Finding Facts from Text

Information Extraction Technology

DoReMi / University of Helsinki 2006
Roman Yangarber

Finding facts

- **What**
  - What are facts
  - What it means to find facts
- **Why**
  - Why is it important
  - Why is it difficult
- **How**
  - Demos
- **Topics and Research...**
Finding facts

- **Factual information** ⇆ **Textual documents**
  - documents in human language
  - from many sources, on many topics: general news, business, science/medicine, etc.

What is a fact

- **Basic**: Entities and Names: identify all
  - persons, organizations, locations,
  - artefacts, medicines/drugs, diseases, ...

- **Why is even this already useful? Examples:**
  - find all persons related to person X
  - find all companies related to company Y
  - find all diseases in country Z
  - try with IR/Google ...

- **Complex**: Relationships and events
  - how entities relate to each other
    - organizations employ people
  - how they interact:
    - who was affected, how, when, where
What it means to find a fact

- unstructured  ➔ structured representation
- plain text  ➔ spreadsheet, database table

Example: “Executive Search”

- George Garrick, 60 years old, president of the London-based European Information Services Inc., was appointed chief executive officer of Nielsen Marketing Research, USA.
Example: Executive Search

George Garrick, 60 years old, president of the London-based European Information Services Inc., was appointed chief executive officer of Nielsen Marketing Research, USA.

<table>
<thead>
<tr>
<th>Position</th>
<th>Company</th>
<th>Location</th>
<th>Person</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>President</td>
<td>European Information Services, Inc.</td>
<td>London</td>
<td>George Garrick</td>
<td>Out</td>
</tr>
<tr>
<td>CEO</td>
<td>Nielsen Marketing Research</td>
<td>USA</td>
<td>George Garrick</td>
<td>In</td>
</tr>
</tbody>
</table>

Example: Epidemics

Viet Nam: 2 additional deaths confirmed; total now 50

Asia's [human] death toll from avian influenza rose to 50 on Wed 6 Apr 2005, when Vietnamese health officials and a hospital doctor confirmed 2 additional deaths in Viet Nam. A 10-year-old girl, who tested positive for the H5N1 virus, died of lung failure hours after she was admitted to St. Paul's Hospital in Hanoi on 27 Mar 2005, a hospital doctor said on

Rule/Pattern: * confirm N death [in Loc]
Why is it important

- Once facts are in database
  - can search for them more easily
  - can process them intelligently
  - find global patterns and trends

- Certain queries are not served well by keywords alone

- Information explosion

Why IE is useful

- Semantic index into document collection
  - For known scenarios, more reliable than keyword index

- Example: answer query like
  - Where does a given disease appear?
IE and IR

IR: keywords
“Hire / fire / executive…”

Additional processing

IE
“Who works where?”

IE vs IR: Focused Search

- Not *spontaneous, random* search
- Users spend much time on *persistent, focused* search – repeated pursuit of facts that are important in their analysis/research
- User places higher *value* on information related to long-standing interest, to which s/he has a long-term commitment, than on information related to one-time interest
Why is it difficult: e.g., reference

- **Language is complex**
  - George Garrick, 40 years old, has served as president of Sony, Inc. for 13 years.
  - The company announced his resignation effective October.

<table>
<thead>
<tr>
<th>Position</th>
<th>Company</th>
<th>Date</th>
<th>Person</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>President</td>
<td>Sony, Inc.</td>
<td>October, 2000</td>
<td>George Garrick</td>
<td>Out</td>
</tr>
</tbody>
</table>

PULS System

- Pattern–based Understanding and Learning system
- Platform for research and development
Topics

- **structure** of IE system(s)
- problems and challenges of **customization** to new domains
  - formulation of task
  - event definition
  - automatic acquisition of **domain knowledge**

- improving quality of facts via aggregating information across document boundaries
  - downstream processing

PULS System architecture

- **diagram**
Performance

- Accuracy measurement
- Many factors compounded:
  - Name classification
  - Reference resolution
  - Coverage of event patterns
  - Elided elements in events

Example applications

- Database of global epidemics
- Database of corporate executives
- Corporate mergers and acquisitions
- Lawsuits / Legal action, Bankruptcy
- Terrorist attacks
- Natural disasters
- Space launches: rockets, missiles, ...
- Industrial repair/maintenance reports
Example application: ProMED-PULS

- Online incremental database
- Start from plain text
- Extract database records:
  - Disease name
  - Location
  - Date
  - Number of victims
  - Kind of victim/descriptor: people, animals, plants
  - Victim status: sick, dead

Demo

- Epidemic surveillance
- Business news
Current work

- Help in building/customizing knowledge bases
- Favor unsupervised/weakly supervised techniques
  - Reduce manual labor
  - Allows us to use much larger corpora for training
- Unsupervised acquisition of semantic knowledge
  - Learning semantic patterns
  - Learning semantic lexicons/names

Current work

- Cross-document fact validation
  - Notion of Confidence – local vs global
  - Aggregate information across documents
  - Correct errors made in earlier stages of pipeline
- More generally, how can we verify a filled slot
  - No functional dependency between attributes
    - (e.g., any disease can occur anywhere)
- Can be viewed as “deeper understanding” of the domain
  - E.g., reason about epidemics from individual incidents
Applications

- Applications form good base for research
  - Observe performance improvements in real setting
  - Provide large fact base, for cross-document integration

Redundancy

- IE on a large scale
  - in contrast with the traditional study of IE, focusing on the smaller-scale, laboratory setting.
- Applying IE methods to a large collection of text attempts to exploit massive redundancy among the facts contained in the collection
  - Redundancy is inherent in the stream of emerging events, whether the topic is general news, science/medicine, business, etc.
In-depth Topics

- Motivation
  - Problem domain
  - Need semantic knowledge
  - What is a pattern?
  - What is a name?

- Learning semantic patterns
- Learning semantic lexicons
- Learning global trends in extracted data
  - For automatic recovery from errors