CS639: Data Management for Data Science

Lecture 21: Information Extraction

Theodoros Rekatsinas
So far...

1. Manage data of various forms (structured, key-values, documents)
   1. RDBMS
   2. MadReduce
   3. Key-value Stores

2. How to learn models that capture the distribution of observed data
   1. Statistics and Statistical Inference
   2. Linear Classifiers
   3. Decision Trees
   4. Unsupervised/Supervised learning
   5. Optimization
Until the end of the semester...

1. Information extraction and Data Integration

2. Communicating insights
   1. Visualizations and Privacy
Information Extraction

1. Extracting knowledge from unstructured data (e.g., text)

2. Recognize Named Entities in unstructured data

3. Clean and normalize extractions
What is Information Extraction?

Goal: Mine knowledge from unstructured data
Growth of Unstructured Text Data

Unstructured data, primarily text, account for more than 80% of the data collected by organizations.
Knowledge in unstructured data

- News
  - Social media post
  - Web pages
- Financial reports
  - Medical records
  - Legal acts
- Customer reviews
  - Tech support memos
  - Field service notes

Get overview of recent news events
Obtain insights from data for decision support
Summarize user feedbacks for quality control
Knowledge from Unstructured Data (Example)
Personalized medicine

intellectual disability with impaired speech development and aggressive behavior

83 candidate genes in her exome with rare variants

AC018470.1, ACAP3, ADAP1, AMPD1, ASPM, ASXL2, BAZ1B, BHLHE22, BTBD9, C17orf104, C17orf74, C19orf26, C1orf87, C2orf81, CCNL2, CDH10, CHD6, CNOT3, COL6A5, DCHS2, DEAF1, DNM1, FAM216B, FAM73B, FAM83H, FAM84B, FAT3, FBXO25, FCRLB, FLJ00104, FRS2, GRK7, HEPHL1, HOXD11, IL19, INSRR, IQCC, KIAA0825, LAMA5, LAMC3, LGR6, MAST4, MBD6, MBLAC2, MCM10, MDH2, METRN, MSL2, N4BP3, NCKAP5, NUP50, NYNRIN, ORC3, PDCD2L, PDXP, PLEKHG1, PLIN2, POU3F2, PXMP2, RAB11FIP1, RASSF1, RIMS1, RTKN2, SASS6, SERPINA3, SH3BP1, SHB, SLC2A9, SLC38A8, SON, SP8, SPTBN5, SRRM2, TAAR1, TARSL2, TET2, TRIM72, TSPAN15, TSPYL4, WDR20, XPNPEP1, ZFYVE16, ZNF469, ZSCAN29
Personalized medicine

Phenotype → Gene variant?

Mutation → Gene

→ Disease

Which gene is at fault?
Personalized medicine

Phenotype → Gene variant → Mutation → Gene → Disease

Find right article (1hr/variant)

Which gene is at fault?

PubMed

25 million articles

National Library of Medicine

Doctor Google
Personalized medicine

Phenotype → Gene variant?

Mutation → Gene

Disease

Find right article (1hr/variant)

Which gene is at fault?

Can we build a machine to read these articles?

PubMed

25 million articles
Personalized medicine

Phenotype \rightarrow \text{Gene variant?} \rightarrow \text{Mutation} \rightarrow \text{Gene} \rightarrow \text{Disease}

Query KB (Instantaneous)

\text{Knowledge Base}

<table>
<thead>
<tr>
<th>Gene</th>
<th>Phenotype</th>
</tr>
</thead>
<tbody>
<tr>
<td>DEAF1</td>
<td>Intellectual Disability</td>
</tr>
</tbody>
</table>

\text{Knowledge Base Construction (KBC)}

\text{Cheaper}

\text{Faster}

\text{Scalable}
Knowledge Extraction from Unstructured Data

1. Step 1: Identify Entities of interest
2. Step 2: Identify relations that these entities participate in
3. Step 3(*): Identify events
Entities

Can computational systems identify real-world entities of different categories from given corpora?

- **Organization**: United States, Red Cross, US government...
- **Person**: Ray Nagin, Mayor, President Bush...
- **Location**: New Orleans, Louisiana, Washington DC...

- **Criticism of government response to the hurricane**...
Can computational systems capture different relations between the entities from given corpora?

American Airlines, a unit of AMR corp., immediately matched the move, spokesman Tim Wagner said. United Airlines, a unit of UAL corp., said the increase took effect Thursday night.

<table>
<thead>
<tr>
<th>Entity 1</th>
<th>Relation</th>
<th>Entity 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>American Airlines</td>
<td>is_subsidiary_of</td>
<td>AMR</td>
</tr>
<tr>
<td>Tim Wagner</td>
<td>is_employee_of</td>
<td>American Airlines</td>
</tr>
<tr>
<td>United Airlines</td>
<td>is_subsidiary_of</td>
<td>UAL</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Events

Can computational systems identify real-world event of different types from given corpora?

Terrorism Template

LOCATION
CHILE: MOLINA (CITY)

TYPE
ROBBERY

Date
07 JAN 90

text corpus

Criticism of government response to the hurricane...
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...
What is Information Extraction

Information Extraction = segmentation + classification + clustering + association

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”
What is Information Extraction

Information Extraction = segmentation + classification + association + clustering

As a family of techniques:

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 family of techniques:

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...
What is Information Extraction

As a family of techniques:

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...
Traditional Rule-Based Systems

Domain experts → Extraction rules → Text corpus

Handcraft

... cities such as NPList ...

City
NPList[0]
NPList[1]
...

Extraction rules

“The tour includes major cities such as [New York], [Los Angeles], and [Dallas]”

City
New York
Los Angeles
Dallas

Entities
Supervised Machine Learning-Based Systems (state-of-the-art)
IE as Supervised Learning

- **Training set**
  - TSS
  - TSS
  - TSS
  - TSS
  - Not TSS
  - Not TSS
  - Not TSS
  - Not TSS

- **Testing set**
  - 1. Not TSS
  - 2. TSS
  - 3. TSS
  - 4. Not TSS
  - 5. Not TSS
  - 6. TSS
  - 7. Not TSS
  - 8. TSS

- **Model**
  - Machine learning algorithm
  - Prediction algorithm

**Predicted labels**
- 1. Not TSS
- 2. TSS
- 3. TSS
- 4. Not TSS
- 5. Not TSS
- 6. TSS
- 7. Not TSS
- 8. TSS
IE as Supervised Learning

News articles

“President Barack Obama and his wife Michelle…”

Learning and Inference

Structured Output

<table>
<thead>
<tr>
<th>Mention1</th>
<th>Mention2</th>
<th>IsSpouse</th>
</tr>
</thead>
<tbody>
<tr>
<td>Michelle Obama</td>
<td>Barack Obama</td>
<td>T</td>
</tr>
</tbody>
</table>
Candidate Extraction

Michelle Obama is married to President Barack Obama.

<table>
<thead>
<tr>
<th>Mention</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Michelle Obama</td>
<td>PERSON</td>
</tr>
<tr>
<td>Barack Obama</td>
<td>PERSON</td>
</tr>
<tr>
<td>President</td>
<td>TITLE</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Mention1</th>
<th>Mention2</th>
<th>HasSpouse</th>
</tr>
</thead>
<tbody>
<tr>
<td>Michelle Obama</td>
<td>Barack Obama</td>
<td></td>
</tr>
</tbody>
</table>
Candidate Extraction++

Remember:
The goal is to maximize recall!

Regular expressions

```
/\^#?([α-f0-9]{6}|[α-f0-9]{3})\$/
```
Feature Extraction

Michelle Obama is married to President Barack Obama.

<table>
<thead>
<tr>
<th>Mention1</th>
<th>Mention2</th>
<th>HasSpouse</th>
</tr>
</thead>
<tbody>
<tr>
<td>Michelle Obama</td>
<td>Barack Obama</td>
<td></td>
</tr>
</tbody>
</table>
Feature Extraction

Michelle Obama is married to President Barack Obama.

<table>
<thead>
<tr>
<th>Mention1</th>
<th>Mention2</th>
<th>HasSpouse</th>
</tr>
</thead>
<tbody>
<tr>
<td>Michelle Obama</td>
<td>Barack Obama</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Mention1</th>
<th>Mention2</th>
<th>feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>M. Obama</td>
<td>B. Obama</td>
<td>PERSON - mary - PERSON</td>
</tr>
<tr>
<td>M. Obama</td>
<td>B. Obama</td>
<td>Distance=3</td>
</tr>
</tbody>
</table>
Feature Extraction

Previously users would write features by hand

Michelle Obama is married to President Barack Obama.

- Word_in_between[“marry”]
- Distance<=5

...Now, most users rely on automated methods

Recursive Neural Networks (RNNs)

Treedlib (our library)

...However, these automated methods all rely on having a large (but noisy?) labeled training set!
Distant Supervision

Leverage existing knowledge bases, dictionaries to obtain training data via matching to the input corpus

Michelle Obama is married to President Barack Obama.

Spousal Relationship

<table>
<thead>
<tr>
<th>Person 1</th>
<th>Person 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Barack Obama</td>
<td>Michelle Obama</td>
</tr>
<tr>
<td>Nicolas Sarkozy</td>
<td>Carla Bruni</td>
</tr>
<tr>
<td>Hillary Clinton</td>
<td>Bill Clinton</td>
</tr>
</tbody>
</table>
Distant Supervision

Corpus Text

Bill Gates founded Microsoft in 1975.
Bill Gates, founder of Microsoft, ...
Bill Gates attended Harvard from ...
Google was founded by Larry Page ...

Freebase

(Bill Gates, Founder, Microsoft)
(Larry Page, Founder, Google)
(Bill Gates, CollegeAttended, Harvard)

Training Data

[Adapted example from Luke Zettlemoyer]
Distant Supervision

Corpus Text
- Bill Gates founded Microsoft in 1975.
- Bill Gates, founder of Microsoft, ...
- Bill Gates attended Harvard from ...
- Google was founded by Larry Page ...

Freebase
- (Bill Gates, Founder, Microsoft)
- (Larry Page, Founder, Google)
- (Bill Gates, CollegeAttended, Harvard)

Training Data
- (Bill Gates, Microsoft)
- Label: Founder
- Feature: X founded Y

[Adapted example from Luke Zettlemoyer]
Distant Supervision

Corpus Text

Bill Gates founded Microsoft in 1975.
Bill Gates, founder of Microsoft, ...
Bill Gates attended Harvard from ...
Google was founded by Larry Page ...

Freebase

(Bill Gates, Founder, Microsoft)
(Larry Page, Founder, Google)
(Bill Gates, CollegeAttended, Harvard)

Training Data

(Bill Gates, Microsoft)
Label: Founder
Feature: X founded Y
Feature: X, founder of Y

[Adapted example from Luke Zettlemoyer]
Distant Supervision

Corpus Text

Bill Gates founded Microsoft in 1975.
Bill Gates, founder of Microsoft, ...
Bill Gates attended Harvard from ...
Google was founded by Larry Page ...

Freebase

(Bill Gates, Founder, Microsoft)
(Larry Page, Founder, Google)
(Bill Gates, CollegeAttended, Harvard)

Training Data

(Bill Gates, Microsoft)
Label: Founder
Feature: X founded Y
Feature: X, founder of Y

(Bill Gates, Harvard)
Label: CollegeAttended
Feature: X attended Y

For negative examples, sample unrelated pairs of entities.

[Adapted example from Luke Zettlemoyer]
IE as supervised learning
Fonduer: An example state-of-the-art system
Fonduer: An example state-of-the-art system

Richly formatted data

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Symbol</th>
<th>Value</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Collector-emitter voltage</td>
<td>$V_{CEO}$</td>
<td>40</td>
<td>V</td>
</tr>
<tr>
<td>Collector-base voltage</td>
<td>$V_{CBO}$</td>
<td>60</td>
<td></td>
</tr>
<tr>
<td>Emitter-base voltage</td>
<td>$V_{EBO}$</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>Collector current</td>
<td>$I_C$</td>
<td>200</td>
<td>mA</td>
</tr>
<tr>
<td>Total power dissipation</td>
<td>$P_{tot}$</td>
<td>330</td>
<td>mV</td>
</tr>
<tr>
<td>$T_S \leq 71^\circ C$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$T_S \leq 115^\circ C$</td>
<td></td>
<td>250</td>
<td></td>
</tr>
<tr>
<td>Junction temperature</td>
<td>$T_J$</td>
<td>150</td>
<td>°C</td>
</tr>
<tr>
<td>Storage temperature</td>
<td>$T_{stg}$</td>
<td>-65 ... 150</td>
<td></td>
</tr>
</tbody>
</table>

Data model

Fonduer automatically parses the richly formatted data into the data model that:
- Preserves structure/semantics across modalities
- Unifies a diverse variety of formats and styles
- Serves as the formal representation in KBC
Fonduer: An example state-of-the-art system

Signals from different modalities can be useful to find the information.