Information Extraction

Original version by Raymond J. Mooney

University of Texas at Austin

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

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

Named Entity Recognition

- Specific type of information extraction in which the goal is to extract formal 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.

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**

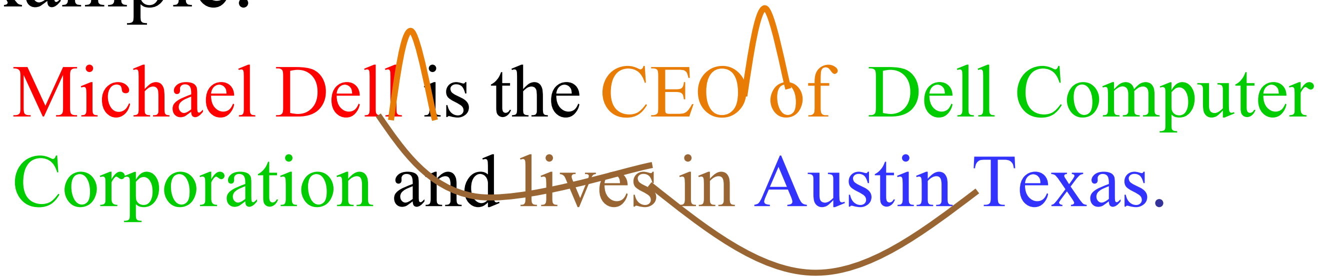
U.S. Supreme Court quashes 'illegal' Guantanamo trials

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.

Relation Extraction

- Once entities are recognized, identify specific relations between entities
  - Employed-by
  - Located-at
  - Part-of
- Example:



  - Michael Dell is the CEO of Dell Computer Corporation and lives in Austin Texas.

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).
- Established standard evaluation methodolgy using development (training) and test data and metrics: precision, recall, F-measure.

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*).
- However, output is intended for human consumption, not machine interpretation.
- An IE system for such generated pages allows the web site to be viewed as a structured database.
- An extractor for a semi-structured web site is sometimes referred to as a *wrapper*.
- Process of extracting from such pages is sometimes referred to as *screen scraping*.

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.
  - Programming language
- Some domains may allow multiple extracted templates per document.
  - Multiple apartment listings in one ad

Pattern-Matching Rule Extraction

- Another approach to building IE systems is to use pattern-matching rules for each field to identify the strings to extract for that field.
- When building web extraction systems (wrappers) manually, it is common to write regular expression patterns (in a language like Perl) to identify the desired regions of the text.
- Works well when a fairly fixed local context is sufficient to identify extractions, as in extracting from web pages generated by a program or very stylized text like classified ads.

Regular Expressions

- Language for composing complex patterns from simpler ones.
  - An individual character is a regex.
  - Union: If e_1 and e_2 are regexes, then (e_1 | e_2) is a regex that matches whatever either e_1 or e_2 matches.
  - Concatenation: If e_1 and e_2 are regexes, then e_1 e_2 is a regex that matches a string that consists of a substring that matches e_1 immediately followed by a substring that matches e_2
  - Repetition (Kleene closure): If e_1 is a regex, then e_1* is a regex that matches a sequence of zero or more strings that match e_1

Regular Expression Examples

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

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]

Evaluating IE Accuracy

- Always evaluate performance on independent, manually-annotated test data not used during system development.
- Measure for each test document:
  - Total number of correct extractions in the solution template: *N*
  - Total number of slot/value pairs extracted by the system: *E*
  - Number of extracted slot/value pairs that are correct (i.e. in the solution template): *C*
- Compute average value of metrics adapted from IR:
  - Recall = C/N
  - Precision = C/E
  - F-Measure = Harmonic mean of recall and precision

ACE 2002 Newspaper Corpus

- 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

ACE 2002 Newspaper Corpus

- Compared
  - ERK: string subsequence kernel extractor
  - K4: The tree dependency kernel from [Culotta et. al, 2004].

Method	Precision	Recall	F-measure
ERK	**73.9**	**35.2**	**47.7**
K4	70.3	26.3	38.0

Text Mining

- Automatically extract information from a large corpus to build a large database or knowledge-base of useful information.
- For example, we have used our trained protein interaction extractor to mine biomedical journal abstracts:
  - Input: 753,459 Medline abstracts that reference “human”
  - Output: Database of 6,580 interactions between 3,737 human proteins

Active Learning

- Annotating training documents for each application is difficult and expensive.
- Random selection can waste effort on annotating documents that do not help the learner.
- Best to focus human effort on annotating the *most informative documents*.
- Active learning methods pick only the most informative examples for training.
- At each step, select the example that is estimated to be the most useful for improving the current learner and then ask the human oracle to annotate this example.

Uncertainty Sampling

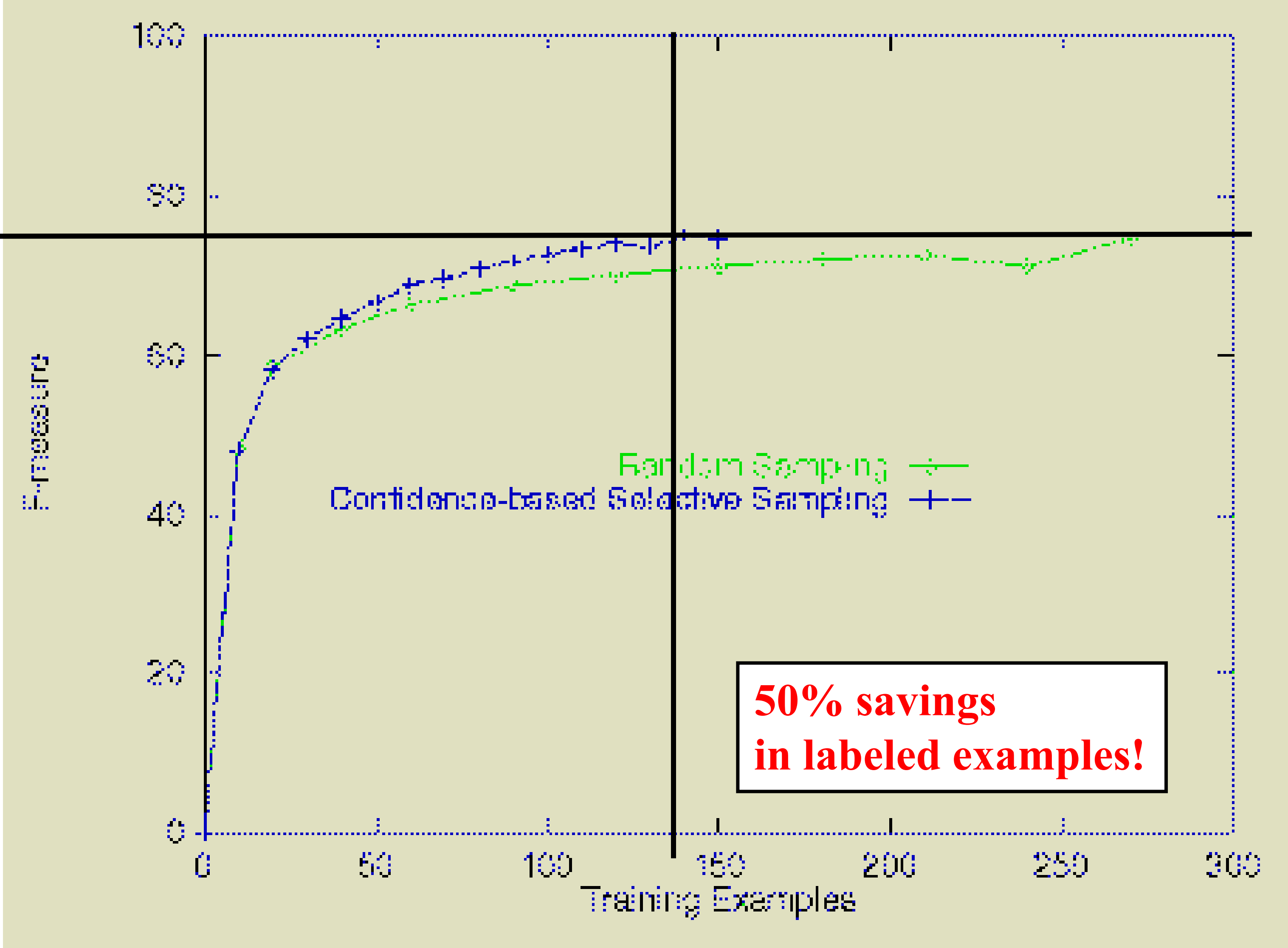
- Assume learned system can provide confidence in its predicted labelings of examples.
- From a pool of unlabeled data, pick as most informative, the unlabelled example about which the current learned system is most uncertain.

```
Let $D$ be a set of unlabeled examples
Until desired accuracy is reached
    Apply current learned system, $L$, to all examples in $D$
    From $D$, select the example, $E$, whose label is most uncertain
    Ask the user to label $E$ and remove it from $D$.
    Add $E$ to the training set and  retrain $L$
```

Rapier Uncertainty Sampling Results

max

perf



50% savings
in labeled examples!

Information Extraction Issues

- Effectively exploiting global information
- Better active learning methods
- Integrating entity and relation extraction
- Unsupervised IE
- Semi-supervised IE
- Adaptation and transfer to new tasks
- Mining extracted data to find cross-document regularities.
- Use resulting mined knowledge to improve IE