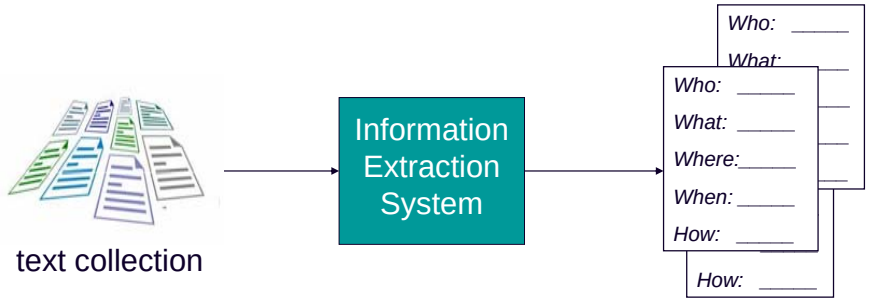
Shallow semantic analysis: Information extraction

- **Introduction**
  - Task definition
  - Evaluation
  - IE system architecture
- **Acquiring extraction patterns**

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... (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

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

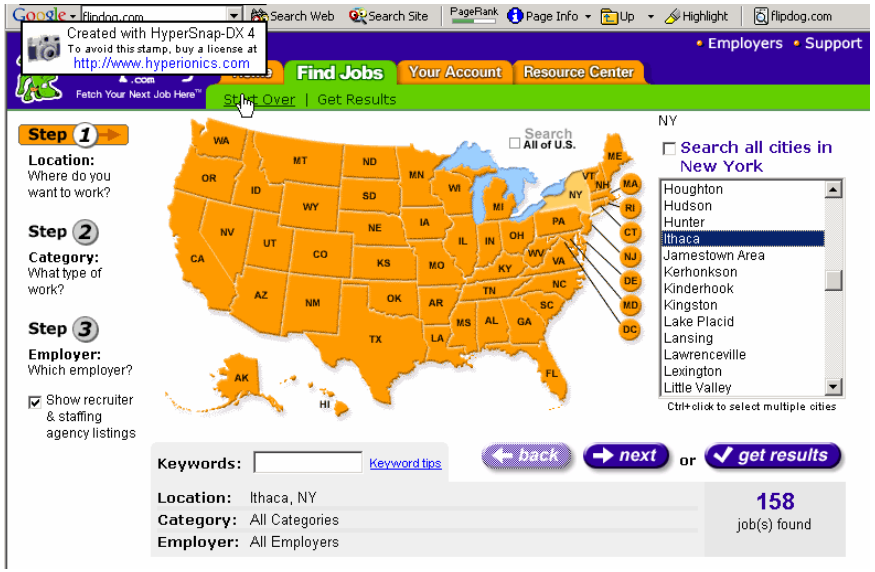
Extracted job template

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

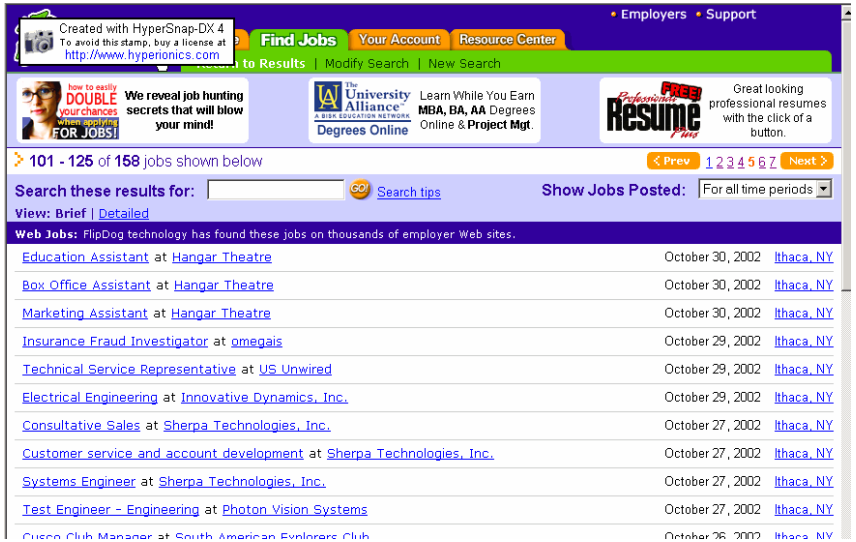
Web extraction

- **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*.**

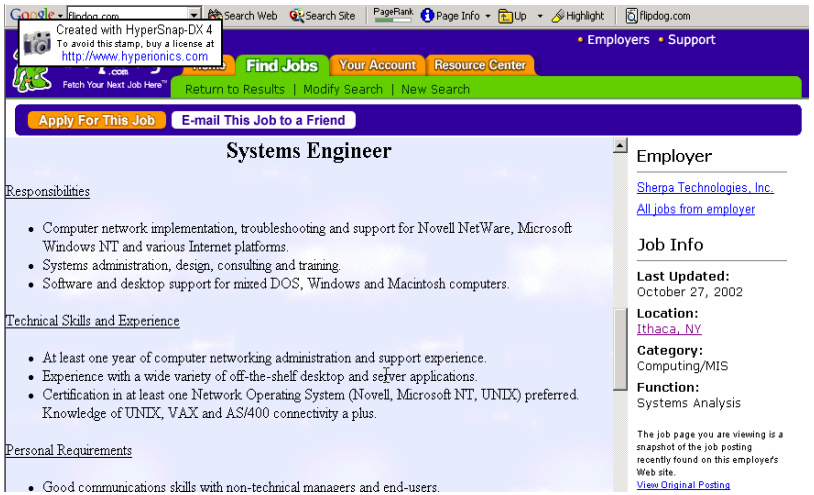
Flipdog.com



Flipdog.com



Posting



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

- **Text fill: substring from the document**
- **Set fill: 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

Shallow semantic analysis: Information extraction

- **Introduction**
  - Task definition
  - Evaluation
  - IE system architecture
- **Acquiring extraction patterns**

Evaluating IE systems

- **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

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 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

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 treated as domain-specific text understanding**
  - key system components need to be re-built for each new domain
  - difficult and time-consuming to build
    - ~6 months/system for IE from unstructured text
  - requires the expertise of computational linguists

Corpus-based statistical/machine learning methods

- acquire linguistic knowledge by applying statistical and symbolic learning methods; derive training examples from the texts themselves
- automate the construction of each IE system component
- improve robustness of final systems while maintaining (or at least approaching) the accuracies of handcrafted systems

Shallow semantic analysis: Information extraction

- **Introduction**
  - Task definition
  - Evaluation
  - IE system architecture
- **Acquiring extraction patterns**
- **Named entity detection**

Natural disasters example

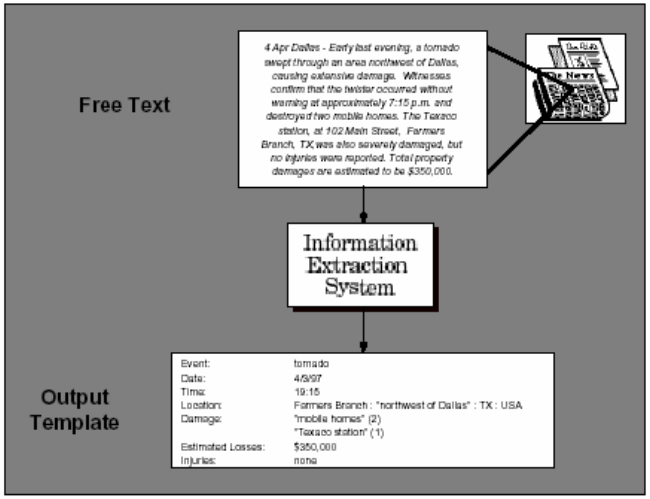
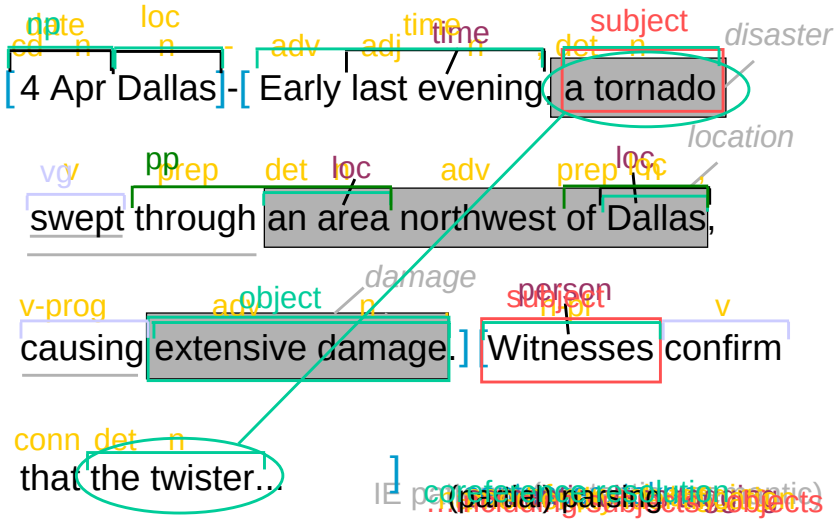
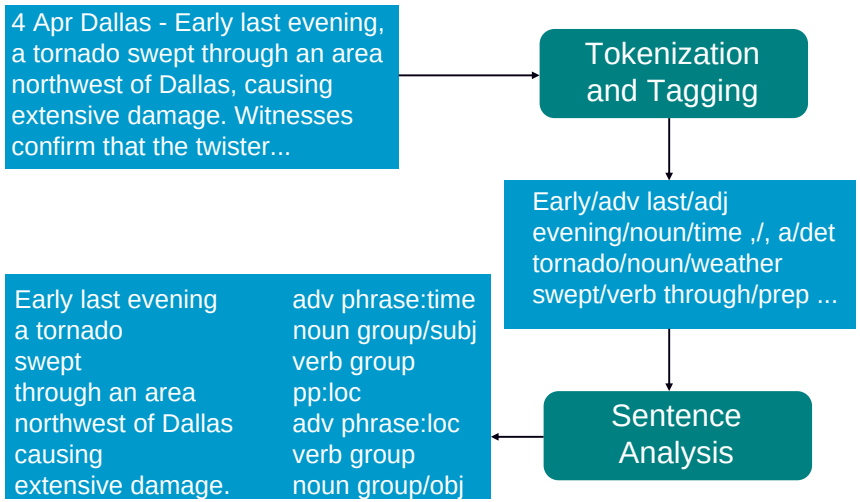


Figure 1: Information Extraction System in the Domain of Natural Disasters.

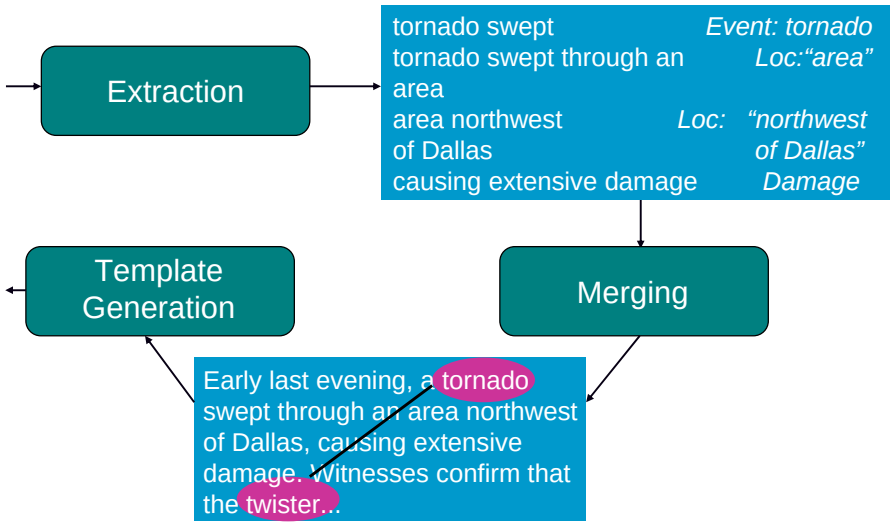
IE system components



Stages of processing



Stages of processing



Shallow semantic analysis: Information extraction

- **Introduction**
  - Task definition
  - Evaluation
  - IE system architecture
- **Acquiring extraction patterns**
  - Learning approaches
    - Semi-automatic methods for extraction from unstructured text
    - Fully automatic methods for extraction from structured text

IE system: input

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"

Issues for learning extraction patterns

- **Training data is difficult to obtain**
  - IE answer keys provide some supervisory information --- string to be extracted and its label --- but often not enough
  - No direct means for learning "set fills"
  - Training examples must encode the output of earlier levels of syntactic and semantic analysis
    - No standard training set available
    - When earlier components change, examples must be regenerated
  - Standard "off-the-shelf" learning algorithms tend to work less well than those specifically tailored to the task

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?

Learning IE patterns

- **Methods vary with respect to**
  - The class of pattern learned (e.g. lexically based regular expression, syntactic-semantic pattern)
  - Training corpus constraints
  - Amount and type of human feedback required
  - Degree of pre-processing necessary
  - Background knowledge presumed

Autoslog [Riloff 1993]

- **Learns syntactico-semantic patterns (originally 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)…to/for (prep) <obj> = disaster-event-location/ victim

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

Advantages/Disadvantages

- **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 original Autoslog 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)

Autoslog-TS

- Human review of learned patterns required
- Also requires labeling the semantic category of the extracted slot filler

Information extraction

- **Introduction**
  - Task definition
  - Evaluation
  - IE system architecture
- **Acquiring extraction patterns**
  - Learning approaches
    - Semi-automatic methods for extraction from unstructured text
    - Fully automatic methods for extraction from structured text

Supervised ML Methods for IE

- **Covering algorithms**
  - E.g. Crystal [Soderland et al., 1995]

- **Sequence tagging algorithms**
  - HMM's
  - Conditional Random Fields (CRFs)
  - Relational Markov Networks (RMNs)