Information Extraction from the World Wide Web
Information Extraction

Definition:

The automatic extraction of *structured* information from *unstructured* documents.

Different from Information Retrieval (search engines): users are presented with retrieved documents, ranked by relevance. User must open web pages and extract information.

Overall Goals:

– Making information more accessible *to people*
– Making information more *machine-processable*

Practical Goal: Build large knowledge bases
Example: The Problem (seeking jobs)

Martin Baker, a person

Genomics job

Employers job posting form
Example: A Solution
foodscience.com-Job2

JobTitle: Ice Cream Guru
Employer: foodscience.com
JobCategory: Travel/Hospitality
JobFunction: Food Services
JobLocation: Upper Midwest
Contact Phone: 800-488-2611
DateExtracted: January 8, 2001
Source: www.foodscience.com/jobs_midwest.html
OtherCompanyJobs: foodscience.com-Job1
Create a database with the Extracted Job Information

Job Openings:
Category = Food Services
Keyword = Baker
Location = Continental U.S.

<table>
<thead>
<tr>
<th>Job Title</th>
<th>Company / Location</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Food Pantry Workers</td>
<td>Lutheran Social Services</td>
<td>October 11, 2002</td>
</tr>
<tr>
<td>Cooks</td>
<td>Lutheran Social Services</td>
<td>October 11, 2002</td>
</tr>
<tr>
<td>Bakers Assistants</td>
<td>Fine Catering by Russell Morin</td>
<td>October 11, 2002</td>
</tr>
<tr>
<td>Baker's Helper</td>
<td>Bird-in-Hand</td>
<td>October 11, 2002</td>
</tr>
<tr>
<td>Assistant Baker</td>
<td>Gourmet To Go</td>
<td>October 11, 2002</td>
</tr>
<tr>
<td>Host/Hostess</td>
<td>Sharis Restaurants</td>
<td>October 10, 2002</td>
</tr>
<tr>
<td>Cooks</td>
<td>Alta's Rustler Lodge</td>
<td>October 10, 2002</td>
</tr>
<tr>
<td>Line Attendant</td>
<td>Sun Valley Corporation</td>
<td>October 10, 2002</td>
</tr>
<tr>
<td>Food Service Worker II</td>
<td>Garden Grove Unified School District</td>
<td>October 10, 2002</td>
</tr>
<tr>
<td>Night Cooks/Baker</td>
<td>SONOCO</td>
<td>October 10, 2002</td>
</tr>
<tr>
<td>Cooks/Prep Cooks</td>
<td>GrandView Lodge</td>
<td>October 10, 2002</td>
</tr>
<tr>
<td>Line Cook</td>
<td>Lone Mountain Ranch</td>
<td>October 10, 2002</td>
</tr>
<tr>
<td>Production Baker</td>
<td>Whole Foods Market</td>
<td>October 06, 2002</td>
</tr>
<tr>
<td>Cake Decorator/Baker</td>
<td>Mandalay Bay Hotel and Casino</td>
<td>October 09, 2002</td>
</tr>
<tr>
<td>Shift Supervisors</td>
<td>Brueggers Bagels</td>
<td>October 08, 2002</td>
</tr>
</tbody>
</table>
Data Mining the Extracted Job Information

U.S. Job Supply Increases Amid Rising Unemployment

The Job Opportunity Index™ (JOI) increased for the first time in three months in October — climbing 0.7 point to 28.4 and signifying a slight increase in U.S. job supply. However, numerous factors, including a dramatic half-point increase in the national unemployment rate, made October anything but normal.

Special Offer! Find out how you can earn a free subscription to the JOI Report on U.S. Labor Markets through a limited-time JOI Subscriber Referral Program!
Example 2: IE from Research Papers


Peter Norvig Robert Wilensky University of California, Berkeley Computer... Thirteenth International Conference on Computational Linguistics, Volume 3

NEC ResearchIndex

Abstract: this paper critically evaluate three recent abductive interpretation models, those of Charniak and Goldman (1989), Hobbs, Stickle, Martin and Edwards (1988), and Ng and Mooney (1990). These three models add the important property of commensurability: all types of evidence are represented in a common currency that can be compared and combined. While commensurability is a desirable property, and there is a clear need for a way to compare alternate explanations, it appears that a single scalar measure is not enough to account for all types of processing. We present other problems for the abductive approach, and some tentative solutions. (Update)

Context of citations to this paper: More

... (break slight modification of the one given in [Ng and Mooney, 1990]) The new definition remedies the anomaly reported in [Norvig and Wilensky, 1990] of occasionally preferring spurious interpretations of greater depths. Table 1: Empirical Results Comparing Coherence and...

... costs as probabilities, specifically within the context of using abduction for text interpretation, are discussed in Norvig and Wilensky (1990). The use of abduction in disambiguation is discussed in Key et al. 1990) We will assume the following: 13 a. Only literals...

Cited by: More
Translation Mismatch in a Hybrid MT System - Gawron (1999) (Correct)
Abduction and Mismatch in Machine Translation - Gawron (1999) (Correct)
Interpretation as Abduction - Hobbs, Stickle, Apel, Martin (1990) (Correct)

Active bibliography (related documents): More All
0.1: Critiquing Effective Decision Support in Time-Critical Domains - Gertner (1995) (Correct)
0.1: Decision Analytic Networks in Artificial Intelligence - Matzkevich, Abramson (1995) (Correct)
...
This filing covers the period from December 1996 to September 1997.

ENRON GLOBAL POWER & PIPELINES L.L.C.
CONSOLIDATED BALANCE SHEETS
(IN THOUSANDS, EXCEPT SHARE AMOUNTS)

<table>
<thead>
<tr>
<th></th>
<th>SEPTEMBER 30, 1997</th>
<th>DECEMBER 31, 1996</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ASSETS</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Current Assets</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cash and cash equivalents</td>
<td>$ 54,262</td>
<td>$ 24,582</td>
</tr>
<tr>
<td>Accounts receivable</td>
<td>8,473</td>
<td>6,301</td>
</tr>
<tr>
<td>Current portion of notes receivable</td>
<td>1,470</td>
<td>1,394</td>
</tr>
<tr>
<td>Other current assets</td>
<td>336</td>
<td>404</td>
</tr>
<tr>
<td><strong>Total Current Assets</strong></td>
<td>$71,730</td>
<td>$32,681</td>
</tr>
<tr>
<td>Investments in to Unconsolidated Subsidiaries</td>
<td>$286,340</td>
<td>$298,530</td>
</tr>
<tr>
<td>Notes Receivable</td>
<td>16,059</td>
<td>12,111</td>
</tr>
<tr>
<td><strong>Total Assets</strong></td>
<td>$374,408</td>
<td>$343,843</td>
</tr>
</tbody>
</table>

**LIABILITIES AND SHAREHOLDERS' EQUITY**

Current Liabilities
- Accounts payable: $ 13,461 $ 11,277
- Accrued taxes: 1,910 1,488
- **Total Current Liabilities**: 15,371 49,348

Deferred Income Taxes 525 4,301

The U.S. energy markets in 1997 were subject to significant fluctuation.

Data mine these reports for:
- suspicious behavior,
- to better understand what is normal.
What is “Information Extraction”

As a task: Filling slots in a database from sub-segments of text.

October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access."

Richard Stallman, founder of the Free Software Foundation, countered saying…
What is “Information Extraction”

As a task: Filling slots in a database from sub-segments of text.

October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access."

Richard Stallman, founder of the Free Software Foundation, countered saying…

<table>
<thead>
<tr>
<th>NAME</th>
<th>TITLE</th>
<th>ORGANIZATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bill Gates</td>
<td>CEO</td>
<td>Microsoft</td>
</tr>
<tr>
<td>Bill Veghte</td>
<td>VP</td>
<td>Microsoft</td>
</tr>
<tr>
<td>Richard Stallman</td>
<td>founder</td>
<td>Free Soft...</td>
</tr>
</tbody>
</table>
What is “Information Extraction”

As a sequence of techniques: Information Extraction = segmentation + classification + clustering + slot filling

October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access."

Richard Stallman, founder of the Free Software Foundation, countered saying...

Microsoft Corporation
CEO
Bill Gates
Microsoft
Gates
Microsoft
Bill Veghte
Microsoft
VP
Richard Stallman
founder
Free Software Foundation

Identify named entities
What is “Information Extraction”

As a family of techniques:

Information Extraction = segmentation + classification + association + slot filling

October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access."

Richard Stallman, founder of the Free Software Foundation, countered saying…

Microsoft Corporation
CEO
Bill Gates
Microsoft
Gates
Microsoft
Bill Veghte
Microsoft
VP
Richard Stallman
founder
Free Software Foundation

classify entities according to categories, e.g. person, role, company
What is “Information Extraction”

As a family of techniques:

Information Extraction = segmentation + classification + association + slot filling

October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access."

Richard Stallman, founder of the Free Software Foundation, countered saying…

associate related entities

Microsoft Corporation
CEO
Bill Gates

Microsoft
Gates

Microsoft
Bill Veghte
Microsoft
VP

Richard Stallman
founder
Free Software Foundation
What is “Information Extraction”

As a family of techniques:

Information Extraction = segmentation + classification + association + slot filling

October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access."

Richard Stallman, founder of the Free Software Foundation, countered saying…

<table>
<thead>
<tr>
<th>NAME</th>
<th>TITLE</th>
<th>ORGANIZATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bill Gates</td>
<td>CEO</td>
<td>Microsoft</td>
</tr>
<tr>
<td>Bill Veghte</td>
<td>VP</td>
<td>Microsoft</td>
</tr>
<tr>
<td>Richard Stallman</td>
<td>founder</td>
<td>Free Software Foundation</td>
</tr>
</tbody>
</table>
Why IE from the Web?

• Science
  – Grand old dream of AI: Build large KB* and reason with it. IE from the Web enables the creation of this KB.
  – IE from the Web is a complex problem that inspires new advances in machine learning.

• Profit
  – Many companies interested in leveraging data currently “locked in unstructured text on the Web”.
  – Not yet a monopolistic winner in this space.

* KB = “Knowledge Base”
Information Extraction in Applications

- Structured Search
- Opinion Mining/Sentiment Extraction
- Data Mining over Extracted Relationships
What makes IE from the Web Different?
Partly avoid deep NLP analysis, exploit formatting & linking

Apple to Open Its First Retail Store in New York City

MACWORLD EXPO, NEW YORK--July 17, 2002--Apple’s first retail store in New York City will open in Manhattan's SoHo district on Thursday, July 18 at 8:00 a.m. EDT. The SoHo store will be Apple’s largest retail store to date and is a stunning example of Apple's commitment to offering customers the world's best computer shopping experience.

"Fourteen months after opening our first retail store, our 31 stores are attracting over 100,000 visitors each week," said Steve Jobs, Apple's CEO. "We hope our SoHo store will surprise and delight both Mac and PC users who want to see everything the Mac can do to enhance their digital lifestyles."

The directory structure, link structure, formatting & layout of the Web is its own new grammar.
Astro Teller is the CEO and co-founder of BodyMedia. Astro holds a Ph.D. in Artificial Intelligence from Carnegie Mellon University, where he was inducted as a national Hertz fellow. His M.S. in symbolic and heuristic computation and B.S. in computer science are from Stanford University. His work in science, literature and business has appeared in international media from the New York Times to CNN to NPR.

Dr. Steven Minton - Founder/CTO
Dr. Minton is a fellow of the American Association of Artificial Intelligence and was the founder of the Journal of Artificial Intelligence Research. Prior to founding Fetch, Minton was a faculty member at USC and a project leader at USC’s Information Sciences Institute. A graduate of Yale University and Carnegie Mellon University, Minton has been a Principal Investigator at NASA Ames and taught at Stanford, UC Berkeley and USC.

Frank Huybrechts - COO
Mr. Huybrechts has over 20 years of

Non-grammatical snippets, rich formatting & links

Barto, Andrew G.  (413) 545-2109 barbar@cs.umass.edu  CS276
Professor.
Computational neuroscience, reinforcement learning, adaptive motor control, artificial neural networks, adaptive and learning control, motor development.

Berger, Emery D.  (413) 577-4211 emery@cs.umass.edu  CS344
Assistant Professor.

Brock, Oliver  (413) 577-0334 ollie@cs.umass.edu  CS246
Assistant Professor.

Clarke, Lori A.  (413) 545-1328 clarke@cs.umass.edu  CS304
Professor.
Software verification, testing, and analysis; software architecture and design.

Cohen, Paul R.  (413) 545-3638 cohen@cs.umass.edu  CS278
Professor.
Planning, simulation, natural language, agent-based systems, intelligent data analysis, intelligent user interfaces.
Landscape of IE Tasks (2/4):
Wrapper induction

Wrapper extracts content of a particular information source and translates it into a relational form.

Web site specific

Formatting
E.g.: Amazon.com Book Pages

Genre specific

Layout
E.g. Resumes

Wide, non-specific

Language
E.g. University Names

Wrapper induction

Web site specific

E.g.: Amazon.com, Book Pages

Genre specific

E.g.: Resumes

Wide, non-specific

E.g.: University Names

Wrapper extracts content of a particular information source and translates it into a relational form.
E.g. word patterns:

<table>
<thead>
<tr>
<th>Closed set</th>
<th>Regular set</th>
</tr>
</thead>
<tbody>
<tr>
<td>U.S. states</td>
<td>U.S. phone numbers</td>
</tr>
<tr>
<td>He was born in <strong>Alabama</strong>…</td>
<td>Phone: (413) 545-1323</td>
</tr>
<tr>
<td>The big <strong>Wyoming</strong> sky…</td>
<td>The CALD main office can be reached at 412-268-1299</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Complex pattern</th>
<th>Ambiguous patterns, needing context and many sources of evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>U.S. postal addresses</td>
<td>Person names</td>
</tr>
<tr>
<td>University of Arkansas  P.O. Box 140  Hope, AR  71802</td>
<td>…was among the six houses sold by <strong>Hope Feldman</strong> that year.</td>
</tr>
<tr>
<td>Headquarters:  1128 Main Street, 4th Floor  Cincinnati, Ohio 45210</td>
<td>Pawel Lake, Software Engineer at WhizBang Labs.</td>
</tr>
</tbody>
</table>
Landscape of IE Tasks (4/4): Pattern Combinations

<table>
<thead>
<tr>
<th>Single entity</th>
<th>Binary relationship</th>
<th>N-ary record</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Person:</strong> Jack Welch</td>
<td><strong>Relation:</strong> Person-Title</td>
<td><strong>Relation:</strong> Succession</td>
</tr>
<tr>
<td><strong>Person:</strong> Jeffrey Immelt</td>
<td><strong>Person:</strong> Jack Welch</td>
<td><strong>Company:</strong> General Electric</td>
</tr>
<tr>
<td><strong>Title:</strong> CEO</td>
<td><strong>Title:</strong> CEO</td>
<td><strong>Out:</strong> Jack Welch</td>
</tr>
<tr>
<td><strong>Location:</strong> Connecticut</td>
<td><strong>Company:</strong> General Electric</td>
<td><strong>In:</strong> Jeffrey Immelt</td>
</tr>
<tr>
<td><strong>Location:</strong> Connecticut</td>
<td><strong>Location:</strong> Connecticut</td>
<td></td>
</tr>
</tbody>
</table>

“Named entity” extraction

Jack Welch will retire as CEO of General Electric tomorrow. The top role at the Connecticut company will be filled by Jeffrey Immelt.
Evaluation of Single Entity Extraction

TRUTH:
Michael Kearns and Sebastian Seung will start Monday’s tutorial, followed by Richard M. Karpe and Martin Cooke.

PREDICTED:
Michael Kearns and Sebastian Seung will start Monday’s tutorial, followed by Richard M. Karpe and Martin Cooke.

Precision = \( \frac{\text{# correctly predicted segments}}{\text{# predicted segments}} = \frac{2}{6} \)

Recall = \( \frac{\text{# correctly predicted segments}}{\text{# true segments}} = \frac{2}{4} \)

F1 = Harmonic mean of Precision & Recall = \( \frac{1}{\left(\frac{1}{P} + \frac{1}{R}\right) / 2} \)
Parade of IE tasks (and some technique)
Information Extraction (IE) is the process of extracting structured information from unstructured machine-readable documents.

- **Source Selection**: ?
- **Tokenization & Normalization**: 05/01/67 → 1967-05-01
- **Named Entity Recognition**: ...married Elvis on 1967-05-01
- **Instance Extraction**:
  - Elvis Presley: singer
  - Angela Merkel: politician
Source extraction: the web

(1 trillion Web sites)
Finding the Sources

• The document collection can be given a priori (Closed Information Extraction) e.g., a specific document, all files on my computer, ...

• We can aim to extract information from the entire Web (Open Information Extraction)

• For this, we need to crawl the Web. The system can find by itself the source documents e.g., by using an Internet search engine such as Google
Sources: structured

<table>
<thead>
<tr>
<th>Name</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>D. Johnson</td>
<td>30714</td>
</tr>
<tr>
<td>J. Smith</td>
<td>20934</td>
</tr>
<tr>
<td>S. Shenker</td>
<td>20259</td>
</tr>
<tr>
<td>Y. Wang</td>
<td>19471</td>
</tr>
<tr>
<td>J. Lee</td>
<td>18969</td>
</tr>
<tr>
<td>A. Gupta</td>
<td>18884</td>
</tr>
<tr>
<td>R. Rivest</td>
<td>18038</td>
</tr>
</tbody>
</table>

File formats:
• TSV file (values separated by tabulator)
• CSV (values separated by comma)
Sources: semi-structured

File formats:
- XML file (Extensible Markup Language)
- YAML (Yaml Ain’t a Markup Language)
Sources: unstructured

Founded in 1215 as a colony of Genoa, Monaco has been ruled by the House of Grimaldi since 1297, except when under French control from 1789 to 1814. Designated as a protectorate of Sardinia from 1815 until 1860 by the Treaty of Vienna, Monaco's sovereignty ...

File formats:
• HTML file
• text file
• word processing document

<table>
<thead>
<tr>
<th>Event</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Foundation</td>
<td>1215</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Sources: mixed

<table>
<thead>
<tr>
<th>Name</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>Barte</td>
<td>Professor</td>
</tr>
</tbody>
</table>

**Barto, Andrew G.**  (413) 545-2109  barto@cs.umass.edu  CS276

Professor.
Computational neuroscience, reinforcement learning, adaptive motor control, artificial neural networks, adaptive and learning control, motor development.

**Berger, Emery D.**  (413) 577-4211  emery@cs.umass.edu  CS344

Assistant Professor.
Summary of Sources

• We can extract from the entire Web, or from certain Internet domains, thematic domains or files.
• We have to deal with character encodings (ASCII, Code Pages, UTF-8,…) and detect the language.
• Our documents may be structured, semi-structured or unstructured.
**Information Extraction** (IE) is the process of extracting structured information from unstructured machine-readable documents.
Tokenization

- Tokenization is the process of splitting a text into tokens.
- A token is
  - a word
  - a punctuation symbol
  - a url
  - a number
  - a date
  - or any other sequence of characters regarded as a unit

In 2011, President Sarkozy spoke this sample sentence.
Normalization

• Problem: We might extract different literals (numbers, dates, etc.) that mean the same.

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Elvis Presley</td>
<td>1935-01-08</td>
</tr>
<tr>
<td>Elvis Presley</td>
<td>08/01/35</td>
</tr>
</tbody>
</table>

• Solution: Normalize the literals, i.e., convert equivalent literals to one standard form:

- 08/01/35
- 01/08/35
- 8th Jan. 1935
- January 8th, 1935

- 1.67m
- 1.67 meters
- 167 cm
- 6 feet 5 inches
- 3 feet 2 toenails

1935-01-08
1.67m
Normalization

- Conceptually, normalization groups tokens into equivalence classes and chooses one representative for each class.

```
resume
résumé, resume, Resume
1935-01-08
  8th Jan 1935, 01/08/1935
```

Take care not to normalize too aggressively:

```
bush
Bush
```
Information Extraction (IE) is the process of extracting structured information from unstructured machine-readable documents.
Named Entity Recognition (NER)

- Named Entity Recognition (NER) is the process of finding entities (people, cities, organizations, dates, ...) in a text.

Elvis Presley was born in 1935 in East Tupelo, Mississippi.
NER: closed set

- If we have an exhaustive set of the entities we want to extract, we can use closed set extraction:
- Comparing every string in the text to every string in the set.

... in Tupelo, Mississippi, but ...
States of the USA
{ Texas, Mississippi,... }

... while Germany and France were opposed to a 3rd World War, ...
Countries of the World (?)
{ France, Germany, USA,... }

May not always be trivial...
... was a great fan of France Gall, whose songs...
NER: patterns

• If the entities follow a certain pattern, we can use patterns

... was born in **1935**. His mother...
... started playing guitar in **1937**, when...
... had his first concert in **1939**, although...

| Phone numbers | Office: 01 23 45 67 89 | Mobile: 06 19 35 01 08 | Home: 09 77 12 94 65 |

| Years         | (4 digit numbers) | 1935, 1937, 1939 |
Patterns

- A pattern is a string that generalizes a set of strings.

- Sequences of the letter ‘a’:
  - \(a^+\)
  - \(a\ a\ aaaaa\ aaaa\ aaaaaaaaa\)

- ‘a’, followed by ‘b’ s:
  - \(ab^+\)
  - \(abbbbbb\ ab\ abbb\)

- Digits:
  - \(0\ 1\ 2\ 3\ 4\ 5\ 6\ 7\ 8\ 9\)
  - \(0\ 9\ 1\ 6\ 2\ 5\ 7\ 4\ 8\ 3\)

- Sequence of digits:
  - \((0\ 1\ 2\ 3\ 4\ 5\ 6\ 7\ 8\ 9)^+\)
  - \(987\ 6543\ 5321\ 5643\)
Regular Expressions (RegEx)

- A regular expression (regex) over a set of symbols $\Sigma$ is:
  1. the empty string
  2. or the string consisting of an element of $\Sigma$ (a single character)
  3. or the string $AB$ where $A$ and $B$ are regular expressions (concatenation)
  4. or a string of the form $(A|B)$, where $A$ and $B$ are regular expressions (alternation)
  5. or a string of the form $(A)^*$, where $A$ is a regular expression (Kleene star, like $(A)^+$ without the empty string)

- For example, with $\Sigma=$\{a,b\}, the following strings are regular expressions:

\[ a \quad b \quad ab \quad aba \quad (a|b) \]
RegEx (2)

- Matching: a string matches a regex of a single character if the string consists of just that character

  ![Diagram](image1.png)

- A string matches a regular expression of the form \((A)^*\) if it consists of zero or more parts that match A

  ![Diagram](image2.png)
RegEx (3)

• Matching: a string matches a regex of the form (A|B) if it matches either A or B

```
(a | b)  (a | (b)*)
 a     b
```

左右箭头表示匹配的字符串

• a string matches a regular expression of the form AB if it consists of two parts, where the first part matches A and the second part matches B

```
ab  b(a)*
ab  baa baaaaa
```

左右箭头表示匹配的字符串
RegEx (4)

• Given an ordered set of symbols $\Sigma$, we define $[x-y]$ for two symbols $x$ and $y$, $x<y$, to be the alternation $x|...|y$ (meaning: any of the symbols in the range)
  
  $[[0-9]] = 0|1|2|3|4|5|6|7|8|9$

• $A+$ for a regex $A$ to be $A(A)^*$ (meaning: one or more $A$'s)
  
  $[0-9]^+ = [0-9][0-9]^*$

• $A\{x,y\}$ for a regex $A$ and integers $x<y$ to be $A...A|A...A|A...A|...|A...A$ (meaning: $x$ to $y$ $A$'s) $f\{4,6\} = $ ffff | fffff | fffff

• $A?$ for a regex $A$ to be $(|A)$ (meaning: an optional $A$) $ab?? = a(|b)$

• . to be an arbitrary symbol from $\Sigma$ (wild char)
Names & Groups in RegEx

When using regular expressions in a program, it is common to name them:

- String digits="[0-9]+";
- String separator="( |-)";
- String pattern=digits+separator+digits;

- Parts of a regular expression can be singled out by bracketed groups (brackets; "( )" or "//"):  
  - String input="The cat caught the mouse."
  - String pattern="The ([a-z]+) caught the ([a-z]+)\."
A Simple Exercise

• Write a regular expression to find all instances of the determiner “the”:

  /the/
  /[tT]he/
  /\b[tT]he\b/
  /(^|[^a-zA-Z][tT]he[^a-zA-Z])/  

*The recent attempt by the police to retain their current rates of pay has not gathered much favor with the southern factions.*
A Simple Exercise

• Write a regular expression to find all instances of the determiner “the”:

```
/\b\[tT]\he\b/  
/\[^a-zA-Z]\[tT]\he[^a-zA-Z]/
```

*The recent attempt by the police to retain their current rates of pay has not gathered much favor with southern factions.*
A Simple Exercise

• Write a regular expression to find all instances of the determiner “the”:

\/[tT]he/ (also capital T)
\[tT]he\b/
\(^[^a-zA-Z][tT]he[^a-zA-Z]\)/

*The* recent attempt by *the* police to retain *their* current rates of pay has not *gathered* much favor with *the* southern factions.
A Simple Exercise

• Write a regular expression to find all instances of the determiner “the”:

```
/the/
/[tT]he/
/^b[t|T]he\b/ (begin end string)
/([^a-zA-Z][tT][^a-zA-Z]/
```

*The recent attempt by the police to retain their current rates of pay has not gathered much favor with the southern factions.*
A Simple Exercise

• Write a regular expression to find all instances of the determiner “the”:
  /
  /the/
  /\b[\d\D]the\b/
  /((^|[^a-zA-Z])\[\d\D]\bhe[^a-zA-Z]}/
  (nothing or no characters before or after)

The recent attempt by the police to retain their current rates of pay has not gathered much favor with the southern factions.
More high-level examples

• Create rules to extract locations
  – Capitalized word + \{city, center, river\} indicates location
    Ex. New York city
    Hudson river
  – Capitalized word + \{street, boulevard, avenue\} indicates location
    Ex. Fifth avenue

<table>
<thead>
<tr>
<th>Metacharacters</th>
<th>Repetition</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>char</strong></td>
<td><strong>a</strong></td>
</tr>
<tr>
<td></td>
<td><strong>a</strong></td>
</tr>
<tr>
<td>^</td>
<td>*</td>
</tr>
<tr>
<td>$</td>
<td>+</td>
</tr>
<tr>
<td>.</td>
<td>?</td>
</tr>
<tr>
<td>*</td>
<td>{m}</td>
</tr>
<tr>
<td>+</td>
<td>{m,}</td>
</tr>
<tr>
<td>?</td>
<td>{m,n}</td>
</tr>
<tr>
<td></td>
<td>repetition?</td>
</tr>
<tr>
<td>alternative</td>
<td>zero or more <strong>a</strong>’s</td>
</tr>
<tr>
<td>grouping; “storing”</td>
<td>one or more <strong>a</strong>’s</td>
</tr>
<tr>
<td>set of characters</td>
<td>zero or one <strong>a</strong>’s (i.e., optional <strong>a</strong>)</td>
</tr>
<tr>
<td>repetition modifier</td>
<td>exactly <strong>m</strong> <strong>a</strong>’s</td>
</tr>
<tr>
<td>quote or special</td>
<td>at least <strong>m</strong> <strong>a</strong>’s</td>
</tr>
<tr>
<td></td>
<td>at least <strong>m</strong> but at most <strong>n</strong> <strong>a</strong>’s</td>
</tr>
<tr>
<td></td>
<td>same as repetition but the shortest match is taken</td>
</tr>
</tbody>
</table>
## Perl regex

### Special notations with \\`

<table>
<thead>
<tr>
<th>Single characters</th>
<th>“Zero-width assertions”</th>
</tr>
</thead>
<tbody>
<tr>
<td>\t</td>
<td>\b “word” boundary</td>
</tr>
<tr>
<td>\n</td>
<td>\B not a “word” boundary</td>
</tr>
<tr>
<td>\r</td>
<td></td>
</tr>
<tr>
<td>\xhh</td>
<td></td>
</tr>
</tbody>
</table>

### Matching

<table>
<thead>
<tr>
<th>Character</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>\w</td>
<td>matches any <em>single</em> character classified as a “word” character (alphanumeric or “_”)</td>
</tr>
<tr>
<td>\W</td>
<td>matches any non-“word” character</td>
</tr>
<tr>
<td>\s</td>
<td>matches any whitespace character (space, tab, newline)</td>
</tr>
<tr>
<td>\S</td>
<td>matches any non-whitespace character</td>
</tr>
<tr>
<td>\d</td>
<td>matches any digit character, equiv. to [0-9]</td>
</tr>
<tr>
<td>\D</td>
<td>matches any non-digit character</td>
</tr>
</tbody>
</table>
Perl regex

Character sets: specialities inside [...]

**Different meanings** apply inside a character set ("character class") denoted by [... so that, *instead* of the normal rules given here, the following apply:

<table>
<thead>
<tr>
<th><strong>[characters]</strong></th>
<th>matches any of the characters in the sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>[x-y]</strong></td>
<td>matches any of the characters from <em>x</em> to <em>y</em> (inclusively) in the ASCII code</td>
</tr>
<tr>
<td><strong>[--]</strong></td>
<td>matches the hyphen character “–”</td>
</tr>
<tr>
<td><strong>\n</strong></td>
<td>matches the newline; other <em>single character denotations with * apply normally, too</em></td>
</tr>
<tr>
<td><strong>[^something]</strong></td>
<td>matches any character <em>except</em> those that [something] denotes; that is, immediately after the leading “[”, the circumflex “^” means “not” applied to all of the rest</td>
</tr>
<tr>
<td>expression</td>
<td>matches...</td>
</tr>
<tr>
<td>------------</td>
<td>------------</td>
</tr>
<tr>
<td>abc</td>
<td>abc (that exact character sequence, but anywhere in the string)</td>
</tr>
<tr>
<td>^abc</td>
<td>abc at the <em>beginning</em> of the string</td>
</tr>
<tr>
<td>abc$</td>
<td>abc at the <em>end</em> of the string</td>
</tr>
<tr>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>^abc</td>
<td>abc$</td>
</tr>
<tr>
<td>ab{2,4}c</td>
<td>an a followed by two, three or four b’s followed by a c</td>
</tr>
<tr>
<td>ab{2,}c</td>
<td>an a followed by at least two b’s followed by a c</td>
</tr>
<tr>
<td>ab*c</td>
<td>an a followed by any number (zero or more) of b’s followed by a c</td>
</tr>
<tr>
<td>ab+c</td>
<td>an a followed by one or more b’s followed by a c</td>
</tr>
<tr>
<td>ab?c</td>
<td>an a followed by an optional b followed by a c; that is, either abc or ac</td>
</tr>
<tr>
<td>a.c</td>
<td>an a followed by any single character (not newline) followed by a c</td>
</tr>
<tr>
<td>a\c</td>
<td>a.c exactly</td>
</tr>
<tr>
<td>[abc]</td>
<td>any one of a, b and c</td>
</tr>
<tr>
<td>[Aa]bc</td>
<td>either of Abc and abc</td>
</tr>
<tr>
<td>[abc]+</td>
<td>any (nonempty) string of a’s, b’s and c’s (such as a, abba, acbabcaca)</td>
</tr>
<tr>
<td>[^abc]+</td>
<td>any (nonempty) string which does <em>not</em> contain any of a, b and c (such as defg)</td>
</tr>
<tr>
<td>\d\d</td>
<td>any two decimal digits, such as 42; same as \d{2}</td>
</tr>
<tr>
<td>\w+</td>
<td>a “<em>word</em>”: a nonempty sequence of alphanumeric characters and low lines (underscores), such as foo and 12bar8 and foo_1</td>
</tr>
<tr>
<td>100\s*mk</td>
<td>the strings 100 and mk optionally separated by any amount of white space (spaces, tabs, newlines)</td>
</tr>
<tr>
<td>abc\b</td>
<td>abc when followed by a word boundary (e.g. in abc! but not in abcd)</td>
</tr>
</tbody>
</table>
NER RegEx examples (in Pearl)

- Software name extraction: “*one or more capitalized words followed by a version number*” (Mac OS X v.10.6.8)

```
([A-Z]\w*\s*)+[Vv]?(\d+\.?)+.  
```

- one or more capitalized words followed by space
- followed by (0 or 1) instances of V or v
- followed by one or more digits, one or zero “.” followed by anything else
Create regular expressions to extract:

Telephone number

blocks of digits separated by hyphens

RegEx = (\d+\-\d+)\d+

• matches valid phone numbers like 900-865-1125 and 725-1234
• incorrectly extracts social security numbers 123-45-6789
• fails to identify numbers like 800.865.1125 and (800)865-CARE

Improved RegEx = (\d{3}[\-\.)]{1,2}\d{4}[\dA-Z]{4}
Another example

John James, Jr. Smith
John James Smith, Jr.
John James Smith Jr.
John, Jr. Smith
John Smith, Jr.
John Smith Jr.

(<first name regexp>)\s(<optional middle regexp>)\s(<optional Jr.|Sr.|II|III|IV>)\s(<last name regexp>)\s(<optional Jr.|Sr.|II|III|IV>)
Matching RegEx

- A regex can be matched efficiently by a Finite State Machine (Finite State Automaton, FSA, FSM).

Regex: $ab^*c$

Implicitly: All unmentioned inputs go to some artificial failure state.

Accepting states usually depicted with double ring.
RegEx summary

• Regular expressions
  – can express a wide range of patterns
  – can be matched efficiently
  – are employed in a wide variety of applications (e.g., in text editors, NER systems, normalization, UNIX grep tool etc.)

Input:
• Manual design of the regex

Condition:
• Entities follow a pattern
Sliding Windows

• What if we do not want to specify regexes by hand? Use sliding windows:
  • Sliding windows method is based on ML learning algorithms and annotated datasets (sentences annotated with named entities of selected types)

*Information Extraction: Tuesday 10:00 am, Rm 407b*
Sliding Windows

Choose certain features (properties) of windows that could be important:

- window contains colon, comma, or digits
- window contains week day, or certain other words
- window starts with lowercase letter
- window contains only lowercase letters
- ...
Feature Vectors

Information Extraction: Tuesday 10:00 am, Rm 407b

Prefix window  Content window  Postfix window

Prefix colon  
Prefix comma  
...
Content colon  
Content comma  
...
Postfix colon  
Postfix comma

Features  Feature Vector

The **feature vector** represents the presence or absence of features of one content window (and its prefix window and postfix window).
Now, we need a corpus (set of documents) in which the entities of interest have been manually labeled.

From this corpus, compute the feature vectors with labels:

- **NLP class**: Wednesday, \textcolor{red}{7:30am} and Thursday all day, \textcolor{red}{rm 667}

\[\begin{array}{c}
1 \\
0 \\
0 \\
0 \\
1 \\
\end{array} \quad \begin{array}{c}
1 \\
1 \\
0 \\
0 \\
0 \\
\end{array} \quad \begin{array}{c}
1 \\
0 \\
1 \\
1 \\
1 \\
\end{array} \quad \begin{array}{c}
1 \\
0 \\
0 \\
0 \\
1 \\
\end{array} \quad \begin{array}{c}
1 \\
0 \\
1 \\
0 \\
1 \\
\end{array}
\]

- **Nothina**  **Nothina**  **Time**  **Nothing**  **Location**
Machine Learning

Information Extraction: Tuesday 10:00 am, Rm 407b

Use the labeled feature vectors as training data for Machine Learning

\[
\begin{bmatrix}
1 \\
0 \\
0 \\
1 \\
1 \\
1
\end{bmatrix}
\]

classify

\[
\begin{bmatrix}
1 \\
0 \\
0 \\
1 \\
1 \\
0
\end{bmatrix}
\]

Result

Time

Nothing Location
Sliding Windows Exercise

- What features would you use to recognize person names?

```
Elvis Presley married Ms. Priscilla at the Aladdin Hotel.
```

- **UpperCase**
- **hasDigit**
- **...**

```
UpperCase  hasDigit  First term is a person title (Mr. Dr...)
0          0          0
0          0          1
0          1          1
1          1          1
```

```
1          0          1
1          1          0
1          0          1
```

```
NER summary (but we learned general techniques..)

• Named Entity Recognition (NER) is the process of finding entities (people, cities, organizations, ...) in a text.
  – We have seen different techniques
  – Closed-set extraction (if the set of entities is known)
  – Extraction with Regular Expressions (if the entities follow a pattern). Can be done efficiently with Finite State Automata
  – Extraction with sliding windows / Machine Learning (if the entities share some syntactic features)
Information Extraction (IE) is the process of extracting structured information from unstructured machine-readable documents.

- Source Selection
- Tokenization & Normalization
  - 05/01/67
  - → 1967-05-01
- Named Entity Recognition
  - ...married Elvis on 1967-05-01
- Instance Extraction
- Fact Extraction
- Ontological Information Extraction

Elvis Presley: singer
Angela Merkel: politician
Instance/relation Extraction

• Instance Extraction is the process of extracting entities with their class (i.e., concept, set of similar entities)

Elvis was a great artist, but while all of Elvis’ colleagues loved the song “Oh yeah, honey”, Elvis did not perform that song at his concert in Hintertuepflingen.

<table>
<thead>
<tr>
<th>Entity</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elvis</td>
<td>artist</td>
</tr>
<tr>
<td>Oh yeah, honey</td>
<td>song</td>
</tr>
<tr>
<td>Hintertuepflingen</td>
<td>location</td>
</tr>
</tbody>
</table>
Instance/relation extraction with patterns

- Sentences express class membership in very predictable patterns. Use these patterns for instance extraction.

**Pattern:** X was a great Y

---

**Example:**

Elvis was a great artist, but while all of Elvis’ colleagues loved the song “Oh yeah, honey”, Elvis did not perform that song at his concert in Hintertüpfelingen.

<table>
<thead>
<tr>
<th>Entity</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elvis</td>
<td>artist</td>
</tr>
</tbody>
</table>
Instance/relation extraction using patterns

Elvis was a great artist

Many scientists, including Einstein, started to believe that matter and energy could be equated.

He adored Madonna, Celine Dion and other singers, but never got an autograph from any of them.

Many US citizens have never heard of countries such as Guinea, Belize or France.

- X was a great Y
- Ys, such as X1, X2, ...
- X1, X2, … and other Y
- many Ys, including X

Can write RegEx for each pattern
Instance/relation extraction using patterns

• Manually writing patterns is difficult
• Lexical patterns do not generalize, e.g.
  – Elvis was a great pianist
  – Elvis was a pianist
  – Elvis, the pianist..
• Learn patterns with ML techniques
• Generalise patterns using HMM or lattices
Learning hypernym relations from definitions
Information Extraction (IE) is the process of extracting structured information from unstructured machine-readable documents.
Fact extraction

- Fact Extraction is the process of extracting pairs (triples,...) of entities together with the relationship of the entities.
Fact Extraction

• Approaches:
  • Fact extraction from tables
    (if the corpus contains lots of tables)
  • Wrapper induction
    (for extraction from one Internet domain)
  • Pattern matching
    (for extraction from natural language documents)
  • ... and many others...
ReVerb (http://openie.cs.washington.edu/)

Pattern Based approach: read “Identifying Relations for Open Information Extraction” EMNLP 2011
Freely downloadable and available on-line

Example Queries:
- What kills bacteria?
- Who built the Pyramids?
- What did Thomas Edison invent?
- What contains antioxidants?

Typed Example Queries:
- What countries are located in Africa?
- What actors starred in which films?
- What is the symbol of which country?
- What foods are grown in which countries?
- What drug ingredients has the FDA approved?
nails

Extracted Synonyms:

a nail

eXtracted from these sentences:

A hammer is used for nails, and a screwdriver is used for screws, so each type of geometric structure requires its own geometry. (via Google)

Round inches can generally be substituted for square inches in geometry when calculating the area of circles. Just as circular mils are used to measure the area of wire, so round inches can be used to measure the areas of circles. Of course, just as there is no exact way to measure circular areas in terms of straight lines, there is no way to exactly measure the area of squares and rectangles using circular geometry. Just as a hammer is used for nails, and a screwdriver is used for screws, so each type of geometric structure requires its own geometry. (via ClueWeb09)

The bottom line: just as a hammer is used for nails and a screwdriver for screws, assembly and compiled code each have a place in embedded applications. (via Google)

It is just choosing the right tool for the job, and if and do-its are like a hammer and screwdriver a hammer is used for a nail and a screwdriver for a screw. (via ClueWeb09)
Information Extraction (IE) is the process of extracting structured information from unstructured machine-readable documents.

- Instance Extraction
- Named Entity Recognition
- Tokenization & Normalization
- Source Selection
- Fact Extraction
- Ontological Information Extraction

<table>
<thead>
<tr>
<th>Person</th>
<th>Nationality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Angela Merkel</td>
<td>German</td>
</tr>
</tbody>
</table>
Ontology Extraction

- An ontology is consistent knowledge base without redundancy
- Every entity appears only with exactly the same name
- There are no semantic contradictions

<table>
<thead>
<tr>
<th>Person</th>
<th>Nationality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Angela Merkel</td>
<td>German</td>
</tr>
<tr>
<td>Merkel</td>
<td>Germany</td>
</tr>
<tr>
<td>A. Merkel</td>
<td>French</td>
</tr>
</tbody>
</table>

| Entity        | Relation     | Entity        |
|---------------|--------------|
| Angela Merkel | citizenOf    | Germany       |
Ontology Extraction

- Ontological Information Extraction (IE) aims to create or extend an ontology.

<table>
<thead>
<tr>
<th>Entity</th>
<th>Relation</th>
<th>Entity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Angela Merkel</td>
<td>citizenOf</td>
<td>Germany</td>
</tr>
</tbody>
</table>

Angela Merkel is the German chancellor.... Merkel was born in Germany...

...A. Merkel has French nationality...

<table>
<thead>
<tr>
<th>Person</th>
<th>Nationality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Angela Merkel</td>
<td>German</td>
</tr>
<tr>
<td>Merkel</td>
<td>Germany</td>
</tr>
<tr>
<td>A. Merkel</td>
<td>French</td>
</tr>
</tbody>
</table>
Ontological IE Challenges

• Challenge 1

Map names to names that are already known

<table>
<thead>
<tr>
<th>Entity</th>
<th>Relation</th>
<th>Entity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Angela Merkel</td>
<td>citizenOf</td>
<td>Germany</td>
</tr>
</tbody>
</table>

Merkel
Angie
A. Merkel
Ontological IE Challenges

• Challenge 2

Be sure to map the names to the right known names

<table>
<thead>
<tr>
<th>Entity</th>
<th>Relation</th>
<th>Entity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Angela Merkel</td>
<td>citizenOf</td>
<td>Germany</td>
</tr>
<tr>
<td>Una Merkel</td>
<td>citizenOf</td>
<td>USA</td>
</tr>
</tbody>
</table>

Merkel is great!
Ontological IE Challenges

- Challenge 3

Map to known relationships

<table>
<thead>
<tr>
<th>Entity</th>
<th>Relation</th>
<th>Entity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Angela Merkel</td>
<td>citizenOf</td>
<td>Germany</td>
</tr>
</tbody>
</table>

... has nationality ...
... has citizenship ...
... is citizen of ...
Ontological IE Challenges

- Challenge 4
- Find hypernymy relations

Chancellor of Germany
Ontolearn (IJCAI 2011, CL 2013)

Inducing lexical taxonomies from scratch
Taxonomy Learning Workflow

Web glossaries & documents

Terminology extraction

Domain terms

Definition & hypernym extraction

Upper terms

Domain filtering

Graph pruning

Induced taxonomy

Hypernym graph

Domain corpus
Taxonomy Learning Workflow

- Domain terms
  - Terminology extraction
  - Definition & hypernym extraction
  - Domain filtering

- Upper terms

Web glossaries & documents

- Hypernym graph

Induced taxonomy
Terminology Extraction

Domain Corpus

Domain terms

flow network
mesh generation
hash function
maximum likelihood
pattern recognition
information processing

several on-line terminology extractors
(e.g. TermExtractor http://hal.di.uniroma1.it/termextractor/public/demo.faces)
Taxonomy Learning Workflow

- Domain terms
- Upper terms
- Web glossaries & documents
- Terminology extraction
- Definition & hypernym extraction
- Domain filtering
- Graph pruning
- Hypernym graph
- Induced taxonomy
In graph theory, a flow network is a directed graph.

Global Cash Flow Network is a business opportunity to make money online.

A flow network is a network with two distinguished vertices.
In graph theory, a flow network is a directed graph.

A flow network is a network with two distinguished vertices.

**Definition & Hypernym Extraction**

+ Domain Filtering

### Domain Terms

- **Flow network**
- **Directed graph**
- **Network**

### Domain Corpus & Web Glossaries & Documents

- **Definition extraction**
  - Flow network is a network with two distinguished vertices.

- **Hypernym extraction**
  - Directed graph
Definition & Hypernym Extraction + Domain Filtering

A directed graph is a graph where ...
A directed graph is a data structure ...

hyponym extraction

graph

flow network
data structure

directed graph

network
data structure
Hypernym Extraction Algorithm

Based on **Word-Class Lattices**, i.e. lattice-based definition models learned by means of a greedy definition alignment algorithm

- Determine whether a sentence is **definitional**
- If so, returns the **hypernym(s)** of the defined term
Performance in Definition Extraction

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>P</th>
<th>R</th>
<th>F₁</th>
<th>A</th>
</tr>
</thead>
<tbody>
<tr>
<td>WCL-1</td>
<td>99.88</td>
<td>42.09</td>
<td>59.22</td>
<td>76.06</td>
</tr>
<tr>
<td>WCL-3</td>
<td>98.81</td>
<td>60.74</td>
<td>75.23</td>
<td>83.48</td>
</tr>
<tr>
<td>Star patterns</td>
<td>86.74</td>
<td>66.14</td>
<td>75.05</td>
<td>81.84</td>
</tr>
<tr>
<td>Bigrams</td>
<td>66.70</td>
<td>82.70</td>
<td>73.84</td>
<td>75.80</td>
</tr>
<tr>
<td>Random BL</td>
<td>50.00</td>
<td>50.00</td>
<td>50.00</td>
<td>50.00</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>P</th>
<th>R⁺</th>
</tr>
</thead>
<tbody>
<tr>
<td>WCL-1</td>
<td>98.33</td>
<td>39.39</td>
</tr>
<tr>
<td>WCL-3</td>
<td>94.87</td>
<td>56.57</td>
</tr>
<tr>
<td>Star patterns</td>
<td>44.01</td>
<td>63.63</td>
</tr>
<tr>
<td>Bigrams</td>
<td>46.60</td>
<td>45.45</td>
</tr>
<tr>
<td>Random BL</td>
<td>50.00</td>
<td>50.00</td>
</tr>
</tbody>
</table>

Wikipedia

UKWac corpus

Outperforms existing methods for definition extraction
Precision in Hypernym Extraction

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Full</th>
<th>Substring</th>
</tr>
</thead>
<tbody>
<tr>
<td>WCL-1</td>
<td>42.75</td>
<td>77.00</td>
</tr>
<tr>
<td>WCL-3</td>
<td>40.73</td>
<td>78.58</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Full</th>
<th>Substring</th>
</tr>
</thead>
<tbody>
<tr>
<td>WCL-1</td>
<td>86.19 (206)</td>
<td>96.23 (230)</td>
</tr>
<tr>
<td>WCL-3</td>
<td>89.27 (383)</td>
<td>96.27 (413)</td>
</tr>
<tr>
<td>Hearst</td>
<td>65.26 (62)</td>
<td>88.42 (84)</td>
</tr>
</tbody>
</table>

Pattern-based methods achieve much lower recall: 62 vs. 383 hypernyms extracted from UKWac
The iterative growth of the hypernym graph
Iteration I
Taxonomy Learning Workflow

- **Domain terms**
- **Definition & hypernym extraction**
- **Domain filtering**
- **Graph pruning**

Input:
- Domain corpus
- Web glossaries & documents

Output:
- Induced taxonomy
- Hypernym graph
Graph Pruning

• **Goal:**
  – From *noisy* hypernym graph
  – To *full-fledged* taxonomy

• **How:**
  – A *graph-based* algorithm
Graph Pruning: Preliminary step

Given the hypernym graph, we disconnect:

• false roots (root nodes not in the set of upper terms)

• false leaves (leaf nodes not in the initial terminology)
Graph Pruning: Preliminary step

Given the hypernym graph, we disconnect:

• **false roots** (root nodes not in the set of upper terms)

• **false leaves** (leaf nodes not in the initial terminology)
Graph Pruning (1)

- Weight each node $v$ by the number of terminological nodes reachable from $v$

  - E.g.:
    - $w(\text{collection}) = 2$
    - $w(\text{graph}) = 3$
For each path $p$ from an upper term $r$ to a node $v$, we calculate its cumulative weight:

$$\omega(p) = \sum_{v' \in p} w(v')$$

E.g.:

- $\omega(p) = 5+3 = 8$
- $\omega(p') = 5+3+2 = 10$
Graph Pruning (3)

- Assign to each edge \((h, v)\) the maximum cumulative weight among all the paths from any upper term to node \(v\)

\[
w(h, v) = \max_{r \in U} \max_{p \in \Gamma(r, h)} \omega(p)
\]

- E.g.:
  - \(w(\text{graph}, \text{binary tree}) = 8\)
  - \(w(\text{tree}, \text{binary tree}) = 10\)
Graph Pruning (4)

- Apply the Chu-Liu Edmonds [1967] algorithm
Graph Pruning (4)

- Apply the Chu-Liu Edmonds [1967] algorithm
- As a result we obtain a tree-like taxonomy
From the Noisy Hypernym Graph...
...to a Tree-like Taxonomy
Pruning Recovery

• Many small connected components will be returned

• To recover from excessive pruning we apply a simple heuristic:
  – Let \( r \) be the root of such a component
  – Select the best-ranking edge \((v, r)\) according to the domain score of the corresponding definition
  – If no edge exists, we use string inclusion
    • E.g., \( r=\text{binary tree} \) and \( v=\text{tree} \)
Evaluation

• Taxonomy evaluation is a hard task
• We performed two different experiments:
  – Inducing an Artificial Intelligence (AI) taxonomy
  – Reproducing existing sub-hierarchies in WordNet
Experiment 1: Inducing an AI Taxonomy

- **Corpus**: IJCAI 2009 proceedings (334 papers)
- **Domain terminology**: via term extraction
  - 374 initial domain terms
- **Upper terms**: we manually selected 13 terms (process, abstraction, algorithm)
  - Used as a stopping criterion for definition/hypernym extraction
- **Definition and hypernym extraction**: IJCAI corpus + Google define
  - 715 nodes and 1025 edges
Experiment 1: Inducing an AI Taxonomy

- Final graph: 427 nodes, 426 edges
- Compression ratio (against unpruned graph): 0.60 (nodes), 0.41 (edges)
Experiment 1: Inducing an AI Taxonomy

- Final graph: 427 nodes, 426 edges
- Compression ratio (against unpruned graph): 0.60 (nodes), 0.41 (edges)
- Manual evaluation of edge precision: 81.5\% (347/426)
  - Note: many hypernyms would be equally valid (collaborative assessment?)
- The AI Taxonomy is available to the community:
  - http://lcl.uniroma1.it/taxolearn
Experiment 2: Evaluation against WordNet

- **Same evaluation strategy** as in Kozareva & Hovy (EMNLP 2010)
  - **Domains**: animals, plants, vehicles
- **No terminology extraction**: we use the terminology provided by K&H for each domain
- **Upper terms**: those in the synsets of animal#n#1, plant#n#2, vehicle#n#1
- **Definition and hypernym extraction**: no domain corpus, just Google define
## Experiment 2: Evaluation against WordNet

- **Statistics and manual evaluation:**

<table>
<thead>
<tr>
<th></th>
<th>animals</th>
<th>plants</th>
<th>vehicles</th>
</tr>
</thead>
<tbody>
<tr>
<td>node compression ratio</td>
<td>0.48</td>
<td>0.40</td>
<td>0.41</td>
</tr>
<tr>
<td>edge compression ratio</td>
<td>0.49</td>
<td>0.35</td>
<td>0.35</td>
</tr>
<tr>
<td>coverage of initial terminology</td>
<td>0.71</td>
<td>0.79</td>
<td>0.73</td>
</tr>
<tr>
<td>precision by hand of edges not in WN (100 randomly chosen)</td>
<td>0.76</td>
<td>0.82</td>
<td>0.92</td>
</tr>
<tr>
<td>precision by hand of nodes not in WN (all)</td>
<td>0.70</td>
<td>0.77</td>
<td>0.69</td>
</tr>
</tbody>
</table>

- Example values for each category:
  - Animals: node compression ratio = 0.48, edge compression ratio = 0.49, coverage of initial terminology = 0.71, precision by hand of edges not in WN (100 randomly chosen) = 0.76, precision by hand of nodes not in WN (all) = 0.70
  - Plants: node compression ratio = 0.40, edge compression ratio = 0.35, coverage of initial terminology = 0.79, precision by hand of edges not in WN (100 randomly chosen) = 0.82, precision by hand of nodes not in WN (all) = 0.77
  - Vehicles: node compression ratio = 0.41, edge compression ratio = 0.35, coverage of initial terminology = 0.73, precision by hand of edges not in WN (100 randomly chosen) = 0.92, precision by hand of nodes not in WN (all) = 0.69
Conclusions

• An algorithm to learn a lexical taxonomy truly from scratch
  – Based on the idea of exploiting the scholarly knowledge from definitions
  – A graph-based approach to learn a “clean” taxonomy
  – Can be applied to any domain of interest with little effort
References

References


