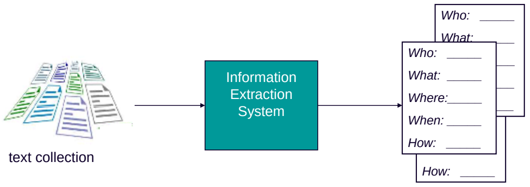
Information Extraction

(Slides are modified from Claire Cardie & Ray Mooney)

Information extraction



text collection

Information Extraction System

Who: _____
What: _____
Where: _____
When: _____
How: _____

Information extraction (IE)

- Identify specific pieces of information (data) in a unstructured or semi-structured textual document.
- Transform unstructured information in a corpus of documents or web pages into a structured database.
- Applied to different types of text:
  - Newspaper articles
  - Web pages
  - Scientific articles
  - Newsgroup messages
  - Classified ads
  - Medical notes

IE system: natural disasters

Disaster Type: earthquake

- location: *Afghanistan*
- date: *today*
- magnitude: *6.9*
- magnitude-confidence: high
- epicenter: *a remote part of the country*
- damage:
  - human-effect:
    - victim: *Thousands of people*
    - number: *Thousands*
    - outcome: dead
    - confidence: medium
    - confidence-marker: *feared*
  - physical-effect:
    - object: *entire villages*
    - outcome: damaged
    - confidence: medium
    - confidence-marker: *Details now hard to come by / reports say*

PAKISTAN MAY BE PREPARING FOR ANOTHER TEST

Thousands of people are feared dead following... (voice-over) ...a powerful earthquake that hit Afghanistan today. The quake registered 6.9 on the Richter scale, centered in a remote part of the country. (on camera) Details now hard to come by, but reports say entire villages were buried by the quake.

Document no.: ABC19980530.1830.0342
Date/time: 05/30/1998 18:35:42.49

Sample Job Posting

Subject: US-TN-SOFTWARE PROGRAMMER
Date: 17 Nov 1996 17:37:29 GMT
Organization: Reference.Com Posting Service
Message-ID: <56nigp$mrs@bilbo.reference.com>

SOFTWARE PROGRAMMER

Position available for Software Programmer experienced in generating software for PC-Based Voice Mail systems. Experienced in C Programming. Must be familiar with communicating with and controlling voice cards; preferable Dialogic, however, experience with others such as Rhetorix and Natural Microsystems is okay. Prefer 5 years or more experience with PC Based Voice Mail, but will consider as little as 2 years. Need to find a Senior level person who can come on board and pick up code with very little training. Present Operating System is DOS. May go to OS-2 or UNIX in future.

Please reply to:
Kim Anderson
AdNET
(901) 458-2888 fax
kimander@memphisonline.com

Extracted Job Template

computer_science_job
id: 56nigp$mrs@bilbo.reference.com
title: SOFTWARE PROGRAMMER
salary:
company:
recruiter:
state: TN
city:
country: US
language: C
platform: PC \ DOS \ OS-2 \ UNIX
application:
area: Voice Mail
req_years_experience: 2
desired_years_experience: 5
req_degree:
desired_degree:
post_date: 17 Nov 1996

Information Extraction Tasks:

1. Named Entity Recognition
   - Extract phrases that correspond to people, places, organizations, etc
   - Sequence Tagging + Classification
2. Relation Extraction
   - Extract relations among entities
   - Example:
     - Employed-by
     - Located-at
     - Part-of
     - Michael Dell is the CEO of Dell Computer Corporation and lives in Austin Texas.

Named Entity Recognition 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.

Named Entity Recognition 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.

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: Named Entity Recognition (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

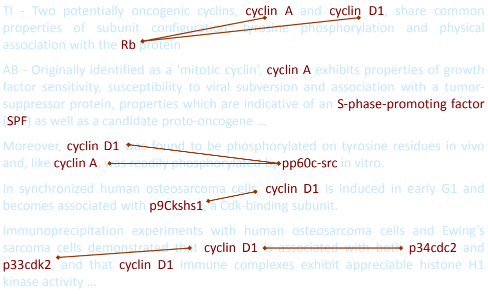
AB - Originally identified as a 'mitotic cyclin', cyclin A exhibits properties of growth factor sensitivity, susceptibility to viral subversion and association with a tumor-suppressor protein, properties which are indicative of an S-phase-promoting factor (SPF) as well as a candidate proto-oncogene …

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 …

Medline Corpus: Relation Extraction Protein Interactions



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 …

Web Extraction

- Many web pages are generated automatically from an underlying database.
- Therefore, the HTML structure of pages is fairly specific and regular (*semi-structured*).
- An IE system for such generated pages allows the web site to be viewed as a structured database.

Amazon Book Description

```
….
</td></tr>
</table>
<b class="sans">The Age of Spiritual Machines : When Computers Exceed Human Intelligence</b><br>
<font face=verdana,arial,helvetica size=-1>
by <a href="/exec/obidos/search-handle-url/index=books&field-author=
            Kurzweil%2C%20Ray/002-6235079-4593641">
Ray Kurzweil</a><br>
</font>
<br>
<a href="http://images.amazon.com/images/P/0140282025.01.LZZZZZZZ.jpg">
<img src="http://images.amazon.com/images/P/0140282025.01.MZZZZZZZ.gif" width=90
   height=140 align=left border=0></a>
<font face=verdana,arial,helvetica size=-1>
<span class="small">
<span class="small">
<b>List Price:</b> <span class=listprice>$14.95</span><br>
<b>Our Price: <font color=#990000>$11.96</font></b><br>
<b>You Save:</b> <font color=#990000><b>$2.99 </b>
(20%)</font><br>
</span>
<p> <br>…
```

Extracted Book Template

Title: The Age of Spiritual Machines :
When Computers Exceed Human Intelligence
Author: Ray Kurzweil
List-Price: $14.95
Price: $11.96
:
:

Template Types

- Slots in template typically filled by a substring from the document.
- Some slots may have a fixed set of pre-specified possible fillers that may not occur in the text itself.
  - Terrorist act: threatened, attempted, accomplished.
  - Job type: clerical, service, custodial, etc.
  - Company type: SEC code
- Some slots may allow multiple fillers.

Pattern-Matching Rule Extraction

- Works well for information extraction from semi-structured text
- Using regular expressions

Regular Expression Examples

- (u|e)nabl(e|ing) matches
  - unable
  - unabling
  - enable
  - enabling
- (un|en)*able matches
  - able
  - unable
  - unenable
  - enununenable

Enhanced Regex's (Perl)

- Special terms for common sets of characters, such as alphabetic or numeric or general "wildcard".
- Special repetition operator (+) for 1 or more occurrences.
- Special optional operator (?) for 0 or 1 occurrences.
- Special repetition operator for specific range of number of occurrences: {min,max}.
  - A{1,5} One to five A's.
  - A{5,} Five or more A's
  - A{5} Exactly five A's

Perl Regex's

- Character classes:
  - \w (word char) Any alpha-numeric (not: \W)
  - \d (digit char) Any digit (not: \D)
  - \s (space char) Any whitespace (not: \S)
  - . (wildcard) Anything
- Anchor points:
  - \b (boundary) Word boundary
  - ^ Beginning of string
  - $ End of string

Perl Regex Examples

- U.S. phone number with optional area code:
  - /\b(\(\d{3}\)\s?)?\d{3}-\d{4}\b/
- Email address:
  - /\b\S+@\S+(\.com|\.edu|\.gov|\.org|\.net)\b/

Simple Extraction Patterns

- Specify an item to extract for a slot using a regular expression pattern.
  - Price pattern: "\b\$\d+(\.\d{2})?\b"
- May require preceding (pre-filler) pattern to identify proper context.
  - Amazon list price:
    - Pre-filler pattern: "<b>List Price:</b> <span class=listprice>"
    - Filler pattern: "\$\d+(\.\d{2})?\b"
- May require succeeding (post-filler) pattern to identify the end of the filler.
  - Amazon list price:
    - Pre-filler pattern: "<b>List Price:</b> <span class=listprice>"
    - Filler pattern: ".+"
    - Post-filler pattern: "</span>"

Adding NLP Information to Patterns

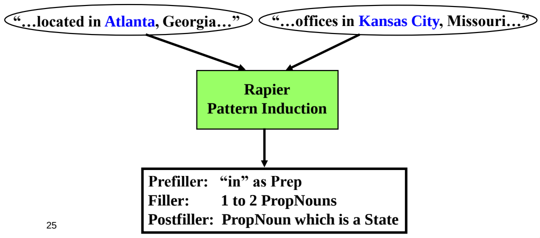
- If extracting from automatically generated web pages, simple regex patterns usually work.
- If extracting from more natural, unstructured, human-written text, some NLP may help.
  - Part-of-speech (POS) tagging
    - Mark each word as a noun, verb, preposition, etc.
  - Syntactic parsing
    - Identify phrases: NP, VP, PP
  - Semantic word categories (e.g. from WordNet)
    - KILL: kill, murder, assassinate, strangle, suffocate
- Extraction patterns can use POS or phrase tags.
  - Crime victim:
    - Prefiller: [POS: V, Hypernym: KILL]
    - Filler: [Phrase: NP]

Pattern-Match Rule Learning

- Writing accurate patterns for each slot for each application requires laborious software engineering.
- Alternative is to use rule induction methods.
- RAPIER system (Califf & Mooney, 1999) learns three regex-style patterns for each slot:
  - Pre-filler pattern
  - Filler pattern
  - Post-filler pattern
- RAPIER allows use of POS and WordNet categories in patterns to generalize over lexical items.

RAPIER Pattern Induction Example

- If goal is to extract the name of the city in which a posted job is located, the least-general-generalization constructed by RAPIER is:



"…located in Atlanta, Georgia…"

"…offices in Kansas City, Missouri…"

Rapier
Pattern Induction

Prefiller: "in" as Prep
Filler: 1 to 2 PropNouns
Postfiller: PropNoun which is a State

25

ELCS
(Extraction using Longest Common Subsequences)

- A method for inducing **pattern-match rules** that extract interactions between previously tagged proteins.
- Each rule consists of a **sequence of words** with allowable **word gaps** between them (similar to Blaschke & Valencia, 2001, 2002).
  - (7) *interactions* (0) *between* (5) **PROT** (9) **PROT** (17) .
- Any pair of proteins in a sentence if tagged as interacting forms a **positive example**, otherwise it forms a **negative example**.
- Positive examples are repeatedly **generalized** to form rules until the rules become overly general and start matching negative examples.

26

Generalizing Rules using Longest Common Subsequence

The self - association site appears to be formed by interactions between helices 1 and 2 of **beta spectrin** *repeat 17 of one dimer with helix 3 of* **alpha spectrin** *repeat 1 of the other dimer to form two combined alpha - beta triple - helical segments .*

Title - Physical and functional interactions between the transcriptional inhibitors **Id3** *and* **ITF-2b** *.*

- (7) *interactions* (0) *between* (5) **PROT** (9) **PROT** (17) .

27

IE as Sequence Labeling

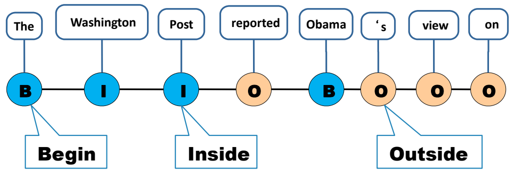
- Can treat IE as a sequence labeling problem.
- Can apply a sliding window classifier using various classification algorithms.
- Can apply probabilistic sequence models:
  - HMM
  - CRF

28

Sequence Tagging

- Prediction using **BIO** tagging

<The Washington Post> reported **<Obama>**'s view on the oil crisis.



The Washington Post reported Obama 's view on

B I I O B O O O

Begin **Inside** **Outside**

Dataset-1: MUC

- DARPA funded significant efforts in IE in the early to mid 1990's.
- Message Understanding Conference (MUC) was an annual event/competition where results were presented.
- Focused on extracting information from news articles:
  - Terrorist events
  - Industrial joint ventures
  - Company management changes
- Information extraction of particular interest to the intelligence community (CIA, NSA).

30

Dataset-2: ACE

- Newspaper article extraction task.
- Documents:
  - 422 training documents
  - 97 test documents
- Extracted information:
  - Entities: Person, Organization, Facility, Location, Geopolitical Entity
  - Relations: Role, Part, Located, Near, Social

31

Evaluating IE systems

- Evaluate performance on independent, manually-annotated test data not used during system development.
- Compute average value of metrics adapted from IR:
  - **Recall** = *# correct extractions / # extractions in gold standard*
  - **Precision** = *# correct extractions / # extractions by system*
  - **F-Measure** = (**F-Score**) Harmonic mean of recall and precision

State of the art

Unrestricted text:
65-70% R; 70-80% P

Semi-structured text:
90+% R/P

MUC [1991-94]

ACE [1991-94]

- terrorist activities
- business joint ventures
- microelectronic chip fabrication
- changes in corporate management
- natural disasters
- summarize medical patient records
- create job-listing databases from newsgroups
- bioinformatics

Issues…

- tension between domain-independent and domain-dependent language processing
  - treating task in a domain-independent way allows the use of general IR/NLP techniques and tools
  - treating task in a domain-dependent way allows for tailoring of techniques for better performance
- IE is generally handled as domain-specific text understanding
  - key system components need to be re-built for each new domain
  - difficult and time-consuming to build if constructed manually
    - Initially, ~6-12 months/system for IE from unstructured text
  - requires the expertise of computational linguists

IE vs. IR vs. full NLU

- IE requires more text-understanding capabilities than the bag-of-words approaches provided by IR techniques
- IE systems often presume that a text categorization system has identified documents relevant to the extraction domain
- IE requires more than document classification
- IE requires a more shallow understanding of the text than a natural language understanding system attempting full/deep semantic analysis.

IR, TC < IE < NLP, NLU