Information Extraction

Prof. Sameer Singh

CS 295: STATISTICAL NLP
WINTER 2017

February 21, 2017
Outline

- What is Information Extraction
- Named Entity Recognition
- Homework 3
Outline

- What is Information Extraction
- Named Entity Recognition
- Homework 3
Making Sense of Text

Query

? 

Search (IR)

Massive Corpus of Unstructured Text

Information Extraction

Structured Representation

Database or Graph

DB Query

Documents

Documents

Documents

Documents

Documents

Documents
News Articles

Query

Which AI startups have been acquired by Tech companies?

Structured Representation

Massive Corpus of News Articles
Fiction

Query

Which two characters are not related by blood?

Structured Representation

Collection of Books

Information Extraction
What is the interaction pathway between YY1 and TIP60?
Applications

Information Extraction → Database or Graph → Question Answering → Visualization & Statistics

Documents

Downstream AI applications
Low-level Info. Extraction
Slightly better...

who is playing in this year's super bowl

About 15,300,000 results (0.66 seconds)

Super Bowl LI
Super Bowl
Sunday, February 5, 6:30 PM on FOX
NRG Stadium, Houston, Texas

New England
Patriots

Atlanta
Falcons

Tickets - Preview

All times are in Eastern Time
The headquarters of BHP Billiton Limited, and the global headquarters of the combined BHP Billiton Group, are located in Melbourne, Australia.
In the industry...

Google Knowledge Graph
  ◦ Google Knowledge Vault

Amazon Product Graph

Facebook Graph API

IBM Watson

Microsoft Satori
  ◦ Project Hanover/Literome

LinkedIn Knowledge Graph

Yandex Object Answer

Diffbot, GraphIQ, Maana, ParseHub, Reactor Labs, SpazioDati
John was born in Liverpool, to Julia and Alfred Lennon.
Role of NLP?

John was born in Liverpool, to Julia and Alfred Lennon.
John was born in Liverpool, to Julia and Alfred Lennon.
John was born in Liverpool, to Julia and Alfred Lennon.
Outline

- What is Information Extraction
- Named Entity Recognition
- Homework 3
- Relation Extraction
An important sub-task: find and classify names in text, for example:

- The decision by the independent MP Andrew Wilkie to withdraw his support for the minority Labor government sounded dramatic but it should not further threaten its stability. When, after the 2010 election, Wilkie, Rob Oakeshott, Tony Windsor and the Greens agreed to support Labor, they gave just two guarantees: confidence and supply.
An important sub-task: **find** and **classify** names in text, for example:

- The decision by the independent MP **Andrew Wilkie** to withdraw his support for the minority **Labor** government sounded dramatic but it should not further threaten its stability. When, after the **2010** election, **Wilkie, Rob Oakeshott, Tony Windsor** and the **Greens** agreed to support **Labor**, they gave just two guarantees: confidence and supply.
An important sub-task: find and classify names in text, for example:

- The decision by the independent MP Andrew Wilkie to withdraw his support for the minority Labor government sounded dramatic but it should not further threaten its stability. When, after the 2010 election, Wilkie, Rob Oakeshott, Tony Windsor and the Greens agreed to support Labor, they gave just two guarantees: confidence and supply.
Detecting Named Entities

How it is done:
• Context is important!
  • Georgia, Washington, ...
  • John Deere, Thomas Cook, ...
  • Princeton, Amazon, ...
• Label whole sentence together
  • Structured prediction again

Uses in Knowledge Extraction:
• Mentions describes the nodes
• Types are incredibly important!
  • Often restrict relations
• Fine-grained types are informative!
  • Brooklyn: city
  • Sanders: politician, senator

John was born in Liverpool, to Julia and Alfred Lennon.
# NER: Entity Types

### Stanford CoreNLP

<table>
<thead>
<tr>
<th>Class</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Location</td>
<td>Buildings, airports, highways, bridges, etc.</td>
</tr>
<tr>
<td>Person</td>
<td>People, including fictional.</td>
</tr>
<tr>
<td>Organization</td>
<td>Companies, agencies, institutions, etc.</td>
</tr>
<tr>
<td>Misc</td>
<td>Nationalities or religious or political groups.</td>
</tr>
</tbody>
</table>

### spaCy.io

<table>
<thead>
<tr>
<th>Class</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>PERSON</td>
<td>People, including fictional.</td>
</tr>
<tr>
<td>NORP</td>
<td>Nationalities or religious or political groups.</td>
</tr>
<tr>
<td>FACILITY</td>
<td>Buildings, airports, highways, bridges, etc.</td>
</tr>
<tr>
<td>ORG</td>
<td>Companies, agencies, institutions, etc.</td>
</tr>
<tr>
<td>GPE</td>
<td>Countries, cities, states.</td>
</tr>
<tr>
<td>LOC</td>
<td>Non-GPE locations, mountain ranges, bodies of water.</td>
</tr>
<tr>
<td>PRODUCT</td>
<td>Objects, vehicles, foods, etc. (Not services.)</td>
</tr>
<tr>
<td>EVENT</td>
<td>Named hurricanes, battles, wars, sports events, etc.</td>
</tr>
<tr>
<td>WORK_OF_ART</td>
<td>Titles of books, songs, etc.</td>
</tr>
<tr>
<td>LANGUAGE</td>
<td>Any named language.</td>
</tr>
</tbody>
</table>
# NER: Entity Types

## Fine-grained Types

<table>
<thead>
<tr>
<th><strong>person</strong></th>
<th>doctor</th>
<th>engineer</th>
<th>monarch</th>
<th>musician</th>
<th>politician</th>
<th>religious_leader</th>
<th>soldier</th>
<th>terrorist</th>
</tr>
</thead>
<tbody>
<tr>
<td>actor</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>architect</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>artist</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>athlete</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>author</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>coach</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>director</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>organization</strong></th>
<th>airline</th>
<th>company</th>
<th>educational_institution</th>
<th>fraternity_sorority</th>
<th>sports_league</th>
<th>sports_team</th>
</tr>
</thead>
<tbody>
<tr>
<td>terrorist_organization</td>
<td>government_agency</td>
<td>government</td>
<td>political_party</td>
<td>educational_department</td>
<td>military</td>
<td>news_agency</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>location</strong></th>
<th>city</th>
<th>country</th>
<th>county</th>
<th>province</th>
<th>railway</th>
<th>road</th>
<th>bridge</th>
</tr>
</thead>
<tbody>
<tr>
<td>body_of_water</td>
<td>island</td>
<td>mountain</td>
<td>glacier</td>
<td>astral_body</td>
<td>cemetery</td>
<td>park</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>product</strong></th>
<th>camera</th>
<th>mobile_phone</th>
<th>computer</th>
<th>software</th>
<th>game</th>
<th>instrument</th>
<th>weapon</th>
</tr>
</thead>
<tbody>
<tr>
<td>engine</td>
<td>airplane</td>
<td>car</td>
<td>ship</td>
<td>spacecraft</td>
<td>train</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>art</strong></th>
<th>written_work</th>
<th>film</th>
<th>newspaper</th>
<th>play</th>
<th>music</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th><strong>event</strong></th>
<th>military_conflict</th>
<th>attack</th>
<th>natural_disaster</th>
<th>election</th>
<th>sports_event</th>
<th>protest</th>
<th>terrorist_attack</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th><strong>building</strong></th>
<th>time</th>
<th>color</th>
<th>award</th>
<th>educational_degree</th>
<th>title</th>
<th>law</th>
<th>ethnicity</th>
<th>language</th>
<th>religion</th>
<th>god</th>
<th>chemical_thing</th>
<th>biological_thing</th>
<th>medical_treatment</th>
<th>disease</th>
<th>symptom</th>
<th>drug</th>
<th>body_part</th>
<th>living_thing</th>
<th>animal</th>
<th>food</th>
<th>website</th>
<th>broadcast_network</th>
<th>broadcast_program</th>
<th>tv_channel</th>
<th>currency</th>
<th>stock_exchange</th>
<th>algorithm</th>
<th>programming_language</th>
<th>transit_system</th>
<th>transit_line</th>
</tr>
</thead>
</table>
Animals with Misleading Names

Electric Eel
- Not an eel.

Mountain Goat
- Not a goat.

Maned Wolf
- Not a wolf.

King Cobra
- Not a cobra. Also, snakes are typically self-governing.

Peacock Mantis Shrimp
- Not a peacock.
- Not a mantis.
- Also, not a shrimp.

Horny Toad
- Not a toad.
- Only thinks of you as a friend.

Mayfly
- Active through the spring and summer.

Eastern Kingbird
- Found in the West.
- Many birds do not recognise its authority.
Sequence Labeling for NER

\[ \arg\max_y \theta \cdot \prod_{i=1}^{n} \phi(x_i, y_i, y_{i-1}) \]
Features: Words and Lexicons

Words

Lexicons

\[ \text{WORD} = "John" \rightarrow k \]

\[ \text{LWORD} = "john" \rightarrow k' \]

People names ← US Census
"Appears in FName.100" \[ \rightarrow k \]

movies, places, ...
Features: Prefixes/Suffixes

John

prefix 1 = "J"
prefix 2 = "Jo"
prefix 4 = "John"
suffix 1 = "n"
Features: Substrings of Words

Cotrimoxazole

Wethersfield

Alien Fury: Countdown to Invasion
Features: Word Shapes

\[\text{Shape}(c) = \begin{cases} 
  X & \text{if A-Z} \\
  x & \text{if a-z} \\
  d & \text{if 0-9} \\
  c & \text{o.w.} 
\end{cases}\]

John       DC-100     CamelCase

Word shapes Xxxx       XX-ddd     XxxxxXxxxx

Short shapes Xx         X-d        XxxXx
Features: Surrounding Context

John

\(i-1\)

Deere

\(i\)

announced

\(i+1\)

NEXT_BIAS
NEXT_WORD=Deere
NEXT_LWORD=deere
NEXT_FIRSTCAP=True
NEXT_SSHAPE=Xx
NEXT_LEXICON=company
...

BIAS
WORD=Deere
LWORD=deere
FIRSTCAP=True
SSHAPE=Xx
LEXICON=company
...

PREV_BIAS
PREV_WORD=Deere
PREV_LWORD=deere
PREV_FIRSTCAP=True
PREV_SSHAPE=Xx
PREV_LEXICON=company
...

John Deere announced
Outline

What is Information Extraction

Named Entity Recognition

Homework 3

Relation Extraction
Sequence Tagging on Twitter

**Parts of Speech**

What a productive day. Not.


**Named Entity Recognition**

‘Breaking Dawn’ Returns to Vancouver on January 11th

O  B-movie  I-movie  O  O  O  B-ggeo-loc  O  O  O
Sequence Tagging Models

Logistic Regression

Conditional Random Fields

Steedman, 2000
What do you have to do?

- Feature Engineering
- Viterbi Algorithm

Test data will be released very close to the deadline!
Upcoming...

Homework
- Homework 3 is due on February 27
- Write-up and data has been released.

Project
- Status report due in 1.5 weeks: March 2, 2017
- Instructions coming soon
- Only 5 pages

Summaries
- Paper summaries: February 28, March 14
- Only 1 page each