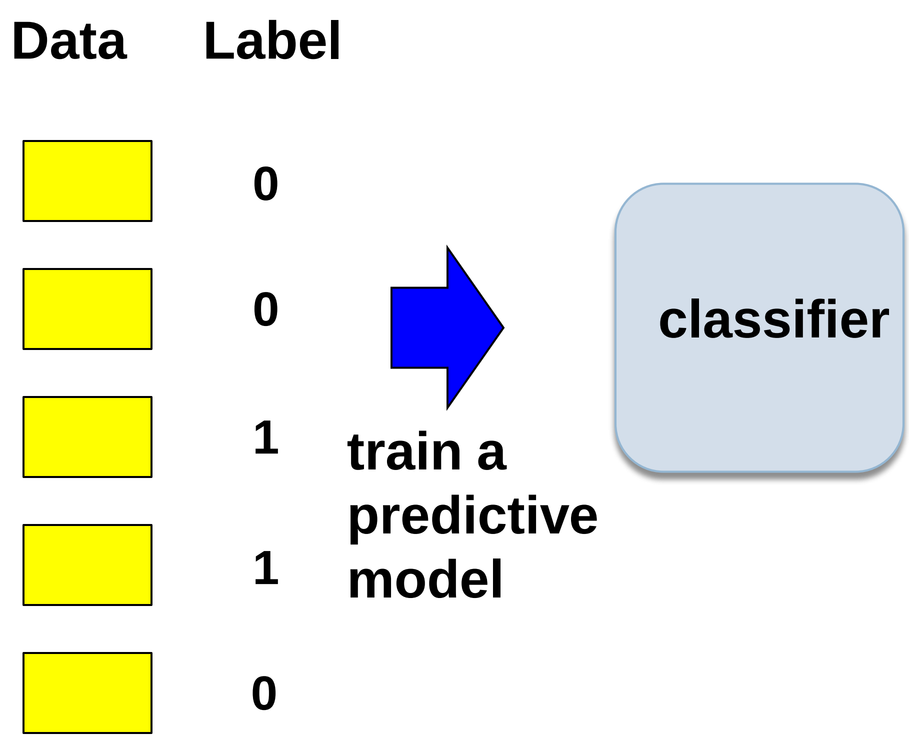
Slide from Kauchak

IE by Boundary Detection

Input: Linear Sequence of Tokens

Date : Thursday , October 25 Time : 4 : 15 - 5 : 30 PM

How can we pose this as a machine learning problem?



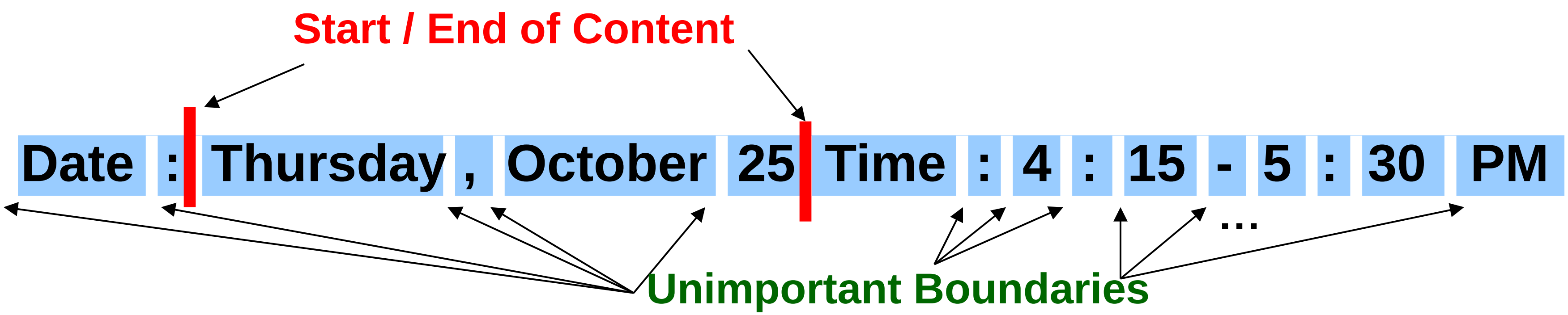
Slide from Kauchak

IE by Boundary Detection

Input: Linear Sequence of Tokens

Date : Thursday , October 25 Time : 4 : 15 - 5 : 30 PM

Method: Identify start and end Token Boundaries



Output: Tokens Between Identified Start / End Boundaries

Date : Thursday , October 25 Time : 4 : 15 - 5 : 30 PM

Slide from Kauchak

Learning: IE as Classification

Learn *TWO* binary classifiers, one for the beginning and one for the end

Begin

Date	:	Thursday	,	October	25	Time	:	4	:	15	-	5	:	30	PM

POSITIVE (1)

End

Date	:	Thursday	,	October	25	Time	:	4	:	15	-	5	:	30	PM

ALL OTHERS NEGATIVE (0)

$$\texttt{Begin}(i) = \begin{cases} 1 & \text{if } i \text{ begins a field} \\ 0 & \text{otherwise} \end{cases}$$

Slide from Kauchak

One approach: Boundary Detectors

A "*Boundary Detectors*" is a pair of token sequences ‹p,s›

- A detector matches a boundary if p matches text before boundary and s matches text after boundary
- Detectors can contain wildcards, e.g. "capitalized word", "number", etc.

<Date: , [*CapitalizedWord*]>

Date: Thursday, October 25

Would this boundary detector match anywhere?

Slide from Kauchak

One approach: Boundary Detectors

A "*Boundary Detectors*" is a pair of token sequences ‹p,s›

- A detector matches a boundary if p matches text before boundary and s matches text after boundary
- Detectors can contain wildcards, e.g. "capitalized word", "number", etc.

<**Date:** , **[*CapitalizedWord*]**>

Date: Thursday, October 25

Slide from Kauchak

Combining Detectors

	Prefix	Suffix
Begin **boundary detector:**	`<a href="`	`http`
End **boundary detector:**	*empty*	`">`

`text<b><a href="http://www.cs.pomona.edu">`

match(es)?

Slide from Kauchak

Combining Detectors

	Prefix	Suffix
Begin **boundary detector:**	<a href="	http
End **boundary detector:**	*empty*	">

text<b><a href="http://www.cs.pomona.edu">

Begin **End**

Slide from Kauchak

Learning: IE as Classification

Learn *TWO* binary classifiers, one for the beginning and one for the end

Begin

Date : Thursday , October 25 Time : 4 : 15 - 5 : 30 PM

POSITIVE (1)

End

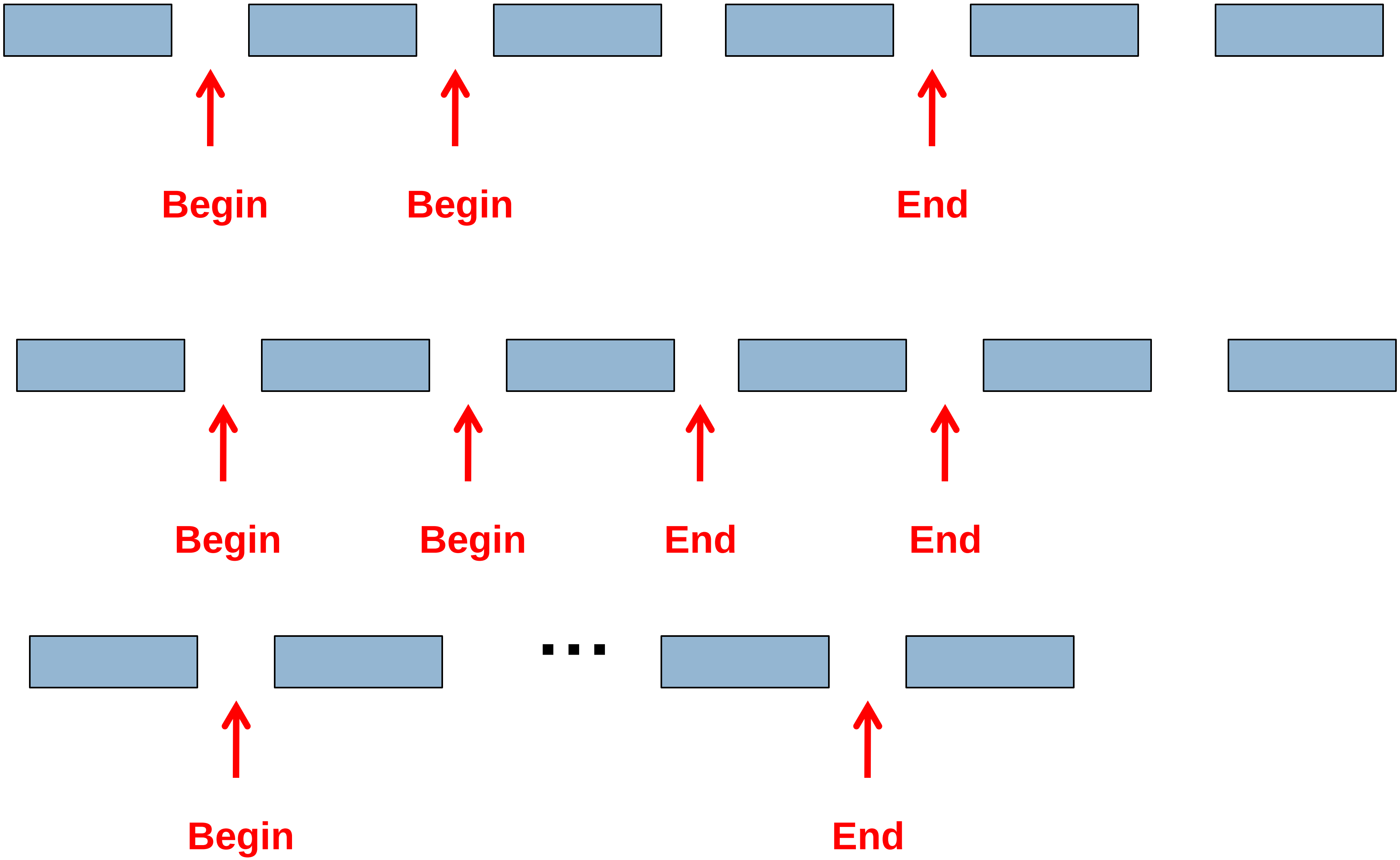
Date : Thursday , October 25 Time : 4 : 15 - 5 : 30 PM

ALL OTHERS NEGATIVE (0)

Say we learn Begin and End, will this be enough?
Any improvements? Any ambiguities?

Slide from Kauchak

Some concerns



Slide from Kauchak

Learning to detect boundaries

- Learn **three** probabilistic classifiers:
  - *Begin(i)* = probability position *i* starts a field
  - *End(j)* = probability position *j* ends a field
  - *Len(k)* = probability an extracted field has length *k*

- Score a possible extraction *(i,j)* by

  *Begin(i) * End(j) * Len(j-i)*

- *Len(k)* is estimated from a histogram data

- *Begin(i) and End(j)* may combine multiple boundary detectors!

Problems with Sliding Windows and Boundary Finders

- Decisions in neighboring parts of the input are made independently from each other.

  - Sliding Window may predict a “seminar end time” before the “seminar start time”.

  - It is possible for two overlapping windows to both be above threshold.

  - In a Boundary-Finding system, left boundaries are laid down independently from right boundaries

Slide from Kauchak

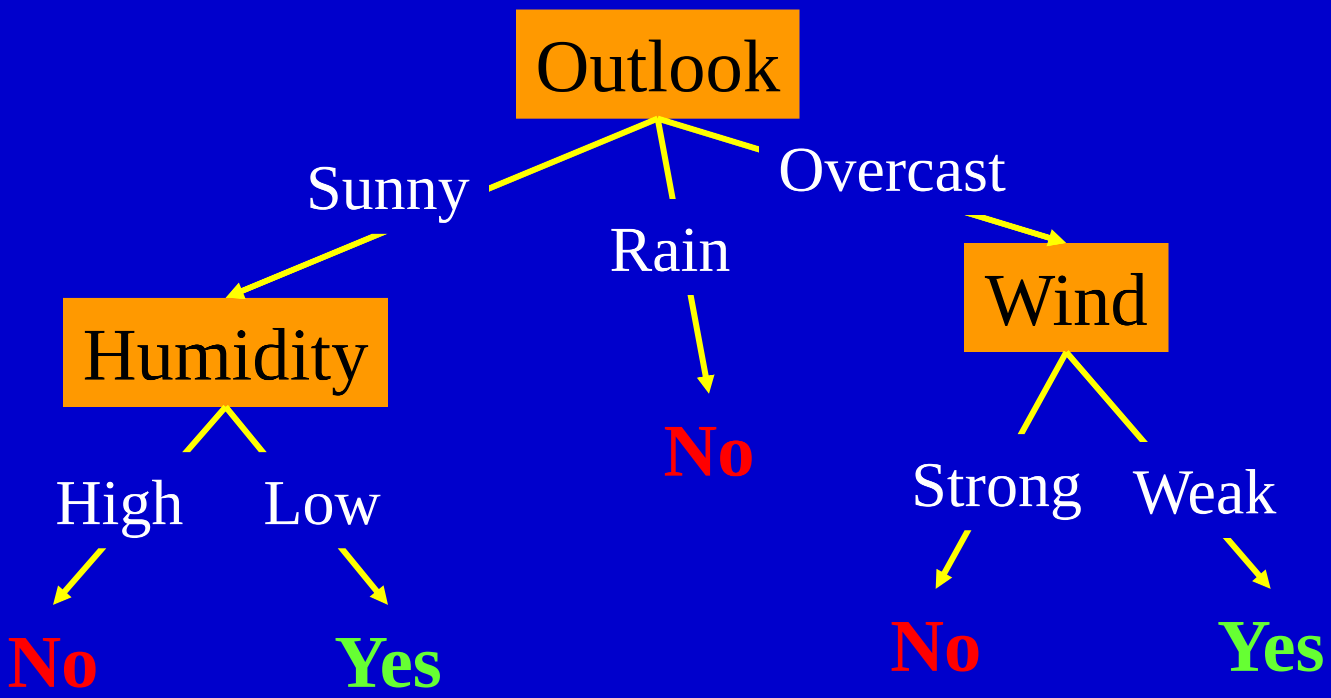
- Slide sources
  - A number of slides were taken from a wide variety of sources (see the attribution at the bottom right of each slide)
  - I'd particularly like to mention Dave Kauchak of Pomona College

Next time: machine learning

- We will take a break from NER and look at classification in general
- We will first focus on learning **decision trees** from training data
  - Powerful mechanism for encoding general decisions
  - Example on next slide

Decision Trees

"Should I play tennis today?"



A decision tree can be expressed as a disjunction of conjunctions

$$(\text{Outlook} = \text{sunny}) \wedge (\text{Humidity} = \text{normal})$$

$$\vee\ (\text{Outlook} = \text{overcast}) \wedge (\text{Wind}=\text{Weak})$$

Slide from Mitchell/Ponce

- Thank you for your attention!