Information extraction from texts

Technical and business challenges
Overview

- Mentis
- Text mining field overview
- Application: Information Extraction
  - Motivation & Overview
Consulting and Software company Created in 2005.

Spin-Off of IRIDIA (artificial Intelligence laboratory at ULB)

Strong partnership with the University

Multi-disciplinary team: Math, physics, IT, Optimization

- 3 PHD
- 7 engineers
Mentis – Some references
Offer High value solutions to help understand, structure, find, and enrich unstructured & Structured Textual information and transform it into knowledge.

Understand
- Deep analysis based on domain specific knowledge bases

Structure
- Use domain specific taxonomies to normalize and structure

Find
- Better Information retrieval through semantic search engines.

Enrich
- Assign entities and facts to words and terms.
Mentis – Core business

- Technology provider
  - Text Analysis Engine
- Outsourcing
  - Expert shopping
  - Process outsourcing
- Solution provider
  - Innovation business (solve the insolvable)
Research communities

Approaches

Methodologies
Levels of Representation

Semantic

- Semantic Models
- Ontologies
- Taxonomies
- Collaborative Tagging

Syntactic

- Cross-modality
- Full Parsing
- Language Models
- Vector space Model

Lexical

- Part Of speech Tags
- Phrases
- Words
- Character
Levels of Representation

Semantic
- Semantic Models
- Ontologies
- Taxonomies
- Collaborative Tagging
- Cross-modality
- Full Parsing
- Language Models
- Vector space Model
- Part Of speech Tags
- Phrases
- Words
- Character

Domain Knowledge - Reasoning
- Unifying Semantics and Data: Flicker, …
- Multi-Lingual Search – Text/Image/Audio
- Machine Translation – Spam Detection
- Classification – Clustering - search
- Named Entity Extraction
- Language identification – Copy detection

Lexical

Syntactic
Levels of Representation

- Semantic Models
  - Ontologies
  - Taxonomies
  - Collaborative Tagging

- Cross-modality
- Full Parsing
- Language Models
- Vector space Model

- Part Of speech Tags
- Phrases
- Words
- Character

- Machine Learning
- IR
- CL
- Semantic Web
Store Organize and Retrieve knowledge
Digital Age

- World Wide Web
- Digital Libraries
- Special purpose content providers (Lexis Nexis)
- Company Intranet and digital assets
- Scientific Literature (siteseer, Medline)
- Patent Database
Types of documents

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

Abstract: this paper we critically evaluate three recent abductive interpretation models, those of Charniak and Goldman (1989); Hobbs, Stickel, 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 Kay 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, Stickel, Appelt, 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)
0.1: A Probabilistic Network of Predicates - Delpezo Lin (1992) (Correct)
Information Extraction
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...
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...
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

aka “named entity extraction”
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…
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”

Information Extraction = segmentation + classification + association + clustering

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...
Analyze

Scanning

Physical Document

Mark-Up and Extraction

Database

Part | Problem | Condition
--- | --- | ---
Fuel Pump | Fails | corroded
Pump Relay | Shorts | Cold weather
Headlight | Fails | Running hot
Engine | Stalls | At low speeds
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.
### 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>
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!
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.
Good Features for IE

Creativity and Domain Knowledge Required!

begins-with-number
begins-with-ordinal
begins-with-punctuation
begins-with-question-word
begins-with-subject
blank
contains-alphanum
contains-bracketed-number
contains-http
contains-non-space
contains-number
contains-pipe

Example word features:

- identity of word
- is in all caps
- ends in “-ski”
- is part of a noun phrase
- is in a list of city names
- is under node X in WordNet or Cyc
- is in bold font
- is in hyperlink anchor
- features of past & future
- last person name was female
- next two words are “and Associates”

contains-question-mark
contains-question-word
ends-with-question-mark
first-alpha-is-capitalized
indented
indented-1-to-4
indented-5-to-10
more-than-one-third-space
only-punctuation
prev-is-blank
prev-begins-with-ordinal
shorter-than-30
### Good Features for IE

**Creativity and Domain Knowledge Required!**

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Is Capitalized</td>
<td>Character n-gram classifier says string is a person name (80% accurate)</td>
</tr>
<tr>
<td>Is Mixed Caps</td>
<td>In stopword list (the, of, their, etc)</td>
</tr>
<tr>
<td>Is All Caps</td>
<td>In honorific list (Mr, Mrs, Dr, Sen, etc)</td>
</tr>
<tr>
<td>Initial Cap</td>
<td>In person suffix list (Jr, Sr, PhD, etc)</td>
</tr>
<tr>
<td>Contains Digit</td>
<td>In name particle list (de, la, van, der, etc)</td>
</tr>
<tr>
<td>All lowercase</td>
<td>In Census lastname list; segmented by P(name)</td>
</tr>
<tr>
<td>Is Initial</td>
<td>In Census firstname list; segmented by P(name)</td>
</tr>
<tr>
<td>Punctuation</td>
<td>In locations lists (states, cities, countries)</td>
</tr>
<tr>
<td>Period</td>
<td>In company name list (“J. C. Penny”)</td>
</tr>
<tr>
<td>Comma</td>
<td>In list of company suffixes (Inc, &amp; Associates, Foundation)</td>
</tr>
</tbody>
</table>

**Word Features**
- lists of job titles,
- Lists of prefixes
- Lists of suffixes
- 350 informative phrases

**HTML/Formatting Features**
- `{begin, end, in} x `{<b>, <i>, <a>, <hN>}` x `{lengths 1, 2, 3, 4, or longer}`
- `{begin, end}` of line
IE Difficult

• Different Languages
  – Morphology is very easy in English, much harder in German.
  – Identifying word and sentence boundaries is fairly easy in European language, much harder in Chinese and Japanese.
  – Some languages use orthography (like English) while others (like Hebrew, Arabic etc) do not have it.

• Different types of style
  – Scientific papers
  – Newspapers
  – Memos
  – Emails
  – Speech transcripts

• Type of Document
  – Tables
  – Graphics
  – Small messages vs. Books
Thanks
Questions?