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.

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
    • physical-effect:
      • object: entire villages
      • outcome: damaged
      • confidence-marker: Details not 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
SAN SALVADOR, 15 JAN 90 (ACAN-EFE) -- 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.
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 |

- 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

MUC (1991-94)

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

IE system components

Natural Disasters Example

Stages of Processing

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: "Witnesses claim that the train reached Yamanashi at approximately 7:35 p.m. and destoryed three houses."

**Concept Node Definition**
- **Concept** = (trigger, object)
- **Trigger** = “destroyed”
- **Position** = direct object
- **Constraints** = (physical-object)
- **Enabling Conditions** = (time) = 7:35

**Instantiation Concept Node**
- **Trigger** = “destroyed”
- **Object** = “three mobile homes”

![Figure 1: Concept Node forｌacking “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

**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:
  \[ Pr \text{(relevant text | text contains i)} = \frac{\text{freq of i in relevant texts}}{\text{frequency of i in corpus}} \]
- 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\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\“\“b\"---
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