Text Mining from User Generated Content

Ronen Feldman
Information Systems Department
School of Business Administration
Hebrew University, Jerusalem, ISRAEL
Ronen.feldman@huji.ac.il

Lyle Ungar
Computer and Information Science
University of Pennsylvania
Philadelphia, PA 19103
ungar@cis.upenn.edu
Outline

- Intro to text mining
  - Information Retrieval (IR) vs. Information Extraction (IE)
- Information extraction (IE)
- Open IE/Relation Extraction
- Sentiment mining
- Wrap-up
Text Mining ≠ Search

Information Retrieval
Find Documents matching the Query

- Actual information buried inside documents
- Long lists of documents

Information Extraction
Display Information relevant to the Query

- 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>Find Material</th>
<th>Internet</th>
<th>Text Mining</th>
</tr>
</thead>
<tbody>
<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>
Outline

- Intro to text mining
- Information extraction (IE)
  - IE Components
- Open IE/Relation Extraction
- Sentiment mining
- Wrap-up
Information Extraction

Theory and Practice
What is Information Extraction?

- IE extracts pieces of information that are salient to the user's needs.
  - Find *named entities* such as persons and organizations
  - Find find *attributes* of those entities or *events* they participate in
  - Contrast IR, which indicates which documents need to be read by a user
- Links between the extracted information and the original documents are maintained to allow the user to reference context.
Applications of Information Extraction

- Infrastructure for IR and for Categorization
- Information Routing
- Event Based Summarization
- Automatic Creation of Databases
  - Company acquisitions
  - Sports scores
  - Terrorist activities
  - Job listings
  - Corporate titles and addresses
Why Information Extraction?

“Who is the CEO of Xerox?”

“Female CEOs of public companies”
Text Sources

- Comments and notes
  - Physicians, Sales reps.
  - Customer response centers
  - Email
  - Word & PowerPoint documents
- Annotations in databases
  - e.g. GenBank, GO, EC, PDB
- The web
  - Blogs, newsgroups
- Newswire and journal articles
  - Medline has 13 million abstracts
- Facebook, tweets, search queries, …
Document Types

- Structured documents
  - Output from CGI
- Semi-structured documents
  - Seminar announcements
  - Job listings
  - Ads
- Free format documents
  - News
  - Scientific papers
  - Blogs, tweets, Facebook status, …
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 the 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>
JERUSALEM - A Muslim suicide bomber blew apart 18 people on a Jerusalem bus and wounded 10 in a mirror-image of an attack one week ago. The carnage could rob Israel's Prime Minister Shimon Peres of the May 29 election victory he needs to pursue Middle East peacemaking. Peres declared all-out war on Hamas but his tough talk did little to impress stunned residents of Jerusalem who said the election would turn on the issue of personal security.
<table>
<thead>
<tr>
<th>MESSAGE: ID</th>
<th>TST-REU-0001</th>
</tr>
</thead>
<tbody>
<tr>
<td>SECSOURCE: SOURCE</td>
<td>Reuters</td>
</tr>
<tr>
<td>SECSOURCE: DATE</td>
<td>March 3, 1996, 11:30</td>
</tr>
<tr>
<td>INCIDENT: DATE</td>
<td>March 3, 1996</td>
</tr>
<tr>
<td>INCIDENT: LOCATION</td>
<td>Jerusalem</td>
</tr>
<tr>
<td>INCIDENT: TYPE</td>
<td>Bombing</td>
</tr>
<tr>
<td>HUM TGT: NUMBER</td>
<td>&quot;killed: 18&quot;</td>
</tr>
<tr>
<td></td>
<td>&quot;wounded: 10&quot;</td>
</tr>
<tr>
<td>PERP: ORGANIZATION</td>
<td>&quot;Hamas&quot;</td>
</tr>
</tbody>
</table>
IE - Method

- Extract raw text (html, pdf, ps, gif)
- Tokenize
- Detect term boundaries
  - We extracted *alpha 1 type XIII collagen* from …
  - Their *house council* recommended …
- Detect sentence boundaries
- Tag parts of speech (POS)
  - *John/noun saw/verb Mary/noun.*
- Tag named entities
  - Person, place, organization, gene, chemical
- Parse
- Determine co-reference
- Extract knowledge
The Language Analysis Stack

Domain Specific

Events & Facts

Entities
Candidates, Resolution, Normalization

Basic NLP
Noun Groups, Verb Groups, Numbers Phrases, Abbreviations

Metadata Analysis
Title, Date, Body, Paragraph

Sentence Marking

Language Specific

Morphological Analyzer
POS Tagging (per word)
Stem, Tense, Aspect, Singular/Plural
Gender, Prefix/Suffix Separation

Tokenization
Components of IE System

- **Must**
  - Tokenization
  - Morphological and Lexical Analysis
  - Syntactic Analysis
  - Domain Analysis
  - Zoning
  - Part of Speech Tagging
  - Sense Disambiguation
  - Shallow Parsing
  - Deep Parsing
  - Anaphora Resolution
  - Integration

- **Advisable**
  - Deep Parsing

- **Nice to have**
  - Part of Speech Tagging

- **Can pass**
  - Zoning

**Legend:**
- Red: Must
- Purple: Advisable
- Green: Nice to have
- Yellow: Can pass
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."
Business Tagging Example

**Acquisition:**
- **Acquirer:** SAP
- **Acquired:** Virsa Systems

**Company:**
- **Virsa Systems**
- **SAP**

**Professional:**
- **Name:** Shai Agassi
- **Company:** SAP
- **Position:** President of the Product and Technology Group and executive board member

**IndustryTerm:** risk management software

**Product:**
- **Microsoft Outlook**
- **MySAP ERP**

**Company:**
- **Microsoft**

---

Honoring 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."
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
Approaches for Building IE Systems: Knowledge Engineering

- 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!
IE – Templates (hand built)

<victim> was murdered
<victim> was killed
bombed <target>
bomb against <target>
killed with <instrument>
was aimed at <target>

offices in <loc>
operates in <loc>
facilities in <loc>
owned by <company>
<company> has positions
offices of <company>
Approaches for Building IE Systems: Statistical Methods

- The techniques are based on statistics
  - E.g., Conditional Random Fields (CRFs)
  - use almost no linguistic knowledge
  - 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.
Statistical Models

- **Naive Bayes model:**
  - generate class label $y_i$
  - generate word $w_i$ from $Pr(W=w_i \mid Y=y_i)$

- **Logistic regression:** conditional version of Naive Bayes
  - set parameters to maximize $\sum_i \log Pr(y_i \mid x_i)$

- **HMM model:**
  - generate states $y_1, \ldots, y_n$ from $Pr(Y=y_i \mid Y=y_{i-1})$
  - generate words $w_1, \ldots, w_n$ from $Pr(W=w_i \mid Y=y_i)$

- **Conditional version of HMMs**
  - *Conditional Random Fields (CRFs)*

*Conditional:* estimate $p(y/x)$; don’t estimate $p(x)$
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.

Language is complex!
Text Representations (Features)

- Character Trigrams
- Words/Parts of Speech
- Terms, Entities
- Linguistic Phrases
- Parse trees
- Relations, Frames, Scripts
Text mining is hard:
Language is complex

- Synonyms and Orthonyms
  - Bush, HEK

- Anaphora (and Sortal anaphoric noun phrases)
  - It, they, the protein, both enzymes

- Notes are rarely grammatical

- Complex structure

  - The first time I bought your product, I tried it on my dog, who became very unhappy and almost ate my cat, who my daughter dearly loves, and then when I tried it on her, she turned blue!
Text mining is hard

- **Hand-built systems give poor coverage**
  - Large vocabulary
    - Chemicals, genes, names
  - Zipf's law
    - activate is common;
    - colocalize and synergize
    - are not
    - Most words are very rare
  - Can’t manually list all patterns

- **Statistical methods need training data**
  - Expensive to manually label data
Text mining is easy

- Lots of redundant data
- Some problems are easy
  - IR: bag of words works embarrassingly well
  - Latent Semantic Analysis (LSA/SVD) for grading tests
- Incomplete, inaccurate answers often useful
  - Exploratory Data Analysis (EDA)
    - Suggest trends or linkages
Conclusions

- What doesn't work
  - Anything requiring high precision, broad coverage, and full automation

- What does work
  - Text mining with humans “in the loop”
    - Information retrieval, Message routing
    - Trend spotting
  - Specialized extractors
    - Company addresses, Sports scores …

- What will work
  - Using extracted info in statistical models
  - Speech to text
The Bottom Line

- Information extraction works great if you can afford to be 90% accurate
- Generally requires human post-processing for > 95%
  - Unless the system is very highly specialized
Outline

- Intro to text mining
- Information extraction (IE)
- Open IE/Relation Extraction
  - Basic Open IE: TextRunner
  - Advanced Open IE: KnowItAll and SRES
- Sentiment mining
- Wrap-up
IE for the Web

Challenges

- Difficult, ungrammatical sentences
- Unreliable information
- Heterogeneous corpus
- Massive Number of Relations

Advantages

- “Semantically tractable” sentences
- Redundancy

Open IE
[Banko, et al. 2007]
TextRunner Search

http://www.cs.washington.edu/research/textrunner/

[Banko et al., 2007]
## Kills - 42 Results

- **Strong antibiotics** (103), Antibiotics (67), Benzoyl peroxide (50), **175 more...** kills **bacteria**
- Ultraviolet disinfection devices (3), ozone (3), iodine (2), **7 more...** may **kill** **bacteria** and viruses
- Levaquin (21) **kills** a variety of **bacteria**
- INH (4), the medicine (4) **kills** the TB **bacteria**
- Many antibiotics (3), Antibiotics (2), the "bad" **bacteria** (2) also **kills** the "good" **bacteria**
- Infact Doxy (4), only the Doxy (2) **kills** a whole bunch of various **bacteria**
- Treatment (4), Penicillin treatment (2) will **kill** the syphilis bacteria
- SILVER (3), our disinfectant solution (2) **kills** almost all known **bacteria**
- Boiling (2), boil-water alerts (2) will **kill** **bacteria** and parasites
- A food (2), antibiotics (2) can **kill** all **bacteria**
- Anti-bacterial cleaners (4) **kills** 99.9 % of **bacteria** Cleans appliances
- Appropriate treatment (4) **kills** the Shigella **bacteria**
- Artemisinin (3) can **kill** other parasites and **bacteria**
- The chlorine dioxide (3) **kills** the already formed **bacteria**
- This mouthwash (3) **kills** germs and **bacteria**
- Those drugs (3) killed Andrew’s normal gut-protective **bacteria**
- Antibiotics (3) **kill** gonorrhea **bacteria**
- Proper cooking (3) **kills** food poisoning **bacteria**
- That microwaves (2) can **kill** the anthrax **bacteria**
- Hot dry vapor steam (2) **kills** mold, mildew, viruses, **bacteria**
- One application (2) **kills** **bacteria** odors
- Benzoyl peroxide (2) **kills** off **bacteria**
- Iodine (2) will **kill** the lactic **bacteria**
- The boiling (2) **kills** impurities and **bacteria**
- The chlorine (2) **kills** iron **bacteria**
- Ozone (2) **kills** iron **bacteria**
- Ampicillin (2) **kills** susceptible **bacteria**

Any positively offset frequency (2) **kills** all **bacteria**...viruses and parasites
**does not kill** - 1 result

Doxycycline (14), Freezing (11), Refrigeration does not kill bacteria

**to kill** - 6 results

antibiotics (7), water (3), milk (3), 2nd time (2) to kill extracellular bacteria, the ability (2) to kill a wide variety of bacteria, milk (2) to kill harmful bacteria, a second time (2) to kill any bacteria, macrophages (2) to kill the intracellular bacteria

**helps kill** - 2 results

Raw garlic (2), lime juice (2), UV germ, Benzoyl peroxide (3) helps kill skin bacteria

**does n't kill** - 1 result

Freezing (6), Irradiation (4), Antacids (2) does n't kill bacteria

**kill not only** - 1 result

Antibiotics (6), these drugs (3) kill not only harmful bacteria
TextRunner

[Banko, Cafarella, Soderland, et al., IJCAI '07]

100-million page corpus
Open IE

- Relation-Independent Extraction
  - How are relations expressed, in general?
  - Unlexicalized
- Self-Supervised Training
  - Automatically label training examples
- Discover relations on the fly
  - Traditional IE: \((e_1, e_2) \in R?\)
  - Open IE: **What is R?**
Training

- No parser at extraction time
- Use trusted parses to auto-label training examples
- Describe instances without parser-based features
  - Unlexicalized \(\uparrow\) PennTreeBank OK

\[
\begin{align*}
\text{John} &\quad \text{hit} &\quad \text{the} &\quad \text{ball} &\quad \text{with} &\quad \text{a} &\quad \text{bat} \\
\text{NNP} &\quad \text{VBD} &\quad \text{DT} &\quad \text{NN} &\quad \text{IN} &\quad \text{DT} &\quad \text{NN} \\
\end{align*}
\]

- (John, hit, ball)
+ (John, hit with, bat)
- (ball, with, bat)
Features

• Unlexicalized
  • Closed class words OK

• Parser-free
  • Part-of-speech tags, phrase chunk tags
  • ContainsPunct, StartsWithCapital, …

• Type-independent
  • Proper vs. common noun, no NE types
Relation Discovery

• Many ways to express one relation
• Resolver [Yates & Etzioni, HLT ‘07]

(Viacom, acquired, Dreamworks)
(Viacom, ‘s acquisition of, Dreamworks)
(Viacom, sold off, Dreamworks)

(Google, acquired, YouTube)
(Google Inc., ‘s acquisition of, YouTube)

(Adobe, acquired, Macromedia)
(Adobe, ‘s acquisition of, Macromedia)

\[ P(R_1 = R_2) \sim \text{shared objects} \times \text{strSim}(R_1, R_2) \]
## Traditional IE vs. Open IE

<table>
<thead>
<tr>
<th></th>
<th>Traditional IE</th>
<th>Open IE</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Input</strong></td>
<td>Corpus + Relations + Training Data</td>
<td>Corpus + Relation-Independent Heuristics</td>
</tr>
<tr>
<td><strong>Relations</strong></td>
<td>Specified in Advance</td>
<td>Discovered Automatically</td>
</tr>
<tr>
<td><strong>Features</strong></td>
<td>Lexicalized, NE-TYPES</td>
<td>Unlexicalized, No NE types</td>
</tr>
</tbody>
</table>
Questions

- How does OIE fare when the relation set is unknown?
- Is it even possible to learn relation-independent extraction patterns?
- How do OIE and Traditional IE compare when the relation is given?
Eval 1: Open Info. Extraction (OIE)

- CRF gives better recall than Naïve Bayes (NB) Classifiers
- Apply to 500 sentences from Web IE training corpus [Bunescu & Mooney ‘07]
- \( P = \) precision, \( R = \) Recall, \( F_1 = \frac{2PR}{P+R} \)

<table>
<thead>
<tr>
<th></th>
<th>OIE-NB</th>
<th></th>
<th>OIE-CRF</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
<td>R</td>
<td>F1</td>
</tr>
<tr>
<td></td>
<td>86.6</td>
<td>23.2</td>
<td>36.6</td>
</tr>
<tr>
<td>Category</td>
<td>Pattern</td>
<td>RF</td>
<td></td>
</tr>
<tr>
<td>---------------</td>
<td>----------------------------------------------</td>
<td>-----</td>
<td></td>
</tr>
<tr>
<td>Verb</td>
<td>$E_1$ Verb $E_2$</td>
<td>37.8</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$X$ established $Y$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Noun+Prep</td>
<td>$E_1$ NP Prep $E_2$</td>
<td>22.8</td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>the</em> $X$ settlement with $Y$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Verb+Prep</td>
<td>$E_1$ Verb Prep $E_2$</td>
<td>16.0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$X$ moved to $Y$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Infinitive</td>
<td>$E_1$ to Verb $E_2$</td>
<td>9.4</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$X$ to acquire $Y$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Modifier</td>
<td>$E_1$ Verb $E_2$ NP</td>
<td>5.2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$X$ is $Y$ winner</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coordinate$_n$</td>
<td>$E_1$ (and</td>
<td>,</td>
<td>,-</td>
</tr>
<tr>
<td></td>
<td>$X$ - $Y$ deal</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coordinate$_v$</td>
<td>$E_1$ (and</td>
<td>,) $E_2$ Verb</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td>$X$, $Y$ merge</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Appositive</td>
<td>$E_1$ NP (:</td>
<td>,) $E_2$</td>
<td>0.8</td>
</tr>
<tr>
<td></td>
<td>$X$ hometown : $Y$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Relation-Independent Patterns

- 95% could be grouped into 1 of 8 categories
- Dangerously simple
  - Paramount, the Viacom-owned studio, bought Dreamworks
  - Charlie Chaplin, who died in 1977, was born in London
- Precise conditions
  - Difficult to specify by hand
  - Learnable by OIE model
## Results

<table>
<thead>
<tr>
<th>Category</th>
<th>OIE-NB</th>
<th></th>
<th>OIE-CRF</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>P</td>
<td>R</td>
<td>F1</td>
</tr>
<tr>
<td>Verb</td>
<td>100.0</td>
<td>38.6</td>
<td>55.7</td>
<td></td>
</tr>
<tr>
<td>Noun+Prep</td>
<td>100.0</td>
<td>9.7</td>
<td>17.5</td>
<td></td>
</tr>
<tr>
<td>Verb+Prep</td>
<td>95.2</td>
<td>25.3</td>
<td>40.0</td>
<td></td>
</tr>
<tr>
<td>Infinitive</td>
<td>100.0</td>
<td>25.5</td>
<td>40.7</td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>All</td>
<td>86.6</td>
<td>23.2</td>
<td>36.6</td>
<td></td>
</tr>
</tbody>
</table>

Open IE is good at identifying verb and some noun-based relationships; others are hard because they are based on punctuation.
Traditional IE with R1-CRF

- Trained from hand-labeled data per relation
- Lexicalized features, same graph structure
- Many relation extraction systems do this  
  [e.g. Bunescu ACL ‘07, Culotta HLT ’06]

- Question: what is effect of
  - Relation-specific/independent features
  - Supervised vs. Self-supervised Training
    keeping underlying models equivalent
Eval 2: Targeted Extraction

- Web IE corpus from [Bunescu 2007]
  - Corporate-acquisitions (3042)
  - Birthplace (1853)
- Collected two more relations in same manner
  - Invented-Product (682)
  - Won-Award (354)
- Labeled examples by hand
### Results

<table>
<thead>
<tr>
<th>Relation</th>
<th>R1-CRF</th>
<th>OIE-CRF</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
<td>R</td>
</tr>
<tr>
<td>Acquisition</td>
<td>67.6</td>
<td>69.2</td>
</tr>
<tr>
<td>Birthplace</td>
<td>92.3</td>
<td>64.4</td>
</tr>
<tr>
<td>InventorOf</td>
<td>81.3</td>
<td>50.8</td>
</tr>
<tr>
<td>WonAward</td>
<td>73.6</td>
<td>52.8</td>
</tr>
<tr>
<td>All</td>
<td>73.9</td>
<td>58.4</td>
</tr>
</tbody>
</table>

Open IE can match precision of supervised IE *without*

- Relation-specific training
- 100s or 1,000s of examples *per relation*
Summary

- Open IE
  - High-precision extractions without cost of per-relation training
  - Essential when number of relations is large or unknown

- May prefer Traditional IE when
  - High recall is necessary
  - For a small set of relations
  - And can acquire labeled data

- Try it!

http://www.cs.washington.edu/research/textrunner
Outline

- Intro to text mining
- Information extraction (IE)
- Open IE/Relation Extraction
  - Basic Open IE: TextRunner
  - Advanced methods: KnowItAll and SRES
- Sentiment mining
- Wrap-up
Self-Supervised Relation Learning from the Web
KnowItAll (KIA)

- Developed at University of Washington by Oren Etzioni and colleagues
  - (Etzioni, Cafarella et al. 2005).
- 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.
- **Input** is a set of entity classes to be extracted, such as “city”, “scientist”, “movie”, etc.,
- **Output** is a list of entities extracted from the Web.
KnowItAll’s Relation Learning

- The base version uses hand written patterns based on a general Noun Phrase (NP) tagger.
- The patterns used for extracting instances of
  - the Acquisition(Company, Company) relation:
    - NP2 "was acquired by" NP1
    - NP1 "'s acquisition of" NP2
  - the MayorOf(City, Person) relation:
    - NP "', mayor of" <city>
    - <city> "'s mayor" NP
    - <city> "mayor" NP
SRES

- **SRES (Self-Supervised Relation Extraction System)**
  - learns to extract relations from the web in an unsupervised way.
  - takes as input the name of the relation and the types of its arguments
    - And a set of “seed” examples
  - Generates positive and negative examples
  - returns as output a set of extracted instances of the relation
SRES Architecture

Input:
Target Relations Definitions
Web

Output:
Extractions

Sentence Gatherer

Sentences

Seeds Generator

Pattern Learner

NER Filter (optional)

Instance Extractor

Classifier

keywords

seeds

patterns

instances
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
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

- Change 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>”
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 $\text{PatternsSet}(P)$. 
Example Pattern Alignment

- $S_1 = “Toward this end, <Arg_1> in July acquired <Arg_2>”$
- $S_2 = “Earlier this year, <Arg_1> acquired <Arg_2>”$
- After the dynamical programming-based search, the following match will be found:

|   | Toward | (cost 2) | Earlier | (cost 2) | this | (cost 0) | this | (cost 0) | end | (cost 2) | year | (cost 2) | , | (cost 0) | , | (cost 0) | <Arg_1> | (cost 0) | <Arg_1> | (cost 0) | in July | (cost 4) | acquired | (cost 0) | acquired | (cost 0) | <Arg_2> | (cost 0) | <Arg_2> | (cost 0) |
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

• **Remove** from each pattern all function words and punctuation marks that are surrounded by skips on both sides.
  • Thus, the pattern
    • *this*, *<Arg1>* *acquired* *<Arg2>*
  • from the example above will be converted to
    • *<Arg1>* *acquired* *<Arg2>*
Content Based Filtering

• Every pattern must contain at least one word relevant (defined via WordNet) to its predicate.

• For example, the pattern
  \!<Arg1>! \* by \<Arg2>!

• will be removed, while the pattern
  \!<Arg1>! \* purchased \<Arg2>!

• will be kept, because the word “purchased” can be reached from “acquisition” via synonym and derivation links.
Scoring the Patterns

- Score the filtered patterns by their performance on the positive and negative sets.
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
Score Extractions using a Classifier

- Score each extraction using the information on the instance, the extracting patterns and the matches.

- Assume 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 $P$. 
Experimental Evaluation

- We want to answer the following questions:
  1. Can we train SRES’s classifier once, and then use the results on all other relations?
  2. How does SRES’s performance compare with KnowItAll and KnowItAll-PL?
Sample Output

- HP – Compaq merger
  - The Packard Foundation, which holds around ten per cent of HP stock, has decided to vote against the proposed merger with Compaq.
  - Although the merger of HP and Compaq has been approved, there are no indications yet of the plans of HP regarding Digital GlobalSoft.
  - During the Proxy Working Group's subsequent discussion, the CIO informed the members that he believed that Deutsche Bank was one of HP's advisers on the proposed merger with Compaq.
  - It was the first report combining both HP and Compaq results since their merger.
  - As executive vice president, merger integration, Jeff played a key role in integrating the operations, financials and cultures of HP and Compaq Computer Corporation following the 19 billion merger of the two companies.
Cross-Classification Experiment

Precision

Acquisiton

Merger

0.7
0.75
0.8
0.85
0.9
0.95
1
0 50 100 150

Acq.
CEO
Inventor
Mayor
Merger
Results!

![Graph showing precision versus correct extractions for different models: KIA, KIA-PL, SRES, and S_NER.]
Inventor Results
When is SRES better than KIA?

- KnowItAll extraction works well when
  - redundancy is high
  - most instances have a good chance of appearing in simple forms

- SRES is more effective for low-frequency instances due to
  - more expressive rules
  - classifier that inhibits those rules from overgeneralizing.
The Redundancy of the Various Datasets

Datasets redundancy

Average sentences per instance

Acq  Merger  Inventor  CEO  Mayor
Outline

- Intro to text mining
- Information extraction (IE)
- Open IE/Relation Extraction
- Sentiment mining
  - And its relation to IE: CARE
- Wrap-up
Traditional Text Mining is not cost effective nor time efficient

- Why?
  - takes too long to develop
  - too expensive
  - not accurate enough
  - lacks complete coverage
A pure statistical approach is not accurate enough

- Without semantic comprehension - is Apple a fruit or a company?
- ‘We reduced our deficit’ – need proper negation handling
Rule writing approach

• Domain specific
• Very long development cycle
• Expensive process
• No guarantee of full pattern coverage
The Evolution of Information Extraction Technology

- Unsupervised IE
- Generic Grammar Augmented IE
- Hybrid Information Extraction
- Supervised Information Extraction
- Rule Based Information Extraction

Care 1.0
Care 2.0
Care 2.0 + Corpus Based Learning
HMM, CRF
DIAL
Example of Unsupervised IE Results

Actos; weight gain (40 (P: 38, N: 2))

Rel_take_DRU_has_SYM(DRUG, SYMPTOM)

Negative (1)
- I've been taking 15 mg of Actos for just over a year now and so far (knock on wood) I haven't had the weight gain that some others have reported as a side effect.

Positive (8)
- I also have read here about some of you who have been on the Actos and the weight gain you had experienced.

We saw an endo because of all of the weight gain and side effects from taking actos. He was on Actos but went off of it because of weight gain and stomach bloating.
- I really don't want to go back on Actos because of weight gain/fluid retention.
- My doctor wanted me to start Actos for awhile, until the Byetta kicks in, but I stopped Actos in the first place because of weight gain and I said no to restarting that.
- I started taking Actos first on May 2, 2007 and I started Metformin 3 weeks later I can not take the Spironolactone till Aug but I have noticed that I have gained weight with these 2 drugs instead of losing and I got a treadmill and do 30 min every morning when I get up and lately I have been doing 30 min at night too because of the weight gain.
- I have experienced weight gain as well and i am on Actos and insulin and glucophage.
- I guess that everything comes with a price, but I'm wondering if most folks who have tried Actos have experienced weight gain and the other side effects (edema, headaches, nausea, fatigue, etc.).

Rel_SYMofDRU(SYMPTOM, DRUG)

Positive (5)
- I do notice that it increases my hunger, so it is possible that Actos weight gain issues may be from hunger being stimulated.
- I don't think that a lot of us had made the Actos induced weight gain connection. One reported side effect of Actos is weight gain.

I have changed to a new MD and when I discussed my concern over the weight gain with Avandia and then Actos, he...
Actos hasn't caused any weight gain, I am still losing some.

Positive (25)
I also am on Synthroid, Atenolol, Diovan, Lotrel, Lexapro, Vitorin and Prilosec OTC. I didn't realize that Actos can cause a weight gain
as I had never read it as a side effect; however, after reading all of the comments on this site, I now know why my weight has increased over the past few months since taking on it.

I don't take any oral meds, but from what I have read here, Actos causes weight gain because of water retention.

why does the endo think you’re a type 1? oral meds are usually given only to type 2’s, as type 2’s have insulin resistance.
oral meds
treat the insulin resistance. type 1's require insulin.....i take actoplus met- which is actos and metformin.actos is like avandia and
i've had no heart issues.....tho-avandia and actos can cause weight gain....take care, trish
Actos causes edema and weight gain also.
Actos can cause weight gain (so can Avandia, it's cousin)

Now I have started to see a lot of reports of Actos causing weight gain, among other things.
for the record, Actos can and does, cause weight gain/water retention.
I'm on both - what did you hate about Metformin? (Actos causes weight gain, metformin weight loss)
Also I hear that the Actos causes weight gain, so now I am afraid the new pill will cause me to gain weight.
I'm type 1 so only on insulin, but I have heard that Actos can cause weight gain.
Avandia & Actos, especially in combination with insulin, causes fluid retention and/or fat weight gain.

My endocrinologist warned me that Actos can cause significant weight gain.

Actos caused weight gain and fluid retention in my chest.

Metformin causes weight loss, Avandia and Actos causes the birth of new fat cells and weight gain.

……
Sentiment Analysis of Stocks from News Sites
The Need for Event Based SA

*Toyota announces voluntary recall of their highly successful top selling 2010 model-year cars*

- Phrase-level SA:
  - highly successful top selling $\Rightarrow$ positive
  - Or at best neutral
    - Taking into account voluntary recall $\Rightarrow$ negative
- Need to recognize the whole sentence as a “product recall” event!
The CaRE extraction engine processes financial content from various sources such as Reuters, Bloomberg, Market Watch, CNN, Barrons, etc. It involves the following steps:

1. Crawling: Collecting financial content from various sources.
2. Cleaning and Extraction of the Main Textual Content from HTML Pages: Pre-processing to extract the main textual content.
3. Identification of Relevant Sentences to the Main Company: Identifying sentences relevant to the main company.
5. Scoring the Document Set using Decaying Effects: Assigning scores to documents based on their relevance and recency.
7. COMPANY SCORE: Final score and ranking of the documents.

The process starts with the financial content being crawled, cleaned, and pre-processed, followed by the extraction of relevant sentences and sentiment analysis, leading to the scoring of individual documents and the assignment of a company score.
Template Based Approach to Content Filtering

Why Subprime Lenders Are In Trouble

New analysis suggests that subprime lenders lowered their lending standards last year as they competed for business.

by Peter Coy

As the subprime mortgage market goes into steep decline, threatening to drag the whole economy along with it, many people are wondering what could have gone so wrong so quickly. Until recently, after all, delinquency and foreclosure rates on subprime loans were reassuringly low.

The answer may lie in how the quality of these mortgages has changed over the years. While subprime is the term used for loans issued to people with poor credit, not all subprime loans are created equal. And the subprime loans that were originated in 2006 that are turning out to be shockingly weak.

Why 2006? What happened last year that caused credit quality to go into steep decline? Michael Youngblood, head of asset-backed securities research at Friedman, Billings, Ramsey Group (FBG), has been poring over the data. He points out that there was a sudden but little-noticed shift in lenders’ strategy that occurred at the end of 2005. Lenders went from competing for customers on price (by lowering rates) to competing for customers on easy terms (by lowering lending standards).

UNDISCLOSED CHANGE

The change was little-noticed, says Youngblood, because the lenders actively denied it. "To my disappointment as a long-time analyst," says Youngblood, the major lenders insisted that they had not lowered their credit standards long after they had begun to do so. "I met with all of the public subprime lenders in early June... to a man they swore they had maintained 2005..."
Hybrid Sentiment Analysis

- All levels are part of the same rulebook, and are therefore considered simultaneously by CaRE
Dictionary-based sentiment

- Started with available sentiment lexicons
  - Domain-specific and general
  - Improved by our content experts
- Examples
  - Modifiers: attractive, superior, inefficient, risky
  - Verbs: invents, advancing, failed, lost
  - Nouns: opportunity, success, weakness, crisis
  - Expressions: exceeding expectations, chapter 11
- Emphasis and reversal
  - successful, extremely successful, far from successful
Event-Based Sentiment

- Product release/approval/recall, litigations, acquisitions, workforce change, analyst recommendations and many more
- Semantic role matters:
  - Google is being sued/is suing…
- Need to address historical/speculative events
  - Google acquired YouTube in 2006
  - **What if Google buys Yahoo** and the software giant Microsoft remains a single company fighting for the power of the Internet?
Why did we get a Positive Spike?

Articles

- Cliffs Natural Resources Gets Into Dilution
  - Score: +0.76
  - Date: 6/13/2011
  - Source: forbes.com

- (AWM) Analysts upbeat on Cliffs growth, offering of shares
  - Score: +0.40
  - Date: 6/13/2011
  - Source: metalbulletin.com

- UBS Initiates Coverage of Cliffs Natural Resources With a Buy Rating and A $122 Target Price
  - Score: +0.93
  - Date: 6/13/2011
  - Source: fnr.com

- UBS Initiates Coverage on Cliffs Natural Resources
  - Score: +0.86
  - Date: 6/13/2011
  - Source: streetinsider.com

- On The Fly Analyst Initiation Summary
  - Score: +0.76
  - Date: 6/13/2011
  - Source: finance.yahoo.com

- Nokia, LinkedIn: Analysts' New Ratings
  - Score: +0.76
  - Date: 6/13/2011
  - Source: thestreet.com

- UBS Starts Cliffs Natural Resources
  - Score: +0.86
  - Date: 6/13/2011
  - Source: streetinsider.com

Events

- UBS Initiates Coverage of Cliffs Natural Resources (NYSE:CLF) today with a buy rating and a $122 price target. The bank says the company’s seaborne exposure drives strong projected cash flow, as the supply for seaborne iron ore will be tight over the next two years. The Consolidated Thompson acquisition increases Cliffs’ seaborne exposure and diversifies the company’s exposure to U.S. iron ore and coal segments, according to UBS. Cliffs Natural Resources has a potential upside of 50.3% based on a current price of $85.04 and an average consensus analyst price target of $127.63. By Tim Tracy tracy@fnr.com Latest Breaking Video News Morgan Stanley lifted its price target for Air Products and Chemicals (NYSE:APD) to $115 from $105 and maintained its overweight rating on the stock. The bank sees 2011 EPS of $5.00, up from $5.76 per share, and 2012 EPS of $6.63, up from $6.45 per share. Air Products & Clif...
JC Penny

J.C Penney Company Inc (NYSE:JCP)

From: 07/23/2010   To: 09/15/2010

Impact   Price

$29  $28  $27  $26  $25  $24  $23  $22  $21  $20  $19  $18  $17  $16  $15  $14  $13  $12  $11  $10  $9  $8  $7  $6  $5  $4  $3  $2  $1  $0

Impact  Price

Jul 23 24 25 26 27 28 29 30 31  Aug 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15

SNTA 3 Months
Key Developments
Boeing, National Renewable Energy Lab to Collaborate on DOD Energy Security

Boeing Receives Contract for Apache Helicopter Rotor Heads

Boeing [NYSE: BA] today announced that the U.S. Defense Logistics Agency Aviation has awarded the company a three-year contract to manufacture spare main rotor heads for the U.S. Army’s fleet of Apache attack helicopters. The fixed-price contract calls for an initial 100 main rotor heads, at a value of $31 million, and allows for 50 additional rotor heads, which would increase the contract value to $48 million.

Boeing Receives Contract for Apache Helicopter Rotor Heads

Boeing Receives A-10 Structural Inspection Contract from US Air Force

Boeing Awarded $19 Million Follow-on US Navy Field Services Contract

Full Article

Boeing Receives Contract for Apache Helicopter Rotor Heads

MESA, Ariz., Oct. 4, 2010 -- Boeing [NYSE: BA] today announced that the U.S. Defense Logistics Agency Aviation has awarded the company a three-year contract to manufacture spare main rotor heads for the U.S. Army’s fleet of Apache attack helicopters. The fixed-price contract calls for an initial 100 main rotor heads, at a value of $31 million, and allows for 50 additional rotor heads, which would increase the contract value to $48 million.

The Army Integrated Logistics division of Boeing Global Services & Support will perform the work at the Boeing facility in Mesa. "This contract award is another testament to Boeing employees' excellent work and commitment to our customers," said Joseph O'Shaughnessy, Boeing Defense, Space & Security vice president of Army Programs and Logistics.

Boeing award for Apache rotor heads is part of a recent series of DOD contracts awarding the company for other Apache-related work. The company recently received a $28 million contract to provide modified Apache AH-64E Advanced CombatHelicopters with enhanced avionics.

The company also recently was awarded a $13 million contract to provide Apache AH-64D Longbow helicopters to the Swiss Army and a $7 million award to provide Apache AH-64D Longbow helicopters for the German Army.

Boeing’s Defense, Space & Security business provides products, services and solutions to customers across a broad range of defense, aerospace and security markets. It is comprised of four major product/capability areas: Intelligence, Surveillance and Electronic Warfare Systems, Defense Services, Aircraft and Missiles.
Mining Medical User Forums
The Text Mining Process

- **Downloading**: html-pages are downloaded from a given forum site
- **Cleaning**: html-like tags and non-textual information like images, commercials, etc… are cleaned from the downloaded text
- **Chunking**: The textual parts are divided into informative units like threads, messages, and sentences
- **Information Extraction**: Products and product attributes are extracted from the messages
- **Comparisons**: Comparisons are made either by using co-occurrence analysis or by utilizing learned comparison patterns
We downloaded messages from 5 different consumer forums

- diabetesforums.com
- healthboards.com
- forum.lowcarber.org
- diabetes.blog.com
- diabetesdaily.com

** Messages in Diabets.blog.com were focused mainly on Byetta
Side Effects and Remedies

Red lines – side effects/symptoms
Blue lines - Remedies

See what **causes** symptoms and what **relieves** them

See what **positive** and **negative** effects a drug has

See which symptoms are most **complained** about

Created by NodeXL (http://nodexl.codeplex.com)
Drugs Taken in Combination

[Diagram showing various medications connected by lines.]
Several Pockets of drugs that were mentioned frequently together in a message were identified.

Byetta was mentioned frequently with:
- Glucotrol
- Januvia
- Amaryl
- Actos
- Avandia
- Prandin
- Symlin

- Lifts larger than 3
- Width of edge reflects how frequently the two drugs appeared together over and beyond what one would have expected by chance
Drug Usage Analysis

Drug Co-taking – Drugs mentioned as “Taken Together”

There are two main clusters of drugs that are mentioned as “taken together”

- Byetta was mentioned as “taken together” with:
  - Januvia
  - Symlin
  - Metformin
  - Amaryl
  - Starlix

Pairs of drugs that are taken frequently together include:
- Glucotrol--Glucophage
- Glucophage--Stralix
- Byetta--Januvia
- Avandia--Actos
- Glucophage--Avandia

- Lifts larger than 1
- Width of edge reflects how frequently the two drugs appeared together over and beyond what one would have expected by chance
Drug Usage Analysis

Drug Switching – Drugs mentioned as “Switched” to and from

There are two main clusters of diabetes drugs within which consumers mentioned frequently that they “switched” from one drug to another.

Byetta was mentioned as “switched” to and from:
• Symlin
• Januvia
• Metformin

- Lifts larger than 1
- Width of edge reflects how frequently the two drugs appeared together over and beyond what one would have expected by chance
Byetta appeared much more than chance with the following side effects:
- “Nose running” or “runny nose”
- “No appetite”
- “Weight gain”
- “Acid stomach”
- “Vomit”
- “Nausea”
- “Hives”
Drug Terms Analysis
Drug Comparisons on Side Effects

Byetta shares with Januvia the side effects:
- Runny nose
- Nausea
- Stomach ache
- Hives

Byetta shares with Levemir the side effects:
- No appetite
- Hives

Byetta shares with Lantus the side effects:
- Weight gain
- Nose running
- Pain

Note that only Byetta is mentioned frequently with terms like “vomit”, “acid stomach” and “diarrhea”

The main side effects discussed with Januvia:
- Thyroid
- Respiratory infections
- Sore throat

The main side effects discussed with Levemir:
- No appetite
- Hives

The main side effects discussed with Lantus:
- Weight gain
- Nose running
- Pain

- Lifts larger than 1
- Width of edge reflects how frequently the two drugs appeared together over and beyond what one would have expected by chance
Drug Terms Analysis

Byetta – Positive Sentiments

Byetta appeared much more than chance (lift \( >2 \)) with the following positive sentiments:

- “Helps with hunger”
- “No nausea”
- “Easy to use”
- “Works”
- “Helps losing weight”
- “No side effects”
Drug Terms Analysis

Drug Comparisons on Positive Sentiments

The main positive sentiments discussed with Januvia:
• “No nausea”
• “Better blood sugar”
• “Works”
• “No side effects”

The main positive sentiments discussed with Levemir:
• “Easy to use”
• “Fast acting”

The main positive sentiments discussed with Lantus:
• “Fast acting”
• “Works”

Byetta shares with Januvia:
• “Better blood sugar”
• “No nausea”
• “Helps lose weight”
• “No side effects”
• “Works”

Byetta shares with Levemir:
• “Easy to use”
• “Helps lose weight”
• “No side effects”
• “Works”

Byetta shares with Lantus:
• “Easy to use”
• “No side effects”
• “Works”

Note that only Byetta is mentioned frequently with “helps with hunger” (point of difference)

Lifts larger than 0.5
Width of edge reflects how frequently the two drugs appeared together over and beyond what one would have expected by chance
Drug Terms Analysis

Byetta – Other Sentiments

Byetta appeared much more than chance (lift>1.5) with the following sentiments:
• “Twice a day”
• “Discontinue”
• “Injection”
• “Once a day”

Byetta was mentioned moderately with the sentiments:
• “Pancreatitis”
• “Free sample”
Visual care

IE authoring environment
Overall architecture

- Generic Preparation
- Generic Grammar
- Domain-specific Preparation
  - Relations definitions, Domain-specific lexicon, Post-processor definitions
- Information Extraction
Generic preparation stage

- CRF-based sequence classifier training
- PoS-labeled corpus
- NER-labeled corpus
- PoS model
- NER model
- Manually-written grammar
- Generic Rulebook
Domain specific preparation stage

- Unlabeled corpus of domain-specific sentences
- Relations definitions, Domain-specific lexicon, Post-processor definitions
- Generic Rulebook
- CARE-II Engine
- Visual CARE-II
- Manual input
Information extraction stage

- Input documents
- Parsed documents
- Output Relations

- Generic Rulebook
- CARE-II Engine
- Relations definitions, Domain-specific lexicon, Post-processor definitions
- Post-processor, Co-reference resolution
Wrap-up
What did we cover?

- Intro to text mining
  - Information Retrieval (IR) vs. Information Extraction (IE)
- Information extraction (IE)
  - IE Components
- Open IE/Relation Extraction
  - Basic Open IE: Text Runner
  - Advanced Open IE: KnowItAll and SRES
- Sentiment mining
  - its relation to IE
- Visualization of text mining results
  - What is compared to what, and how
- Wrap-up
Text Mining is Big Business

- Part of most big data mining systems
  - SAS, Oracle, SPSS, SAP Fair Isaac, …
- Many sentiment analysis companies
  - the “big boys,” Nielsen Buzzmetrics and dozens of others.
- Sometimes tied to special applications
  - Autonomy - suite of text mining, clustering and categorization solutions for knowledge management
  - Thomson Data Analyzer - analysis of patent information, scientific publications and news
- Open source has more fragmented tools
  - NLTK, Stanford NLP tools, GATE, Lucene, MinorThird
  - RapidMiner/YALE - open-source data and text mining
- Lots more:
  - AeroText - Information extraction in multiple languages
  - LanguageWare - the IBM Tools for Text Mining
  - Attensity, Endeca Technologies, Expert System S.p.A., Nstein Technologies
Summary

- Information Extraction
  - Not just information retrieval
  - Find named entities, relations, events
  - Hand-built vs. trained models
    - CRFs widely used
- Open Information Extraction
  - Unsupervised relation extraction
    - Bootstrap pattern learning
- Sentiment analysis
- Visualize results
  - Link analysis, MDS, …
- Text mining is easy and hard
To learn more

- See www.cis.upenn.edu/~ungar/KDD/text-mining.html