Knowledge Graph Construction from Text

AAAI 2017

Jay Pujara, Sameer Singh, Bhavana Dalvi
Tutorial Overview

Part 1: Knowledge Graphs

Part 2: Knowledge Extraction

Part 3: Graph Construction

Part 4: Critical Analysis
Tutorial Outline

1. Knowledge Graph Primer [Jay]

2. Knowledge Extraction from Text
   a. NLP Fundamentals [Sameer]
   b. Information Extraction [Bhavana]

Coffee Break

3. Knowledge Graph Construction
   a. Probabilistic Models [Jay]
   b. Embedding Techniques [Sameer]

4. Critical Overview and Conclusion [Bhavana]
What is Knowledge Extraction?

Text

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

Literal Facts

Liverpool

birthplace

John Lennon

childOf

Alfred Lennon

childOf

Julia Lennon
NLP
Fundamentals
EXTRACTING STRUCTURES FROM LANGUAGE
What is NLP?

Unstructured
Ambiguous
Lots and lots of it!

Humans can read them, but
... very slowly
... can’t remember all
... can’t answer questions

NLP

“Knowledge”

Structured
Precise, Actionable
Specific to the task

Can be used for downstream applications, such as creating Knowledge Graphs!
What is NLP?

John was born in Liverpool, to Julia and Alfred Lennon.
What is Information Extraction?

John was born in Liverpool, to Julia and Alfred Lennon.
John was born in Liverpool, to Julia and Alfred Lennon.
Tokenization & Sentence Splitting

“Mr. Bob Dobolina is thinkin' of a master plan. Why doesn't he quit?”

[Mr.] [Bob] [Dobolina] [is] [thinkin’] [of] [a] [master] [plan] [.] [Why] [doesn't] [he] [quit] [?]

How it is done:
• Regular expressions, but not trivial
  • Mr., Yahoo!, lower-case
• For non-English, incredibly difficult!
  • Chinese: no “space” character
• Non-trivial for some domains...
  • What is a “token” in BioNLP?

Uses in KG Construction:
• Strictly constrains other NLP tasks
  • Parts of Speech
  • Dependency Parsing
• Directly effects KG nodes/edges
  • Mention boundaries
  • Relations within sentences
Tagging the Parts of Speech

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

How it is done:
• Context is important!
  • run, table, bar, ...
• Label whole sentence together
  • “Structured prediction”
• Conditional Random Fields, ..
• Now: CNNs, LSTMs, ...

Uses in KG Construction:
• Entities appear as nouns
• Verbs are very useful
  • For identifying relations
  • For identifying entity types
• Important for downstream NLP
  • NER, Dependency Parsing, ...

NNP VBD VBD IN NNP TO NNP CC NNP NNP

Uses in KG Construction:
• Entities appear as nouns
• Verbs are very useful
  • For identifying relations
  • For identifying entity types
• Important for downstream NLP
  • NER, Dependency Parsing, ...

11
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 KG Construction:
- Mentions describes the nodes
- Types are incredibly important!
  - Often restrict relations
- Fine-grained types are informative!
  - Brooklyn: city
  - Sanders: politician, senator
# NER: Entity Types

**Stanford CoreNLP**

<table>
<thead>
<tr>
<th>Class</th>
<th>Entities</th>
</tr>
</thead>
<tbody>
<tr>
<td>3 class:</td>
<td>Location, Person, Organization</td>
</tr>
<tr>
<td>4 class:</td>
<td>Location, Person, Organization, Misc</td>
</tr>
<tr>
<td>7 class:</td>
<td>Location, Person, Organization, Money, Percent, Date, Time</td>
</tr>
</tbody>
</table>

**spaCy.io**

<table>
<thead>
<tr>
<th>Category</th>
<th>Description</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>

From Stanford CoreNLP (http://nlp.stanford.edu/software/CRF-NER.shtml)
NER: Entity Types

<table>
<thead>
<tr>
<th>person</th>
<th>organization</th>
<th>terrorist_organization</th>
</tr>
</thead>
<tbody>
<tr>
<td>actor</td>
<td>airline</td>
<td>government_agency</td>
</tr>
<tr>
<td>architect</td>
<td>company</td>
<td>government</td>
</tr>
<tr>
<td>artist</td>
<td>educational_institution</td>
<td>political_party</td>
</tr>
<tr>
<td>athlete</td>
<td>monarch</td>
<td>educational_department</td>
</tr>
<tr>
<td>author</td>
<td>musician</td>
<td>military</td>
</tr>
<tr>
<td>coach</td>
<td>politician</td>
<td>news_agency</td>
</tr>
<tr>
<td>director</td>
<td>religious_leader</td>
<td></td>
</tr>
<tr>
<td></td>
<td>soldier</td>
<td></td>
</tr>
<tr>
<td></td>
<td>terrorist</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>location</th>
<th>product</th>
<th>event</th>
</tr>
</thead>
<tbody>
<tr>
<td>city</td>
<td>engine</td>
<td>military_conflict</td>
</tr>
<tr>
<td>country</td>
<td>airplane</td>
<td>attack</td>
</tr>
<tr>
<td>county</td>
<td>car</td>
<td>natural_disaster</td>
</tr>
<tr>
<td>province</td>
<td>ship</td>
<td>election</td>
</tr>
<tr>
<td>railway</td>
<td>spacecraft</td>
<td>sports_event</td>
</tr>
<tr>
<td>road</td>
<td>train</td>
<td>protest</td>
</tr>
<tr>
<td>bridge</td>
<td></td>
<td>terrorist_attack</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>building</th>
<th>time</th>
<th>chemical_thing</th>
</tr>
</thead>
<tbody>
<tr>
<td>airport</td>
<td>color</td>
<td>biological_thing</td>
</tr>
<tr>
<td>dam</td>
<td>award</td>
<td>medical_treatment</td>
</tr>
<tr>
<td>hospital</td>
<td>educational_degree</td>
<td>disease</td>
</tr>
<tr>
<td>hotel</td>
<td>title</td>
<td>symptom</td>
</tr>
<tr>
<td>library</td>
<td>law</td>
<td>drug</td>
</tr>
<tr>
<td>power_station</td>
<td>ethnicity</td>
<td>body_part</td>
</tr>
<tr>
<td>restaurant</td>
<td>language</td>
<td>living_thing</td>
</tr>
<tr>
<td>sports_facility</td>
<td>religion</td>
<td>animal</td>
</tr>
<tr>
<td>theater</td>
<td>god</td>
<td>food</td>
</tr>
</tbody>
</table>

|                         | camera                               | website                              |
|                         | mobile_phone                         | broadcast_network                    |
|                         | computer                             | broadcast_program                    |
|                         | software                             | tv_channel                           |
|                         | game                                 | currency                             |
|                         | instrument                           | stock_exchange                       |
|                         | weapon                               | algorithm                            |
|                         |                                     | programming_language                 |
|                         |                                      | transit_system                       |
|                         |                                      | transit_line                         |
Dependency Parsing

How it is done:
- **Model**: score trees using features
  - **Lexical**: words, POS, ...
  - **Structure**: distance, ...
- **Prediction**: Search over trees
  - greedy, spanning tree, belief propagation, dynamic prog, ...

Uses in KG Construction:
- Incredibly useful for relations!
  - What verb is attached?
  - Relation to which mention?
- Incredibly useful for attributes!
  - Appositives: “X, the CEO, ...”
- Paths are used as surface relations
John was born in Liverpool, to Julia and Alfred Lennon.

Text Patterns

John, Liverpool
  “was born in”
John, Julia
  “was born in Liverpool, to”
John, Alfred Lennon
  “was born in Liverpool, to Julia and”

Dependency Paths

“was born in”
“was born to”
Within-document Coreference

He...

Lennon..

the Pool

John Lennon...

Mrs. Lennon..

.. his mother ..

Alfred

his father

he

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

How it is done:
  • Mo`del: score pairwise links
    • dep path, similarity, types, ...
    • “representative mention”
  • Prediction: Search over clusterings
    • greedy (left to right), ILP, belief propagation, MCMC, ...

Uses in KG Construction:
  • More context for each entity!
  • Many relations occur on pronouns
    • “He is married to her”
  • Coref can be used for types
    • Nominals: The president, ...
  • Difficult, so often ignored
John Lennon was born in Liverpool, to Julia and Alfred Lennon.
Surface Patterns

Combine tokens, dependency paths, and entity types to define rules.

Bill Gates, the CEO of Microsoft, said ...
Mr. Jobs, the brilliant and charming CEO of Apple Inc., said ...
... announced by Steve Jobs, the CEO of Apple.
... announced by Bill Gates, the director and CEO of Microsoft.
... mused Bill, a former CEO of Microsoft.
and many other possible instantiations...
Rule-Based Extraction

A diagram showing a relationship between Argument 1 and Argument 2, with nodes labeled as 'appos', 'det', 'nmod', 'case', and 'headOf'.

Use a collection of rules as the system itself

Variations:

Source:
- Manually specified
- Learned from Data

Multiple Rules:
- Attach priorities/precedence
- Attach probabilities (more later)

High precision: when it fires, it’s correct
- Easy to explain predictions
- Easy to fix mistakes

However...
- Only work when the rules fire

Poor recall: Do not generalize!
Supervised Extraction

Machine Learning: hopefully, generalizes the labels in the *right way*

Use all of **NLP as features**: words, POS, NER, dependencies, embeddings

However

Usually, a lot of labeled data is needed, which is expensive & time consuming. Requires a lot of **feature engineering**!

\[
P(\text{birthplace}) = 0.75
\]

John was born in **Liverpool**, to Julia and Alfred Lennon.
...during the late 60’s and early 70’s, Kevin Smith worked with several local...

...the term hip-hop is attributed to Lovebug Starski. What does it actually mean...

Like Back in 2008, the Lions drafted Kevin Smith, even though Smith was badly...

... backfield in the wake of Kevin Smith’s knee injury, and the addition of Haynesworth...

The filmmaker Kevin Smith returns to the role of Silent Bob...

Nothing could be more irrelevant to Kevin Smith’s audacious ‘’Dogma’’ than ticking off...

... The Physiological Basis of Politics,” by Kevin Smith, Douglas Oxley, Matthew Hibbing...
Entity Names: Two Main Problems

Entities with Same Name
- Same type of entities share names
  - Kevin Smith, John Smith,
  - Springfield, ...
- Things named after each other
  - Clinton, Washington, Paris,
  - Amazon, Princeton, Kingston, ...
- Partial Reference
  - First names of people, Location
  - instead of team name, Nick names

Different Names for Entities
- Nick Names
  - Bam Bam, Drumpf, ...
- Typos/Misspellings
  - Baarak, Barak, Barrack, ...
- Inconsistent References
  - MSFT, APPL, GOOG...
**Entity Linking Approach**

Washington drops 10 points after game with UCLA Bruins.

<table>
<thead>
<tr>
<th>Candidate Generation</th>
<th>LOC/ORG</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Entity Types</th>
<th>UWashington, Huskies</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Coreference</th>
<th>UCLA Bruins, USC Trojans</th>
</tr>
</thead>
</table>
John Lennon was born in Liverpool, to Julia and Alfred Lennon.