CS674 Natural Language Processing

- **Introduction to Information Extraction**
  - Task definition
  - Evaluation
  - IE system architecture

Monty Python & The Holy Grail

Information needs

- **What was the name of the enchanter played by John Cleese in the movie "Monty Python and the Holy Grail"?**
  - Ad-hoc IR / Google search
  - Question answering systems
- **Describe each character, including the actor who played him or her, in every movie starring John Cleese.**

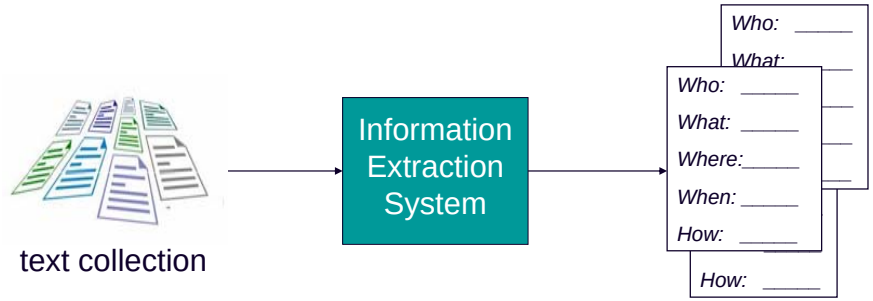
Information extraction need

Describe each character, including the actor who played him or her, in every movie starring John Cleese.

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... (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

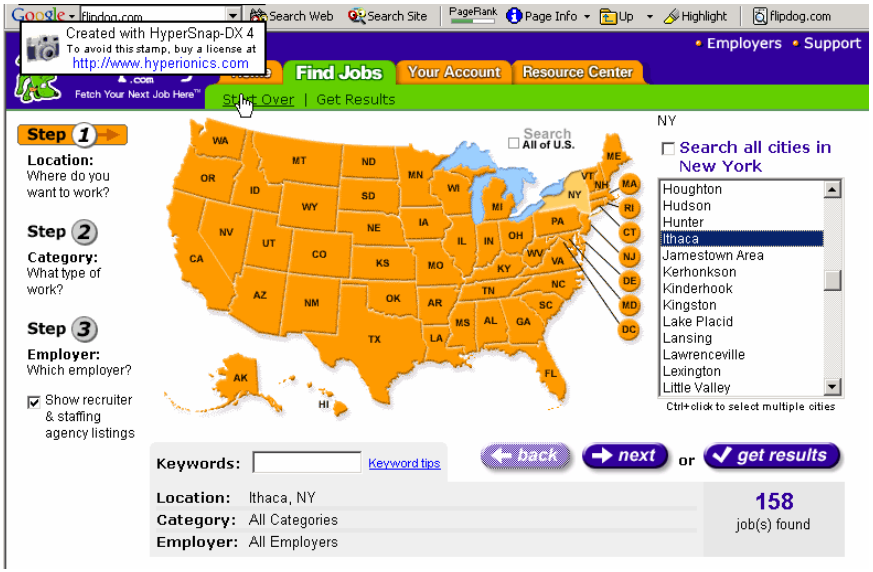
Extracted job template

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

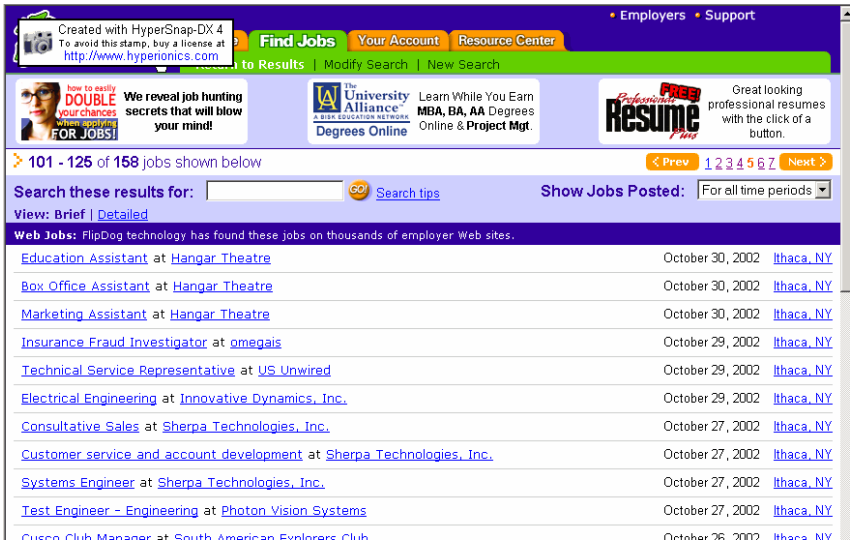
Web extraction

- **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*.**

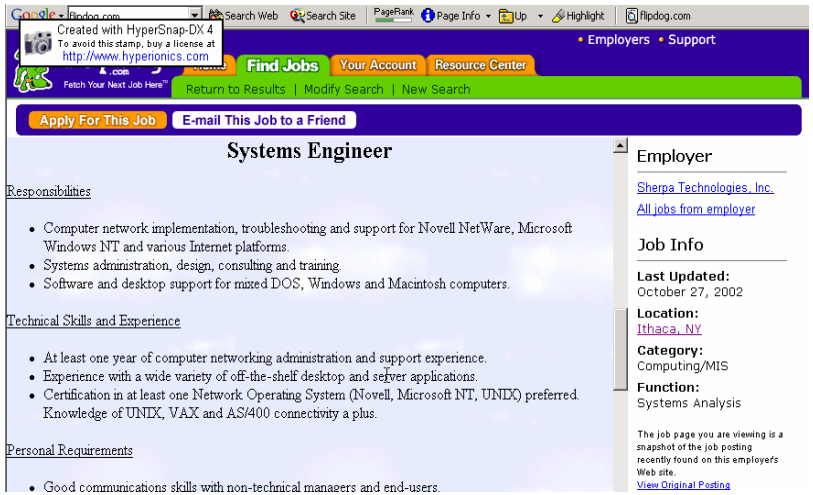
Flipdog.com



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

- **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

MUC

- **DARPA funded significant efforts in IE in the 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

- **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. exist 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**
- **bioinformatics**

[Soderland et al. 1995; Craven et al. 1997; Califf & Mooney 1998;McCallum et al. 2000's; Riloff & X 2000's]

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

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
    - Initially, ~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

Information extraction

- **Introduction**
  - Task definition
  - Evaluation
  - IE system architecture

Natural disasters example

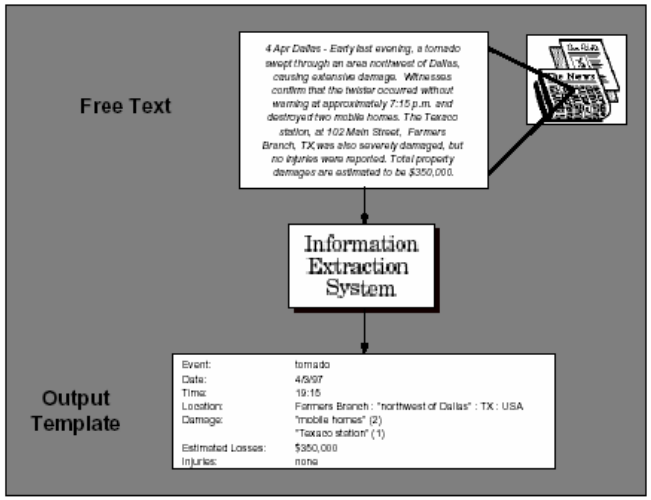
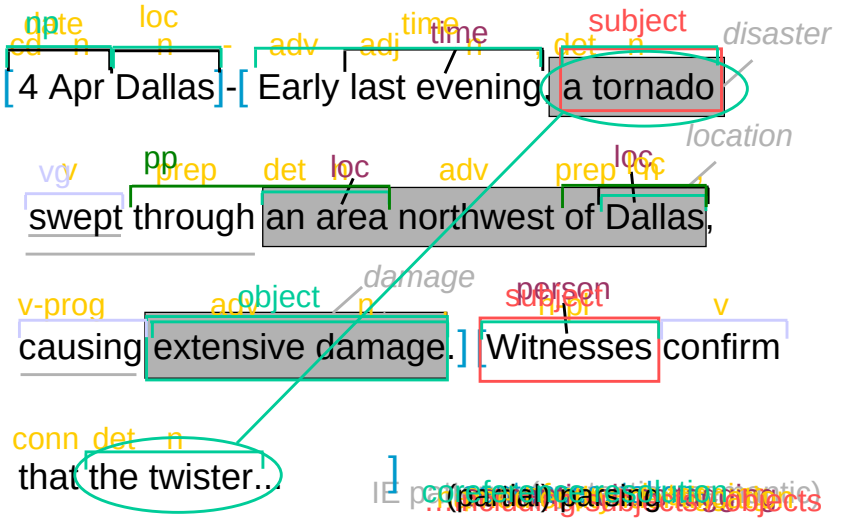


Figure 1: Information Extraction System in the Domain of Natural Disasters.

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 — *Event: tornado*
tornado swept through an area — *Loc:"area"*
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...