Text Mining and Link Analysis

Ronen Feldman
Information Systems Department
Business School
Hebrew University, Jerusalem

Ronen.Feldman@huji.ac.il
Background

• Rapid proliferation of information available in digital format

• People have **less time** to absorb **more information**
The Information Landscape

**Problem**
Lack of tools to handle unstructured data

- Unstructured (Textual): 80%
- Structured (Databases): 20%
Find *Documents* matching the Query

Display *Information* relevant to the Query

- Actual information buried inside documents
- Long lists of documents

- Extract Information from within the documents
- Aggregate over entire collection
Text Mining

**Input**
Documents

**Output**
Patterns
Connections
Profiles
Trends

Seeing the Forest for the Trees
Let Text Mining Do the Legwork for You

<table>
<thead>
<tr>
<th>CAVEMAN</th>
<th>Internet</th>
<th>Text Mining</th>
</tr>
</thead>
<tbody>
<tr>
<td>Find Material</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Read</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Understand</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Consolidate</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Absorb / Act</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
What Is Unique in Text Mining?

• Feature extraction.
• Very large number of features that represent each of the documents.
• The need for background knowledge.
• Even patterns supported by small number of document may be significant.
• Huge number of patterns, hence need for visualization, interactive exploration.
Document Types

• Structured documents
  – Output from CGI

• Semi-structured documents
  – Seminar announcements
  – Job listings
  – Ads

• Free format documents
  – News
  – Scientific papers
Text Representations

- Character Trigrams
- Words
- Linguistic Phrases
- Non-consecutive phrases
- Frames
- Scripts
- Role annotation
- Parse trees
How it Works

Unified Analysis

Role-Based Interfaces

Output

Tagging Platform

Extraction Across Records
Including domain specific entities & relationships

Unstructured Text

Documents
Text, Word, Excel, Email, WWW, PDF

Database

Unstructured Text

XML

Database

Unstructured Text

<PartProblemCondition>
<Part> Fuel Pump </Part>
<Problem> Fails </Problem>
<Condition> Corroded </Condition>
</PartProblemCondition>

Documents
Text, Word, Excel, Email, WWW, PDF

Database

Text Fields

Extract Across Records
Including domain specific entities & relationships

Part Problem Condition

Fuel Pump Fails corroded
Pump Relay Shorts Cold weather
Headlight Fails Running hot
Engine Stalls At low speeds
Components of IE System

- Tokenization
  - Zoning
  - Part of Speech Tagging
  - Sense Disambiguation
  - Shallow Parsing

- Morphological and Lexical Analysis
  - Deep Parsing
  - Anaphora Resolution

- Syntactic Analysis
  - Integration

- Domain Analysis
The Finsbury Park Mosque is the center of radical Muslim activism in England. Through its doors have passed at least three of the men now held on suspicion of terrorist activity in France, England and Belgium, as well as one Algerian man in prison in the United States.

"The mosque's chief cleric, Abu Hamza al-Masri lost two hands fighting the Soviet Union in Afghanistan and he advocates the elimination of Western influence from Muslim countries. He was arrested in London in 1999 for his alleged involvement in a Yemen bomb plot, but was set free after Yemen failed to produce enough evidence to have him extradited."

....
SAP Acquires Virsa for Compliance Capabilities

By Renee Boucher Ferguson

April 3, 2006

Honing its software compliance skills, SAP announced April 3 the acquisition of Virsa Systems, a privately held company that develops risk management software.

Terms of the deal were not disclosed.

SAP has been strengthening its ties with Microsoft over the past year or so. The two software giants are working on a joint development project, Mendocino, which will integrate some MySAP ERP (enterprise resource planning) business processes with Microsoft Outlook. The first product is expected in 2007.

"Companies are looking to adopt an integrated view of governance, risk and compliance instead of the current reactive and fragmented approach," said Shai Agassi, president of the Product and Technology Group and executive board member of SAP, in a statement. "We welcome Virsa employees, partners and customers to the SAP family."
SAP has been strengthening its ties with Microsoft over the past year or so. The two software giants are working on a joint development project: Mendocino, which will integrate some MySAP ERP (enterprise resource planning) business processes with Microsoft Outlook. The first product is expected in 2007.

"Companies are looking to adopt an integrated view of governance, risk and compliance instead of the current reactive and fragmented approach," said Shai Agassi, president of the Product and Technology Group and executive board member of SAP, in a statement. "We welcome Virsa employees, partners and customers to the SAP family."
Honing its software compliance skills, SAP announced April 3 the acquisition of Virsa Systems, a privately held company that develops risk management software. Terms of the deal were not disclosed.

SAP has been strengthening its ties with Microsoft over the past year or so. The two software giants are working on a joint development project, Mendocino, which will integrate some MySAP ERP (enterprise resource planning) business processes with Microsoft Outlook. The first product is expected in 2007.

"Companies are looking to adopt an integrated view of governance, risk and compliance instead of the current reactive and fragmented approach," said Shai Agassi, president of the Product and Technology Group and executive board member of SAP, in a statement. "We welcome Virsa employees, partners and customers to the SAP family."
Leveraging Content Investment

Any type of content
• Unstructured textual content (current focus)
• Structured data; audio; video (future)

In any format
• Documents; PDFs; E-mails; articles; etc
• “Raw” or categorized
• Formal; informal; combination

From any source
• WWW; file systems; news feeds; etc.
• Single source or combined sources
Link Analysis in Textual Networks
A Complete Link Analysis System
Types of Link Analysis Questions:

• Who is Central in the organization?
• Which 3 individuals’ removal or incapacitation would sever this drug-supply network?
• What role or roles does specific individual appear to be playing in a given organization?
• Which communication channels without a terrorist organization are worth monitoring?
• What significant changes have taken place in the supply operation of a given organization since this time last year?
Sample LD Queries In the Terror Domain

- People
  - What <Links> connect <Person1> and <Person2>?  

- Organizations
  - What <Organizations> are in the <path> between <Person1> and <Person2>?  

- Social Relations
  - Does <Person1> have influence over <Person2>?  

- Temporal Relations
  - Was there a <Time-Period> when both <Person1> and <Person2> and belonged to <Organization1>?  

- Composite Relations
  - Could <Person1>, or any associate of <Person1>, have met with <Person2>, or any associate of <Person2>, in <Location>, during <Time-Period>?  

- Scenarios
  - Given that these <Events> have occurred involving <Person1> and <Person1>, what <Scenarios> could they be executing?
### THE HIJACKERS...

<table>
<thead>
<tr>
<th>American Airlines 11</th>
<th>United Airlines 175</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Crashed into WTC (north)</strong></td>
<td><strong>Crashed into WTC (south)</strong></td>
</tr>
<tr>
<td>Mohamed Atta (Egyptian)</td>
<td>Marwan al-Shehhi (United Arab Emirates)</td>
</tr>
<tr>
<td>Received pilot training</td>
<td>Received pilot training</td>
</tr>
<tr>
<td>Waleed M. Alshehri (Saudi)</td>
<td>Fayez Ahmed (Believed to be Saudi)</td>
</tr>
<tr>
<td>Commercial pilot</td>
<td>No picture available</td>
</tr>
<tr>
<td>Wail Alshahri (Saudi)</td>
<td>Ahmed Alghamdi (Possibly Saudi)</td>
</tr>
<tr>
<td>Possible pilot training</td>
<td></td>
</tr>
<tr>
<td>Satam al-Suqami (Nationality unknown)</td>
<td>Hamza Alghamdi (Believed to be Saudi)</td>
</tr>
<tr>
<td></td>
<td>Possible pilot training</td>
</tr>
<tr>
<td></td>
<td>Mohaid Alshehri (Nationality unknown)</td>
</tr>
<tr>
<td></td>
<td>Possible pilot training</td>
</tr>
<tr>
<td>Abdulaziz Alomari* (Saudi)</td>
<td></td>
</tr>
<tr>
<td>Possible pilot training</td>
<td></td>
</tr>
<tr>
<td><strong>Crashed into Pentagon</strong></td>
<td><strong>Crashed in Pennsylvania</strong></td>
</tr>
<tr>
<td>American Airlines 77</td>
<td>United Airlines 93</td>
</tr>
<tr>
<td>Khalid al-Midhar (Nationality unknown)</td>
<td>Ziad Jarrah (Lebanese)</td>
</tr>
<tr>
<td>Received pilot training</td>
<td>Received pilot training</td>
</tr>
<tr>
<td>Majed Moqed (Nationality unknown)</td>
<td>Ahmed Alhaznawi (Saudi)</td>
</tr>
<tr>
<td>Salem Alhamzi* (Saudi)</td>
<td>Ahmed Alnami (Nationality unknown)</td>
</tr>
<tr>
<td>Possible pilot training</td>
<td>Saeed Alghamdi* (Seems to be Saudi)</td>
</tr>
<tr>
<td>Nawaf Alhamzi* (Saudi)</td>
<td></td>
</tr>
<tr>
<td>Hani Hanjour (Saudi)</td>
<td></td>
</tr>
</tbody>
</table>

### AND HOW THEY WERE CONNECTED

<table>
<thead>
<tr>
<th><strong>Attended same technical college</strong></th>
<th><strong>Known to be together in week before attacks</strong></th>
<th><strong>Last known address</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Hamburg, Germany</td>
<td>Mohamed Atta</td>
<td>Hollywood, Florida</td>
</tr>
<tr>
<td>Mohamed Atta</td>
<td>Marwan al-Shehhi</td>
<td>Marwan al-Shehhi</td>
</tr>
<tr>
<td>Ziad Jarrah</td>
<td></td>
<td>Waleed M. Alshehri</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Wail Alshahri</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Satam al-Suqami</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ziad Jarrah</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Hani Hanjour</td>
</tr>
<tr>
<td><strong>Took flight classes together</strong></td>
<td><strong>Attended a gym in Maryland (Sept 2-6), also seen dining together</strong></td>
<td><strong>Other cities in Florida</strong></td>
</tr>
<tr>
<td>Pilot schools in Florida</td>
<td>Khalid al-Midhar</td>
<td>Mohamed Atta</td>
</tr>
<tr>
<td>Mohamed Atta</td>
<td>Majed Moqed</td>
<td>Marwan al-Shehhi</td>
</tr>
<tr>
<td>Marwan al-Shehhi</td>
<td>Salem Alhamzi</td>
<td>Waleed M. Alshehri</td>
</tr>
<tr>
<td></td>
<td>Nawaf Alhamzi</td>
<td>Mohaid Alshehri</td>
</tr>
<tr>
<td></td>
<td>Hani Hanjour</td>
<td></td>
</tr>
<tr>
<td><strong>Pilot schools in San Diego</strong></td>
<td><strong>Bought flight tickets using same address</strong></td>
<td><strong>Outside Florida</strong></td>
</tr>
<tr>
<td>Khalid al-Midhar</td>
<td>Mohamed Atta*</td>
<td>Satam al-Suqami</td>
</tr>
<tr>
<td>Nawaf Alhamzi</td>
<td>Marwan al-Shehhi</td>
<td>Hamza Alghamdi</td>
</tr>
<tr>
<td></td>
<td>Abdulaziz Alomari*</td>
<td>Majed Moqed</td>
</tr>
<tr>
<td></td>
<td>Also used same credit card</td>
<td>Salem Alhamzi</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Nawaf Alhamzi</td>
</tr>
<tr>
<td><strong>Bought flight tickets together</strong></td>
<td><strong>Bought from the same travel agent in Florida</strong></td>
<td></td>
</tr>
<tr>
<td>Mohamed Atta</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ziad Jarrah</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ahmed Alhaznawi</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Running Example
Kamada and Kawai's (KK) Method
Finding the shortest Path (from Atta)
A better Visualization
Centrality
Applications of Centrality

• Targeting
  – Betweenness
  – Business

• Identification of Network Vulnerability
  – Betweenness
  – Point Strength
  – Business
Partitioning of networks
Cores of the Hijackers Graph
Structural Equivalence in the Hijackers
EDis between each pair of terrorists

|     | 1     | 2     | 3     | 4     | 5     | 6     | 7     | 8     | 9     | 10    | 11    | 12    | 13    | 14    | 15    | 16    | 17    | 18    | 19    |
|-----|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| 1   | Nawaq Alhamzi | 0.0 | 1.4 | 9.3 | 9.6 | 3.7 | 2.8 | 3.7 | 4.2 | 2.4 | 4.9 | 3.7 | 3.7 | 3.7 | 4.7 | 4.7 | 4.7 | 4.9 | 3.2 | 3.2 |
| 2   | Khalid Al-Midhar | 1.4 | 0.0 | 9.4 | 8.4 | 4.0 | 2.4 | 3.5 | 4.0 | 2.8 | 4.7 | 4.0 | 4.0 | 4.0 | 4.5 | 4.5 | 4.5 | 4.7 | 3.5 | 3.5 |
| 3   | Mohamed Atta | 9.3 | 9.4 | 0.0 | 2.4 | 9.8 | 9.7 | 10.2 | 7.5 | 9.4 | 7.6 | 9.8 | 9.4 | 9.8 | 7.5 | 7.5 | 7.5 | 7.6 | 9.6 | 9.6 |
| 4   | Marwan Al-Shehhi | 9.6 | 8.4 | 2.4 | 0.0 | 10.7 | 8.7 | 9.3 | 7.6 | 9.5 | 7.2 | 9.5 | 9.1 | 9.5 | 7.1 | 7.1 | 7.1 | 7.2 | 9.3 | 9.3 |
| 5   | Hani Hanjour | 3.7 | 4.0 | 9.8 | 10.7 | 0.0 | 3.2 | 2.0 | 5.3 | 4.0 | 6.8 | 6.0 | 6.3 | 6.3 | 6.6 | 6.6 | 6.6 | 6.8 | 6.0 | 6.0 |
| 6   | Majed Moqed | 2.8 | 2.4 | 9.7 | 8.7 | 3.2 | 0.0 | 1.4 | 4.2 | 3.2 | 5.3 | 4.7 | 5.1 | 5.1 | 5.1 | 5.1 | 5.1 | 5.3 | 4.7 | 4.7 |
| 7   | Salem Alhamzi | 3.7 | 3.5 | 10.2 | 9.3 | 2.0 | 1.4 | 0.0 | 4.9 | 4.0 | 6.2 | 5.7 | 6.0 | 6.0 | 6.0 | 6.0 | 6.0 | 6.2 | 5.7 | 5.7 |
| 8   | Abdulaziz Alomari | 4.2 | 4.0 | 7.5 | 7.6 | 5.3 | 4.2 | 4.9 | 0.0 | 4.0 | 3.2 | 4.9 | 4.5 | 5.3 | 2.8 | 2.8 | 2.8 | 3.2 | 4.9 | 4.9 |
| 9   | Ahmed Alghamdi | 2.4 | 2.8 | 9.4 | 9.5 | 4.0 | 3.2 | 4.0 | 4.0 | 0.0 | 4.7 | 3.5 | 3.5 | 3.5 | 4.5 | 4.5 | 4.5 | 4.7 | 3.5 | 3.5 |
| 10  | Ziad Jarrahi | 4.9 | 4.7 | 7.6 | 7.2 | 6.8 | 5.3 | 6.2 | 3.2 | 4.7 | 0.0 | 4.2 | 3.7 | 4.7 | 1.4 | 1.4 | 1.4 | 0.0 | 4.2 | 4.2 |
| 11  | Hamza Alghamdi | 3.7 | 4.0 | 9.8 | 9.5 | 6.0 | 4.7 | 5.7 | 4.9 | 3.5 | 4.2 | 0.0 | 2.8 | 2.8 | 4.5 | 4.5 | 4.5 | 4.2 | 2.0 | 2.0 |
| 12  | Mohald Alshehri | 3.7 | 4.0 | 9.4 | 9.1 | 6.3 | 5.1 | 6.0 | 4.5 | 3.5 | 3.7 | 2.8 | 0.0 | 2.8 | 3.5 | 3.5 | 3.5 | 3.7 | 2.8 | 2.8 |
| 13  | Saeed Alghamdi | 3.7 | 4.0 | 9.8 | 9.5 | 6.3 | 5.1 | 6.0 | 5.3 | 3.5 | 4.7 | 2.8 | 2.8 | 0.0 | 4.5 | 4.5 | 4.5 | 4.7 | 2.0 | 2.0 |
| 14  | Satam Al Suqami | 4.7 | 4.5 | 7.5 | 7.1 | 6.6 | 5.1 | 6.0 | 2.8 | 4.5 | 1.4 | 4.5 | 3.5 | 4.5 | 0.0 | 0.0 | 0.0 | 1.4 | 4.0 | 4.0 |
| 15  | Waleed M. Alshehri | 4.7 | 4.5 | 7.5 | 7.1 | 6.6 | 5.1 | 6.0 | 2.8 | 4.5 | 1.4 | 4.5 | 3.5 | 4.5 | 0.0 | 0.0 | 0.0 | 1.4 | 4.0 | 4.0 |
| 16  | Wail Alshehri | 4.7 | 4.5 | 7.5 | 7.1 | 6.6 | 5.1 | 6.0 | 2.8 | 4.5 | 1.4 | 4.5 | 3.5 | 4.5 | 0.0 | 0.0 | 0.0 | 1.4 | 4.0 | 4.0 |
| 17  | Fayez Ahmed | 4.9 | 4.7 | 7.6 | 7.2 | 6.8 | 5.3 | 6.2 | 3.2 | 4.7 | 0.0 | 4.2 | 3.7 | 4.7 | 1.4 | 1.4 | 1.4 | 0.0 | 4.2 | 4.2 |
| 18  | Ahmed Alnami | 3.2 | 3.5 | 9.6 | 9.3 | 6.0 | 4.7 | 5.7 | 4.9 | 3.5 | 4.2 | 2.0 | 2.8 | 2.0 | 4.0 | 4.0 | 4.0 | 4.2 | 0.0 | 0.0 |
| 19  | Ahmed Alhaznawi | 3.2 | 3.5 | 9.6 | 9.3 | 6.0 | 4.7 | 5.7 | 4.9 | 3.5 | 4.2 | 2.0 | 2.8 | 2.0 | 4.0 | 4.0 | 4.0 | 4.2 | 0.0 | 0.0 |
Clustering based on structural equivalence
Block Modeling
What is Block Modeling

• Block modeling is an analysis technique that serves to find clusters of vertices that behave in a similar way. Block modeling was introduced by The technique is fairly general and can use a variety of equivalence relations between the vertices. The general block modeling problem is composed of two subproblems:
  – Performing clustering of the vertices; each cluster serves as a block.
  – Calculating the links (and their associated value) between the blocks.
Visualization of the predicates
## Block Model of 4 blocks

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>com</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2</td>
<td>-</td>
<td>com</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>3</td>
<td>com</td>
<td>com</td>
<td>null</td>
<td>-</td>
</tr>
<tr>
<td>4</td>
<td>-</td>
<td>null</td>
<td>null</td>
<td>null</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>0</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>-</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>4</td>
</tr>
</tbody>
</table>
The related graph
Shrinking of the network
# Block Model of 6 blocks

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>com</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2</td>
<td>null</td>
<td>Null</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>3</td>
<td>null</td>
<td>Null</td>
<td>com</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>4</td>
<td>com</td>
<td>Com</td>
<td>Null</td>
<td>com</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>5</td>
<td>com</td>
<td>Null</td>
<td>Null</td>
<td>null</td>
<td>Com</td>
<td>-</td>
</tr>
<tr>
<td>6</td>
<td>null</td>
<td>Null</td>
<td>Null</td>
<td>null</td>
<td>Null</td>
<td>com</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>-</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
The related graph
Information Extraction

Theory and Practice
What is Information Extraction?

• IE does not indicate which documents need to be read by a user, it rather extracts pieces of information that are salient to the user's needs.
• Links between the extracted information and the original documents are maintained to allow the user to reference context.
• The kinds of information that systems extract vary in detail and reliability.
• Named entities such as persons and organizations can be extracted with reliability in the 90th percentile range, but do not provide attributes, facts, or events that those entities have or participate in.
Relevant IE Definitions

• Entity: an object of interest such as a person or organization.
• Attribute: a property of an entity such as its name, alias, descriptor, or type.
• Fact: a relationship held between two or more entities such as Position of a Person in a Company.
• Event: an activity involving several entities such as a terrorist act, airline crash, management change, new product introduction.
## IE Accuracy by Information Type

<table>
<thead>
<tr>
<th>Information Type</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entities</td>
<td>90-98%</td>
</tr>
<tr>
<td>Attributes</td>
<td>80%</td>
</tr>
<tr>
<td>Facts</td>
<td>60-70%</td>
</tr>
<tr>
<td>Events</td>
<td>50-60%</td>
</tr>
</tbody>
</table>
## MUC Conferences

<table>
<thead>
<tr>
<th>Conference</th>
<th>Year</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>MUC 1</td>
<td>1987</td>
<td>Naval Operations</td>
</tr>
<tr>
<td>MUC 2</td>
<td>1989</td>
<td>Naval Operations</td>
</tr>
<tr>
<td>MUC 3</td>
<td>1991</td>
<td>Terrorist Activity</td>
</tr>
<tr>
<td>MUC 4</td>
<td>1992</td>
<td>Terrorist Activity</td>
</tr>
<tr>
<td>MUC 5</td>
<td>1993</td>
<td>Joint Venture and Micro Electronics</td>
</tr>
<tr>
<td>MUC 6</td>
<td>1995</td>
<td>Management Changes</td>
</tr>
<tr>
<td>MUC 7</td>
<td>1997</td>
<td>Spaces Vehicles and Missile Launches</td>
</tr>
</tbody>
</table>
Applications of Information Extraction

- Routing of Information
- Infrastructure for IR and for Categorization (higher level features)
- Event Based Summarization.
- Automatic Creation of Databases and Knowledge Bases.
Approaches for Building IE Systems

• Knowledge Engineering Approach
  – Rules are crafted by linguists in cooperation with domain experts.
  – Most of the work is done by inspecting a set of relevant documents.
  – Can take a lot of time to fine tune the rule set.
  – Best results were achieved with KB based IE systems.
  – Skilled/gifted developers are needed.
  – A strong development environment is a MUST!
Approaches for Building IE Systems

• Automatically Trainable Systems
  – The techniques are based on pure statistics and almost no linguistic knowledge
  – They are language independent
  – The main input is an annotated corpus
  – Need a relatively small effort when building the rules, however creating the annotated corpus is extremely laborious.
  – Huge number of training examples is needed in order to achieve reasonable accuracy.
  – Hybrid approaches can utilize the user input in the development loop.
Mining Discussion Boards
Connections between Running Shoes
The Most Central Shoe
Connecting Cars and Terms
Clustering Results

- **European “Low End”**
  - BMW 3-Series, Mercedes-Benz C-Class, Jaguar X-Type, Saab 9-3, Saab 9-5, Volvo S4, Volvo S6, Volvo S70, Volvo S8
- **Lexus**
  - Lexus ES, Lexus LS400, Lexus LS430, Lexus LS460, Lexus LS500, Lexus LS600h
- **Acura+**
  - Acura MDX, Acura RDX, Acura TL, Acura TSX, Jaguar S-Type, Jaguar XJ, Jaguar XK
- **Audi**
  - Audi A3, Audi A4, Audi A8, Audi Q7, Audi R8, Audi RS4, Audi S4, Audi S6, Audi S8, Audi TT
- **Infinity+**
- **European “High End” +**
  - Acura RL, Audi A6, BMW 5-Series, BMW 7-Series, Infiniti M35, Infiniti M45, Lexus GS, Mercedes-Benz S-Class, Mercedes-Benz E-Class
MDS of Brands Lift
Dendogram on Brands Lift

- Acura
- Infiniti 12
- Lexus 15
- Audi 2
- BMW 8
- Mercedes 18
- Cadillac 5
- Jaguar 13
- Lincoln 16
- Saab 23
- Volvo 28
- Buick 4
- Oldsmobile 21
- Pontiac 22
- Chevrolet 5
- Saturn 24
- Chrysler 7
- Dodge 0
- Hyundai 11
- Kia 14
- Mitsubishi 9
- Suzuki 25
- Volkswagen 27
- Ford 9
- Mazda 17
- Honda 10
- Toyota 26
- Nissan 20
Company Lifts - 6-cluster solution

- **Low end/ Korean and Japanese**
  
  Hyundai, Kia, Mitsubishi, Suzuki

- **American Classics**

  Buick, Chevrolet, Oldsmobile, Pontiac, Saturn

- **German and Luxury Japanese**

  Acura, Audi, BMW, Infiniti, Lexus, Mercedes

- **Rugged**

  Chrysler, Dodge, Volkswagen

- **Luxury American+**

  Cadillac, Jaguar, Lincoln, Saab, Volvo

- **Main Stream**

  Ford, Honda, Mazda, Nissan, Toyota
Digging in Deeper – Main Stream
MDS of Main Stream Japanese Car Models - Lift
Digging in Deeper – Luxury Models
MDS of Luxury Cars Models (Lifts)
Self-Supervised Relation Learning from the Web
KnowItAll (KIA)

• KnowItAll is a system developed at University of Washington by Oren Etzioni and colleagues (Etzioni, Cafarella et al. 2005).

• KnowItAll is an autonomous, domain-independent system that extracts facts from the Web. The primary focus of the system is on extracting entities (unary predicates), although KnowItAll is able to extract relations (N-ary predicates) as well.

• The input to KnowItAll is a set of entity classes to be extracted, such as “city”, “scientist”, “movie”, etc., and the output is a list of entities extracted from the Web.
KnowItAll’s Relation Learning

• The base version of KnowItAll uses only the generic handwritten patterns. The patterns are based on a general Noun Phrase (NP) tagger.

• For example, here are the two patterns used by KnowItAll for extracting instances of the Acquisition(Company, Company) relation:
  – NP2 "was acquired by" NP1
  – NP1 "'s acquisition of" NP2

• And the following are the three patterns used by KnowItAll for extracting the MayorOf(City, Person) relation:
  – NP ", mayor of" <city>
  – <city> "'s mayor" NP
  – <city> "mayor" NP
SRES

• SRES (Self-Supervised Relation Extraction System) which learns to extract relations from the web in an unsupervised way.

• The system takes as input the name of the relation and the types of its arguments and returns as output a set of instances of the relation extracted from the given corpus.
SRES Architecture

**Input:** Target Relations Definitions

- Sentences
- Keywords

**Output:** Ex extractions

- Seeds Generator
- Pattern Learner
- Instance Extractor
- Classifier

**Optional:** NER Filter

**Web Gatherer**
Seeds for Acquisition

• Oracle – PeopleSoft
• Oracle – Siebel Systems
• PeopleSoft – J.D. Edwards
• Novell – SuSE
• Sun – StorageTek
• Microsoft – Groove Networks
• AOL – Netscape
• Microsoft – Vicinity
• San Francisco-based Vector Capital – Corel
• HP – Compaq
Major Steps in Pattern Learning

• The sentences containing the arguments of the seed instances are extracted from the large set of sentences returned by the Sentence Gatherer.
• Then, the patterns are learned from the seed sentences.
  – We need to generate automatically
    • Positive Instances
    • Negative Instances
• Finally, the patterns are post-processed and filtered.
Positive Instances

• The positive set of a predicate consists of sentences that contain an instance of the predicate, with the actual instance’s attributes changed to “<AttrN>”, where $N$ is the attribute index.

• For example, the sentence
  – “The Antitrust Division of the U.S. Department of Justice evaluated the likely competitive effects of Oracle's proposed acquisition of PeopleSoft.”

• will be changed to
  – “The Antitrust Division… …..effects of <Attr1>'s proposed acquisition of <Attr2>.”
Negative Instances II

• We generate the negative set from the sentences in the positive set by changing the assignment of one or both attributes to other suitable entities in the sentence.

• In the shallow parser based mode of operation, any suitable noun phrase can be assigned to an attribute.
Examples

• The Positive Instance
  – “The Antitrust Division of the U.S. Department of Justice evaluated the likely competitive effects of <Attr1>’s proposed acquisition of <Attr2>”

• Possible Negative Instances
  – <Attr1> of the <Attr2> evaluated the likely…
  – <Attr2> of the U.S. … …acquisition of <Attr1>
  – <Attr1> of the U.S. … …acquisition of <Attr2>
  – The Antitrust Division of the <Attr1> ….. acquisition of <Attr2>”
Additional Instances

• we use the sentences produced by exchanging “<Attr1>” and “<Attr2>” (with obvious generalization for n-ary predicates) in the positive sentences.

• If the target predicate is symmetric, like Merger, then such sentences are put into the positive set.

• Otherwise, for anti-symmetric predicates, the sentences are put into the negative set.
Pattern Generation

• The patterns for a predicate $P$ are generalizations of pairs of sentences from the positive set of $P$.
• The function \( \text{Generalize}(S_1, S_2) \) is applied to each pair of sentences $S_1$ and $S_2$ from the positive set of the predicate. The function generates a pattern that is the best (according to the objective function defined below) generalization of its two arguments.
• The following pseudo code shows the process of generating the patterns:

For each predicate $P$

For each pair $S_1, S_2$ from $\text{PositiveSet}(P)$

Let $\text{Pattern} = \text{Generalize}(S_1, S_2)$.

Add $\text{Pattern}$ to $\text{PatternsSet}(P)$. 
The Pattern Language

• The patterns are sequences of tokens, skips, and slots. The tokens can match only themselves, the skips match zero or more arbitrary tokens, and slots match instance attributes.

• Examples of patterns:
  – `<Attr1> * was acquired by <Attr2>`
  – `<Attr1> * merged with * <Attr2>`
  – `<Attr2> is * ceo of * <Attr1>`

• Note, that the sentences from the positive and negative sets of predicates are also patterns, the least general ones since they do not contain skips.
The Generalize Function

- The Generalize($s_1, s_2$) function takes two patterns (e.g., two sentences with slots marked as $<AttrN>$) and generates the least (most specific) common generalization of both.
- The function does a dynamical programming search for the best match between the two patterns.
- The cost of the match is defined as the sum of costs of matches for all elements.
  - two identical elements match at no cost,
  - a token matches a skip or an empty space at cost 2,
  - a skip matches an empty space at cost 1.
  - All other combinations have infinite cost.
- After the best match is found, it is converted into a pattern by copying matched identical elements and adding skips where non-identical elements are matched.
Example

- S1 = “Toward this end, <Arg1> in July acquired <Arg2>”
- S2 = “Earlier this year, <Arg1> acquired <Arg2>”
- After the dynamical programming-based search, the following match will be found:

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Toward</strong></td>
<td><strong>Earlier</strong></td>
<td>(cost 2)</td>
</tr>
<tr>
<td><strong>this</strong></td>
<td><strong>this</strong></td>
<td>(cost 0)</td>
</tr>
<tr>
<td><strong>end</strong></td>
<td><strong>year</strong></td>
<td>(cost 2)</td>
</tr>
<tr>
<td>,</td>
<td>,</td>
<td>(cost 0)</td>
</tr>
<tr>
<td>&lt;Arg1&gt;</td>
<td>&lt;Arg1&gt;</td>
<td>(cost 0)</td>
</tr>
<tr>
<td><strong>in July</strong></td>
<td><strong>acquired</strong></td>
<td>(cost 4)</td>
</tr>
<tr>
<td><strong>acquired</strong></td>
<td><strong>acquired</strong></td>
<td>(cost 0)</td>
</tr>
<tr>
<td>&lt;Arg2&gt;</td>
<td>&lt;Arg2&gt;</td>
<td>(cost 0)</td>
</tr>
</tbody>
</table>
Generating the Pattern

• at total cost = 12. The match will be converted to the pattern
  – * * this * * , <Arg1> * acquired <Arg2>
• which will be normalized (after removing leading and trailing skips, and combining adjacent pairs of skips) into
  – this * , <Arg1> * acquired <Arg2>
Post-processing, filtering, and scoring of patterns

• In the first step of the post-processing we remove from each pattern all function words and punctuation marks that are surrounded by skips on both sides. Thus, the pattern from the example above will be converted to

  , <Arg1> * acquired <Arg2>

• Note, that we do not remove elements that are adjacent to meaningful words or to slots, like the comma in the pattern above, because such anchored elements may be important.
Content Based Filtering

• Every pattern must contain at least one word relevant to its predicate. For each predicate, the list of relevant words is automatically generated from WordNet by following all links to depth at most 2 starting from the predicate keywords. For example, the pattern
  \(<\text{Arg1}> \ast \text{ by } \langle\text{Arg2}\rangle\)
• will be removed, while the pattern
  \(<\text{Arg1}> \ast \text{ purchased } \langle\text{Arg2}\rangle\)
• will be kept, because the word “purchased” can be reached from “acquisition” via synonym and derivation links.
Scoring the Patterns

• The filtered patterns are then scored by their performance on the positive and negative sets.

• We want the scoring formula to reflect the following heuristic: it needs to rise monotonically with the number of positive sentences it matches, but drop very fast with the number of negative sentences it matches.

\[
Score(Pattern) = \frac{|S \in PositiveSet : Pattern \text{ matches } S|}{\left(|S \in NegativeSet : Pattern \text{ matches } S| + 1\right)^2}
\]
Sample Patterns - Inventor

- X, .* inventor .* of Y
- X invented Y
- X, .* invented Y
- when X, .* invented Y
- X's .* invention .* of Y
- inventor .* Y, X
- Y inventor X
- invention .* of Y .* by X
- after X, .* invented Y
- X is .* inventor .* of Y
- inventor .* X, .* of Y
- inventor of Y, .* X,
- X is .* invention of Y
- Y, .* invented .* by X
- Y was invented by X
Sample Patterns – CEO
(Company/X, Person/Y)

- X ceo Y
- X ceo .* Y ,
- former X .* ceo Y
- X ceo .* Y .
- Y , .* ceo of .* X ,
- X chairman .* ceo Y
- Y , X .* ceo
- X ceo .* Y said
- X ' .* ceo Y
- Y , .* chief executive officer .* of X
- said X .* ceo Y
- Y , .* X ' .* ceo
- Y , .* ceo .* X corporation
- Y , .* X ceo
- X ' s .* ceo .* Y ,
- X chief executive officer Y
- Y , ceo .* X ,
- Y is .* chief executive officer .* of X
Shallow Parser mode

• In the first mode of operation (without the use of NER), the predicates may define attributes of two different types: \textit{ProperName} and \textit{CommonNP}.

• We assume that the values of the \textit{ProperName} type are always heads of proper noun phrases. And the values of the \textit{CommonNP} type are simple common noun phrases (with possible proper noun modifiers, e.g. “the Kodak camera”).

• We use a Java-written shallow parser from the OpenNLP (http://opennlp.sourceforge.net/) package. Each sentence is tokenized, tagged with part-of-speech, and tagged with noun phrase boundaries. The pattern matching and extraction is straightforward.
Building a Classification Model

• The goal is to set the score of the extractions using the information on the instance, the extracting patterns and the matches. Assume, that extraction $E$ was generated by pattern $P$ from a match $M$ of the pattern $P$ at a sentence $S$. The following properties are used for scoring:

1. Number of different sentences that produce $E$ (with any pattern).
2. Statistics on the pattern $P$ generated during pattern learning – the number of positive sentences matched and the number of negative sentences matched.
3. Information on whether the slots in the pattern $P$ are anchored.
4. The number of non-stop words the pattern $P$ contains.
5. Information on whether the sentence $S$ contains proper noun phrases between the slots of the match $M$ and outside the match $M$.
6. The number of words between the slots of the match $M$ that were matched to skips of the pattern $P$. 
Building a Classification Model

• During the experiments, it turned out that the pattern statistics (2) produced detrimental results, and the proper noun phrase information (5) did not produce any improvement. The rest of the information was useful, and was turned into the following set of binary features:
  – $f_1(E, P, M, S) = 1$, if the number of sentences producing $E$ is greater than one.
  – $f_2(E, P, M, S) = 1$, if the number of sentences producing $E$ is greater than two.
  – $f_3(E, P, M, S) = 1$, if at least one slot of the pattern $P$ is anchored.
  – $f_4(E, P, M, S) = 1$, if both slots of the pattern $P$ are anchored.
Building a Classification Model

- $f_5$ ... $f_9(E, P, M, S) = 1$, if the number of nonstop words in $P$ is 0, 1 or greater, 2 or greater, ... 4 or greater, respectively
- $f_{10}$ ... $f_{15}(E, P, M, S) = 1$, if the number of words between the slots of the match $M$ that were matched to skips of the pattern $P$ is 0, 1 or less, 2 or less, 3 or less, 5 or less, and 10 or less, respectively.

• As can be seen, the set of features above is rather small, and is not specific to any particular predicate. This allows to train a model using a small amount of labeled data for one predicate, and then to use the model for all other predicates.
Using an NER Component

- In the SRES-NER version the entities of each candidate instance are passed through a simple rule-based NER filter, which attaches a score ("yes", "maybe", or "no") to the argument(s) and optionally fixes the arguments boundaries. The NER is capable of identifying entities of type PERSON and COMPANY (and can be extended to identify additional types).
NER Scores

• The scores mean:
  – “yes” – the argument is of the correct entity type.
  – “no” – the argument is not of the right entity type, and hence the candidate instance should be removed.
  – “maybe” – the argument type is uncertain, can be either correct or no.
Utilizing the NER Scores

• If “no” is returned for one of the arguments, the instance is removed. Otherwise, an additional binary feature is added to the instance's vector:
  – $f_{16} = 1$ iff the score for both arguments is “yes”.

• For bound predicates, only the second argument is analyzed, naturally.
Experimental Evaluation

• We want to answer the following 4 questions:
  1. Can we train SRES’s classifier once, and then use the results on all other relations?
  2. What boost will we get by introducing a simple NER into the classification scheme of SRES?
  3. How does SRES’s performance compare with KnowItAll and KnowItAll-PL?
  4. What is the true recall of SRES?
Training

1. The patterns for a single model predicate are run over a small set of sentences (10000 sentences in our experiment), producing a set of extractions (between 150-300 extractions in our experiments).

2. The extractions are manually labeled according to whether they are correct or no.

3. For each pattern match $M_k$, the value of the feature vector $f_k = (f_1, \ldots, f_{16})$ is calculated, and the label $L_k = \pm 1$ is set according to whether the extraction that the match produced is correct or no.

4. A regression model estimating the function $L(f)$ is built from the training data $\{(f_k, L_k)\}$. We used the BBR, but other models, such as SVM are of course possible.
Testing

1. The patterns for all predicates are run over the sentences.

2. For each pattern match $M$, its score $L(f(M))$ is calculated by the trained regression model. Note that we do not threshold the value of $L$, instead using the raw probability value between zero and one.

3. The final score for each extraction is set to the maximal score of all matches that produced the extraction.
• <e> <arg1>HP</arg1> <arg2>Compaq</arg2>
  – <s><DOCUMENT>Additional information about the <X>HP</X> -<Y>Compaq</Y> merger is available at www.VotetheHPway.com .</DOCUMENT></s>
  – <s><DOCUMENT>The Packard Foundation, which holds around ten per cent of <X>HP</X> stock, has decided to vote against the proposed merger with <Y>Compaq</Y>. </DOCUMENT></s>
  – <s><DOCUMENT>Although the merger of <X>HP</X> and <Y>Compaq</Y> has been approved, there are no indications yet of the plans of HP regarding Digital GlobalSoft.</DOCUMENT></s>
  – <s><DOCUMENT>During the Proxy Working Group's subsequent discussion, the CIO informed the members that he believed that Deutsche Bank was one of <X>HP</X>‘s advisers on the proposed merger with <Y>Compaq</Y>. </DOCUMENT></s>
  – <s><DOCUMENT>It was the first report combining both <X>HP</X> and <Y>Compaq</Y> results since their merger.</DOCUMENT></s>
  – <s><DOCUMENT>As executive vice president, merger integration, Jeff played a key role in integrating the operations, financials and cultures of <X>HP</X> and <Y>Compaq</Y> Computer Corporation following the 19 billion merger of the two companies.</DOCUMENT></s>
Cross-Classification Experiment

![Graphs showing Acquisition and Merger precision with different categories: Acq., CEO, Inventor, Mayor, and Merger.](image-url)
Results!

Acquisition

Merger

Correct Extractions

Precision

Correct Extractions

Precision

KIA  KIA-PL  SRES  S_NER

KIA  KIA-PL  SRES  S_NER
More Results

![Graph for CEOOf](attachment:image1.png)

![Graph for MayorOf](attachment:image2.png)
Inventor Results

![Graph showing precision against correct extractions with lines for KIA, KIA-PL, and SRES]
When is SRES better than KIA?

• KnowItAll extraction works well when redundancy is high and most instances have a good chance of appearing in simple forms that KnowItAll is able to recognize.

• The additional machinery in SRES is necessary when redundancy is low.

• Specifically, SRES is more effective in identifying low-frequency instances, due to its more expressive rule representation, and its classifier that inhibits those rules from overgeneralizing.
The Redundancy of the Various Datasets

![Bar chart showing datasets redundancy]

- Acq
- Merger
- Inventor
- CEO
- Mayor

Average sentences per instance
True Recall Estimates

• It is impossible to manually annotate all of the relation instances because of the huge size of the input corpus.

• Thus, indirect methods must be used. We used a large list of known acquisition and merger instances (that occurred between 1/1/2004 and 31/12/2005) taken from the paid service subscription SBC Platinum.

• For each of the instances in this list we identified all of sentences in the input corpus that contained both instance attributes and assumed that all such sentences are true instances of the corresponding relation.
Under Estimation of the recall

• This is of course an overestimate since in some cases the appearance of both attributes of a true relation instance is just a chance occurrence and does not constitute a true mention of the relation.

• Thus, our estimates of the true recall are pessimistic, and the actual recall is higher.
True Recall Estimates

Acquisition

Merger
Conclusions

• We have presented the SRES system for autonomously learning relations from the Web.

• SRES solves the bottleneck created by classic information extraction systems that either relies on manually developed extraction patterns or on manually tagged training corpus.

• The system relies upon a pattern learning component that enables it to boost the recall of the system.
Future Work

• In our future research we want to try to improve the precision values even at the highest recall levels.

• One of the topics we would like to explore is the complexity of the patterns that we learn. Currently we use a very simple pattern language that just has 3 types of elements, slots, constants and skips. We want to see if we can achieve higher precision with more complex patterns.

• In addition we would like to test SRES on n-ary predicates, and to extend the system to handle predicates that are allowed to lack some of the attributes.

• Another possible research direction is using the Web to validate the extractions interactively.