CS630 Representing and Accessing Digital Information

Information Extraction

Thorsten Joachims
Cornell University

Based on slides from Prof. Claire Cardie

Information Extraction

- **Introduction**
  - Task definition
  - Evaluation
  - IE system architecture
- **Acquiring extraction patterns**
  - Extraction Rules
    - Semi-automatic methods for extraction from unstructured text
    - Fully automatic methods for extraction from structured text
  - Finite-State Models

Information Needs: Example 1

Question: What was the name of the enchanter played by John Cleese in the movie "Monty Python and the Holy Grail"?

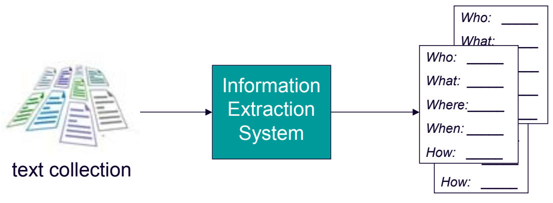
- Ad-hoc IR
- Question answering

Information Needs: Example 2

Question: Describe each movie, character, and actor who played him or her.

motion-picture
title:
date:
plot:
characters: ***movie-role***
actor:
character:
description:
…

Information Extraction



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 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

IE System: Terrorism

SAN SALVADOR, 15 JAN 90 (ACAN-EFE) -- [TEXT] ARMANDO CALDERON SOL, PRESIDENT OF THE NATIONALIST REPUBLICAN ALLIANCE (ARENA), THE RULING SALVADORAN PARTY, TODAY CALLED FOR AN INVESTIGATION INTO ANY POSSIBLE CONNECTION BETWEEN THE MILITARY PERSONNEL IMPLICATED IN THE ASSASSINATION OF JESUIT PRIESTS.

"IT IS SOMETHING SO HORRENDOUS, SO MONSTROUS, THAT WE MUST INVESTIGATE THE POSSIBILITY THAT THE FMLN (FARABUNDO MARTI NATIONAL LIBERATION FRONT) STAGED THIS ASSASSINATION TO DISCREDIT THE GOVERNMENT," CALDERON SOL SAID.

SALVADORAN PRESIDENT ALFREDO CRISTIANI IMPLICATED FOUR OFFICERS, INCLUDING ONE COLONEL, AND FIVE MEMBERS OF THE ARMED FORCES IN THE ASSASSINATION OF SIX JESUIT PRIESTS AND TWO WOMEN ON 16 NOVEMBER AT THE CENTRAL AMERICAN UNIVERSITY.

IE System: Output

1. DATE	- 15 JAN 90
2. LOCATION	EL SALVADOR: CENTRAL AMERICAN UNIVERSITY
3. TYPE	MURDER
4. STAGE OF EXECUTION	ACCOMPLISHED
5. INCIDENT CATEGORY	TERRORIST ACT
6. PERP: INDIVIDUAL ID	"FOUR OFFICERS" "ONE COLONEL" "FIVE MEMBERS OF THE ARMED FORCES"
7. PERP: ORGANIZATION ID	"ARMED FORCES", "FMLN"
8. PERP: CONFIDENCE	REPORTED AS FACT
9. HUM TGT: DESCRIPTION	"JESUIT PRIESTS" "WOMEN"
10. HUM TGT: TYPE	CIVILIAN: "JESUIT PRIESTS" CIVILIAN: "WOMEN"
11. HUM TGT: NUMBER	6: "JESUIT PRIESTS" 2: "WOMEN"
12. EFFECT OF INCIDENT	DEATH: "JESUIT PRIESTS" DEATH: "WOMEN"

IE from Semi-Structured Text

- **Job postings:**
  - Newsgroups: Rapier from austin.jobs
  - Web pages: Flipdog
- **Job resumes:**
  - BurningGlass
  - Mohomine
- **Seminar announcements**
- **Company information from the web**
- **Continuing education course info from the web**
- **University information from the web**
- **Apartment rental ads**

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

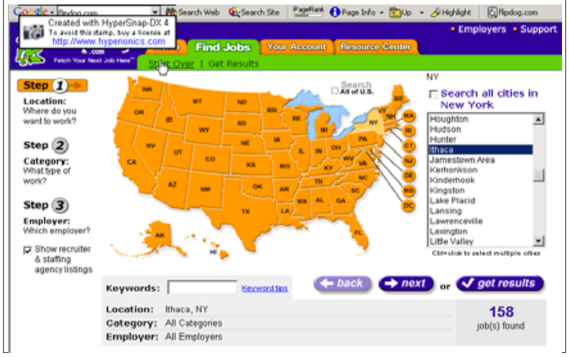
Extracted Job Template

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

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*.**

Flipdog.com



Step 1
Location:
Where do you want to work?

Step 2
Category:
What type of work?

Step 3
Employer:
Which employer?

☑ Show recruiter & staffing agency listings

NY

□ Search all cities in New York

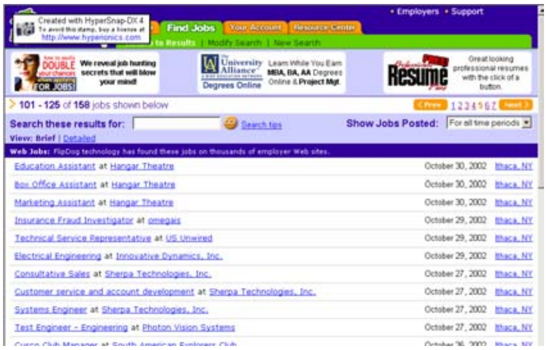
Houghton
Hudson
Hunter
Ithaca
Jamestown Area
Kerhonkson
Kinderhook
Kingston
Lake Placid
Lansing
Lawrenceville
Lexington
Little Valley

Keywords:

Location: Ithaca, NY
Category: All Categories
Employer: All Employers

158
job(s) found

Flipdog.com


Created with HyperSnap-DX 4
To avoid this stamp, buy a license at
http://www.hyperionics.com
• Employers • Support
Find Jobs
Your Account
Resource Center
to Results | Modify Search | New Search
DOUBLE your chances FOR JOBS!
We reveal job hunting secrets that will blow your mind!
University Alliance
Degrees Online
Learn While You Earn MBA, BA, AA Degrees Online & Project Mgt.
Resume
Great looking professional resumes with the click of a button.
Prev
1 2 3 4 5 6 7
Next
Search tips

101 - 125 of 158 jobs shown below

Search these results for:

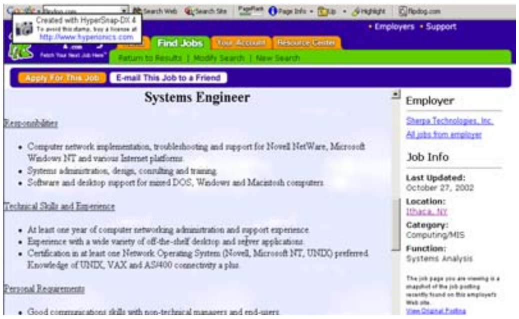
Show Jobs Posted: For all time periods

View: Brief | Detailed

Web Jobs: FlipDog technology has found these jobs on thousands of employer Web sites.

Education Assistant at Hangar Theatre	October 30, 2002	Ithaca, NY
Box Office Assistant at Hangar Theatre	October 30, 2002	Ithaca, NY
Marketing Assistant at Hangar Theatre	October 30, 2002	Ithaca, NY
Insurance Fraud Investigator at omegais	October 29, 2002	Ithaca, NY
Technical Service Representative at US Unwired	October 29, 2002	Ithaca, NY
Electrical Engineering at Innovative Dynamics, Inc.	October 29, 2002	Ithaca, NY
Consultative Sales at Sherpa Technologies, Inc.	October 27, 2002	Ithaca, NY
Customer service and account development at Sherpa Technologies, Inc.	October 27, 2002	Ithaca, NY
Systems Engineer at Sherpa Technologies, Inc.	October 27, 2002	Ithaca, NY
Test Engineer - Engineering at Photon Vision Systems	October 27, 2002	Ithaca, NY
Cusco Club Manager at South American Explorers Club	October 26, 2002	Ithaca, NY

Posting


Search Web
Search Site
PageRank
Page Info
Up
Highlight
flipdog.com
Created with HyperSnap-DX 4
To avoid this stamp, buy a license at
http://www.hyperionics.com
Fetch Your Next Job Here™
• Employers • Support
Find Jobs
Your Account
Resource Center
Return to Results | Modify Search | New Search
Apply For This Job
E-mail This Job to a Friend

Systems Engineer

Responsibilities

- Computer network implementation, troubleshooting and support for Novell NetWare, Microsoft Windows NT and various Internet platforms
- Systems administration, design, consulting and training
- Software and desktop support for mixed DOS, Windows and Macintosh computers

Technical Skills and Experience

- At least one year of computer networking administration and support experience
- Experience with a wide variety of off-the-shelf desktop and server applications
- Certification in at least one Network Operating System (Novell, Microsoft NT, UNIX) preferred Knowledge of UNIX, VAX and AS/400 connectivity a plus

Personal Requirements

- Good communications skills with non-technical managers and end-users

Employer

Sherpa Technologies, Inc.
All jobs from employer

Job Info

Last Updated: October 27, 2002
Location: Ithaca, NY
Category: Computing/MIS
Function: Systems Analysis

The job page you are viewing is a snapshot of the job posting recently found on this employer's Web site.
View Original Posting

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
:
:

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

Template Slot 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

Information Extraction

- **Introduction**
  - Task definition
  - Evaluation
  - IE system architecture
- **Acquiring extraction patterns**
  - Extraction Rules
    - Semi-automatic methods for extraction from unstructured text
    - Fully automatic methods for extraction from structured text
  - Finite-State Models

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

Evaluating IE Systems

- **Evaluate performance on new, manually-annotated test data.**
- **Measure for each test document:**
  - Total number of 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

State of the Art

Unrestricted text:
60-70% R; 65-75% P

Semi-structured text:
90% R/P

MUC
[1991-94]

- **terrorist activities**
- **business joint ventures**
- **microelectronic chip fabrication**
- **changes in corporate management**
- **natural disasters**
- **summarize medical patient records**
- **support automatic classification of legal documents**
- **build knowledge bases from web pages**
- **create job-listing databases from newsgroups**

[Soderland et al. 1995; Craven et al. 1997; Califf & Mooney 1998;…]

IE vs. IR vs. NLP

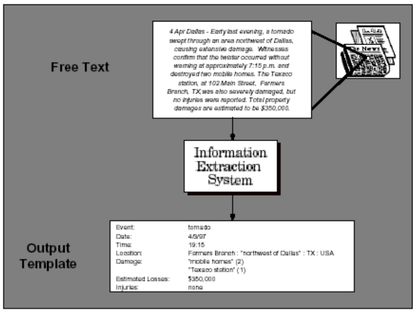
- **IE requires more text-understanding capabilities than the bag-of-words approaches provided by IR techniques**
- **IE requires a more shallow understanding of the text than a natural language understanding system attempting full/deep semantic analysis. IE is domain specific, NLP is general.**
- **IE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

IR, TC < IE < NLP

Information Extraction

- **Introduction**
  - Task definition
  - Evaluation
  - IE system architecture
- **Acquiring extraction patterns**
  - Extraction Rules
    - Semi-automatic methods for extraction from unstructured text
    - Fully automatic methods for extraction from structured text
  - Finite-State Models

Natural Disasters Example

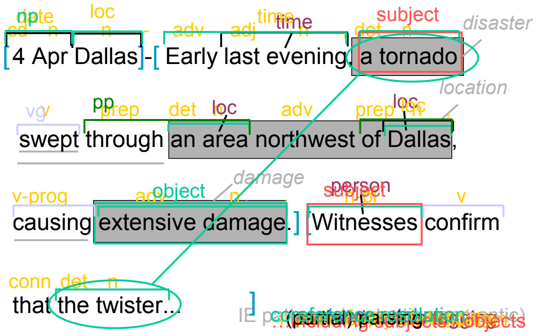


4 Apr Dallas - Early last evening, a tornado swept through an area northwest of Dallas, causing extensive damage. Witnesses confirm that the twister occurred without warning at approximately 7:15 p.m. and destroyed two mobile homes. The Texaco station, at 102 Main Street, Farmers Branch, TX was also severely damaged, but no injuries were reported. Total property damages are estimated to be $350,000.

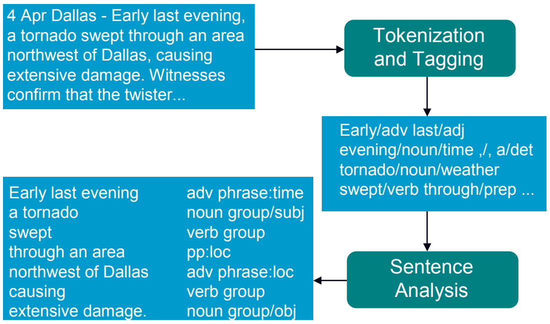
Event:	tornado
Date:	4/3/97
Time:	19:15
Location:	Farmers Branch : "northwest of Dallas" : TX : USA
Damage:	"mobile homes" (2) "Texaco station" (1)
Estimated Losses:	$350,000
Injuries:	none

Figure 1: Information Extraction System in the Domain of Natural Disasters.

IE system components



Stages of Processing



4 Apr Dallas - Early last evening, a tornado swept through an area northwest of Dallas, causing extensive damage. Witnesses confirm that the twister...

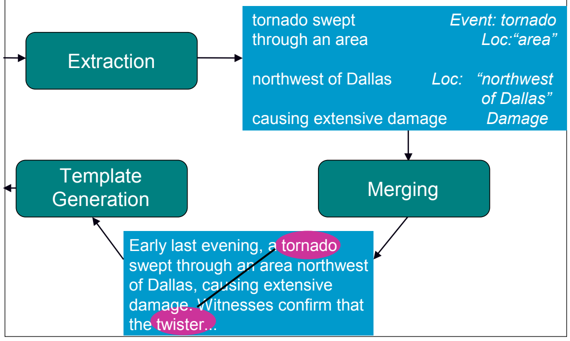
Tokenization and Tagging

Early/adv last/adj evening/noun/time ,/, a/det tornado/noun/weather swept/verb through/prep ...

Sentence Analysis

Early last evening	adv phrase:time
a tornado	noun group/subj
swept	verb group
through an area	pp:loc
northwest of Dallas	adv phrase:loc
causing	verb group
extensive damage.	noun group/obj

Stages of Processing



Extraction

tornado swept through an area	*Event: tornado* *Loc:"area"*
northwest of Dallas	*Loc: "northwest of Dallas"*
causing extensive damage	*Damage*

Merging

Template Generation

Early last evening, a tornado swept through an area northwest of Dallas, causing extensive damage. Witnesses confirm that the twister...

Information Extraction

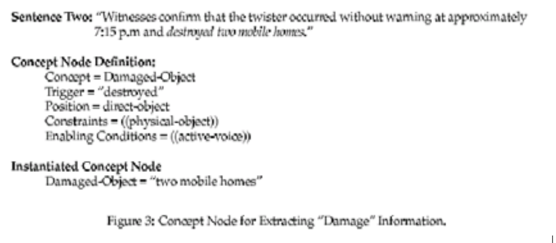
- **Introduction**
  - Task definition
  - Evaluation
  - IE system architecture
- **Acquiring extraction patterns**
  - Extraction Rules
    - Semi-automatic methods for extraction from unstructured text
    - Fully automatic methods for extraction from structured text
  - Finite-State Models

Learning IE Patterns from Examples

- **Goal**
  - Given a training set of documents paired with human-produced filled extraction templates [answer keys],
  - Learn extraction patterns for each slot using an appropriate machine learning algorithm.
- **Options**
  - Memorize the fillers of each slot
  - Generalize the fillers using
    - p-o-s tags?
    - phrase structure (NP, V) and grammatical roles (SUBJ, OBJ)?
    - semantic categories?

Autoslog [Riloff 1993]

- **Learns syntactico-semantic patterns (called "concept nodes")**



Sentence Two: "Witnesses confirm that the twister occurred without warning at approximately 7:15 p.m and *destroyed two mobile homes*."

Concept Node Definition:
Concept = Damaged-Object
Trigger = "destroyed"
Position = direct-object
Constraints = ((physical-object))
Enabling Conditions = ((active-voice))

Instantiated Concept Node
Damaged-Object = "two mobile homes"

Figure 3: Concept Node for Extracting "Damage" Information.

from Cardie [1997]

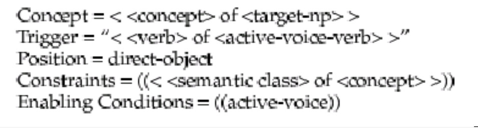
Autoslog Algorithm

- **Noun phrase extraction only**
- **Relies on a small set of pattern templates**

  <active-voice-verb> <direct object>=<target-np>
  <subject>=<target-np> <active-voice-verb>
  <subject>=<target-np> <passive-voice-verb>
  <passive-voice-verb> by <object>=<target-np>
  …
  - Domain-independent
  - So require little modification when switching domains
- **Requires partial parser**
- **Assumes semantic category(ies) for each slot are known, and all potential slot fillers can be tested w.r.t. them**

Autoslog Algorithm

- **Find the sentence from which the noun phrase originated.**
- **Present the sentence to the partial parser.**
- **Apply the pattern templates in order.**
- **When a pattern applies, generate a concept node definition from the matched constituent, its context, the slot type (from the answer key), and the (predefined) semantic class for the filler.**



Concept = < <concept> of <target-np> >
Trigger = "< <verb> of <active-voice-verb> >"
Position = direct-object
Constraints = ((< <semantic class> of <concept> >))
Enabling Conditions = ((active-voice))

Learned Terrorism Patterns

- **<victim> was murdered**
- **<perpetrator> bombed**
- **<perpetrator> attempted to kill**
- **was aimed at <target>**

Natural Disasters Patterns

<subject> = disaster-event (earthquake) registered (active)
registered (active) <direct obj> = magnitude

Yesterday's earthquake registered 6.9 on the Richter scale.

measuring (gerund) <direct obj> = magnitude

measuring 6.9 …

aid (noun)…in/to/for (prep) <obj> = disaster-event-location/ victim

…sending medical aid to Afghanistan…
…sending medical aid to earthquake victims…

Autoslog Algorithm

- **Learns bad patterns as well as good patterns**
  - Too general (e.g. triggered by "is" or "are" or by verbs not tied to the domain)
  - Too specific
  - Just plain wrong
    - Parsing errors
    - Target NPs occur in a prepositional phrase and Autoslog can't determine the trigger (e.g. is it the preceding verb or the preceding NP?)
- **Requires that a person review the proposed extraction patterns, discarding bad ones**
- **No computational linguist needed (?)**
- **Reduced human effort from 1200-1500 hours to ~4.5 hours**
- **F-measure dropped from 50.5 to 48.7 (for one test set); from 41.9 to 41.8 (for a second test set)**

Autoslog-TS

- **Largely unsupervised**
- **Two sets of documents: relevant, not relevant**
- **Apply pattern templates to extract every NP in the texts**
- **Compute *relevance rate* for each pattern *i* :**

Pr (relevant text | text contains i) =
freq of *i* in relevant texts / frequency of *i* in corpus

- **Sort patterns according to relevance rate and frequency**

relevance rate * log (freq)

Covering Algorithms

- **E.g. Crystal** [Soderland et al. 1995]
  - Allows for more complicated patterns
    - Can test target NP or any constituent in its context for
      - presence of any word or sequence of words
      - semantic class of heads or modifiers
- **Covering algorithm: successively generalizes the input examples until the generalization produces errors**
  - Generate the most specific pattern possible for every phrase to be extracted in the training corpus
  - For each pattern, P, find the most similar pattern P' and relax the constraints of each just enough to unify P and P'.
  - Test the new extraction pattern E against the training corpus.
  - If its error rate is < threshold T, add E to the set of patterns, replacing P and P'.
  - Repeat the process on E until the error tolerance is exceeded.
  - Move on to the next pattern, P, in the original set

Information Extraction

- **Introduction**
  - Task definition
  - Evaluation
  - IE system architecture
- **Acquiring extraction patterns**
  - Extraction Rules
    - Semi-automatic methods for extraction from unstructured text
    - Fully automatic methods for extraction from structured text
  - Finite-State Models

Extraction Patterns for Semi-Structured Text

- **If extracting from automatically generated web pages, simple regex patterns usually work.**
- **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>"

slides 47-51 © Ray Mooney

Simple Template Extraction

- **Extract slots in order, starting the search for the filler of the *n*+1 slot where the filler for the *n*th slot ended. Assumes slots always in a fixed order.**
  - Title
  - Author
  - List price
  - …
- **Make patterns specific enough to identify each filler always starting from the beginning of the document.**
- **Rapier system learns three regex-style patterns for each slot: pre-filler, filler, post-filler**

Extraction Patterns for Semi-Structured Text

- **If extracting from more natural, unstructured, human-written text, some NLP will usually 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
  - E.g. Rapier's extraction patterns can use POS or phrase tags.
    - Crime victim:
      - Prefiller: [POS: V, Hypernym: KILL]
      - Filler: [Phrase: NP]

Set Fill Extraction

- **If a slot has a fixed set of pre-specified possible fillers, text categorization can be used to fill the slot.**
  - Job category
  - Company type
- **Treat each of the possible values of the slot as a category, and classify the entire document to determine the correct filler.**

XML and IE

- **If relevant documents were all available in standardized XML format, IE would be unnecessary.**
- **But…**
  - Difficult to develop a universally adopted DTD format for the relevant domain.
  - Difficult to manually annotate documents with appropriate XML tags.
  - Commercial industry may be reluctant to provide data in easily accessible XML format.
- **IE provides a way of automatically transforming semi-structured or unstructured data into an XML compatible format.**

Information Extraction

- **Introduction**
  - Task definition
  - Evaluation
  - IE system architecture
- **Acquiring extraction patterns**
  - Extraction Rules
    - Semi-automatic methods for extraction from unstructured text
    - Fully automatic methods for extraction from structured text
  - Finite-State Models

Finite-State Models

- **Semi-Structured Information Extraction**
  - D. Freitag and A. McCallum. Information Extraction with HMMs and Shrinkage, AAAI Workshop on Machine Learning for Information Extraction, 1999.
  - T. Scheffer, C. Decomain, and S. Wrobel. Active hidden Markov models for information extraction. International Symposium on Intelligent Data Analysis, 2001.
  - J. Lafferty, A. McCallum, and F. Pereira. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. International Conference on Machine Learning, 2001.