Relation Extraction

Prof. Sameer Singh

CS 295: STATISTICAL NLP
WINTER 2017

February 23, 2017
Outline

Introduction to Relation Extraction

Hand-written Patterns

Supervised Machine Learning

Semi and Unsupervised Learning
John was born in Liverpool, to Julia and Alfred Lennon.
Relation Extraction

Company report: “International Business Machines Corporation (IBM or the company) was incorporated in the State of New York on June 16, 1911, as the Computing-Tabulating-Recording Co. (C-T-R)…”

Extracted Complex Relation:

<table>
<thead>
<tr>
<th>Relation</th>
<th>IBM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Company-Founding</td>
<td>IBM</td>
</tr>
<tr>
<td>Location</td>
<td>New York</td>
</tr>
<tr>
<td>Date</td>
<td>June 16, 1911</td>
</tr>
<tr>
<td>Original-Name</td>
<td>Computing-Tabulating-Recording Co.</td>
</tr>
</tbody>
</table>

But we will focus on the simpler task of extracting relation **triples**

- Founding-year(IBM,1911)
- Founding-location(IBM,New York)
The Leland Stanford Junior University, commonly referred to as Stanford University or Stanford, is an American private research university located in Stanford, California... near Palo Alto, California... Leland Stanford... founded the university in 1891.
News Domain

**ROLE:** relates a person to an organization or a geopolitical entity
- subtypes: member, owner, affiliate, client, citizen

**PART:** generalized containment
- subtypes: subsidiary, physical part-of, set membership

**AT:** permanent and transient locations
- subtypes: located, based-in, residence

**SOCIAL:** social relations among persons
- subtypes: parent, sibling, spouse, grandparent, associate
Automated Content Extraction

- PERSON-SOCIAL: Family, Business, Lasting Personal, Located
- PHYSICAL: Near, Located
- GENERAL AFFILIATION: Citizen-Resident-Ethnicity-Religion, Org-Location-Origin
- PART-WHOLE: Subsidiary, Geographical
- ORG AFFILIATION: Founder, Ownership, Membership, Investor, Student-Alum, Employment, Sports-Affiliation
- ARTIFACT: User-Owner-Inventor-Manufacturer

CS 295: STATISTICAL NLP (WINTER 2017)
ACE Relations Examples

Physical-Located PER-GPE
He was in Tennessee

Part-Whole-Subsidiary ORG-ORG
XYZ, the parent company of ABC

Person-Social-Family PER-PER
John’s wife Yoko

Org-AFF-Founder PER-ORG
Steve Jobs, co-founder of Apple...
Geographical Relations

- Geographical Entity (GE)
  - Natural GE
    - mountain
    - river
    - Zugspitze
    - Neckar
    - height (m) 2962
    - length (km) 367
  - country
    - Germany
    - flow_through Stuttgart
    - located_in Berlin
  - Inhabited GE
    - city
    - capital_of Germany
    - flow_through Neckar
    - located_in Stuttgart

- Flow Through Relations:
  - is-a
  - flow_through
## Medical Relations

### UMLS Resource

<table>
<thead>
<tr>
<th>Injury</th>
<th>disrupts</th>
<th>Physiological Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bodily Location</td>
<td>location-of</td>
<td>Biologic Function</td>
</tr>
<tr>
<td>Anatomical Structure</td>
<td>part-of</td>
<td>Organism</td>
</tr>
<tr>
<td>Pharmacologic Substance</td>
<td>causes</td>
<td>Pathological Function</td>
</tr>
<tr>
<td>Pharmacologic Substance</td>
<td>treats</td>
<td>Pathologic Function</td>
</tr>
</tbody>
</table>
Medical Relations

Doppler echocardiography can be used to diagnose left anterior descending artery stenosis in patients with type 2 diabetes

\[
\text{Echocardiography, Doppler DIAGNOSES Acquired stenosis}
\]
Freebase Relations

Thousands of relations and millions of instances!
Manually created from multiple sources including Wikipedia InfoBoxes

<table>
<thead>
<tr>
<th>Relation name</th>
<th>Size</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>/people/person/nationality</td>
<td>281,107</td>
<td>John Dugard, South Africa</td>
</tr>
<tr>
<td>/location/location/contains</td>
<td>253,223</td>
<td>Belgium, Nijlen</td>
</tr>
<tr>
<td>/people/person/profession</td>
<td>208,888</td>
<td>Dusa McDuff, Mathematician</td>
</tr>
<tr>
<td>/people/person/place_of_birth</td>
<td>105,799</td>
<td>Edwin Hubble, Marshfield</td>
</tr>
<tr>
<td>/dining/restaurant/cuisine</td>
<td>86,213</td>
<td>MacAyo’s Mexican Kitchen, Mexican</td>
</tr>
<tr>
<td>/business/business_chain/location</td>
<td>66,529</td>
<td>Apple Inc., Apple Inc., South Park, NC</td>
</tr>
<tr>
<td>/biology/organism_classification_rank</td>
<td>42,806</td>
<td>Scorpaeniformes, Order</td>
</tr>
<tr>
<td>/film/film/genre</td>
<td>40,658</td>
<td>Where the Sidewalk Ends, Film noir</td>
</tr>
<tr>
<td>/film/film/language</td>
<td>31,103</td>
<td>Enter the Phoenix, Cantonese</td>
</tr>
<tr>
<td>/biology/organism-higher_classification</td>
<td>30,052</td>
<td>Calopteryx, Calopterygidae</td>
</tr>
<tr>
<td>/film/film/country</td>
<td>27,217</td>
<td>Turtle Diary, United States</td>
</tr>
<tr>
<td>/film/writer/film</td>
<td>23,856</td>
<td>Irving Shulman, Rebel Without a Cause</td>
</tr>
<tr>
<td>/film/director/film</td>
<td>23,539</td>
<td>Michael Mann, Collateral</td>
</tr>
<tr>
<td>/film/producer/film</td>
<td>22,079</td>
<td>Diane Eskenazi, Aladdin</td>
</tr>
<tr>
<td>/people/deceased_person/place_of_death</td>
<td>18,814</td>
<td>John W. Kern, Asheville</td>
</tr>
<tr>
<td>/music/artist/origin</td>
<td>18,619</td>
<td>The Octopus Project, Austin</td>
</tr>
<tr>
<td>/people/person/religion</td>
<td>17,582</td>
<td>Joseph Chartrand, Catholicism</td>
</tr>
<tr>
<td>/book/author/works_written</td>
<td>17,278</td>
<td>Paul Auster, Travels in the Scriptorium</td>
</tr>
<tr>
<td>/soccer/football_position/players</td>
<td>17,244</td>
<td>Midfielder, Chen Tao</td>
</tr>
<tr>
<td>/people/deceased_person/cause_of_death</td>
<td>16,709</td>
<td>Richard Daintree, Tuberculosis</td>
</tr>
<tr>
<td>/film/film/music</td>
<td>14,070</td>
<td>Stavisky, Stephen Sondheim</td>
</tr>
<tr>
<td>/business/company/industry</td>
<td>13,805</td>
<td>ATS Medical, Health care</td>
</tr>
</tbody>
</table>
Ontological Relations

IS-A (hypernym): subsumption between classes
- Giraffe IS-A ruminant IS-A ungulate IS-A mammal IS-A vertebrate IS-A animal...

Instance-of: relation between individual and class
- San Francisco instance-of city
Outline

- Introduction to Relation Extraction
- Hand-written Patterns
- Supervised Machine Learning
- Semi and Unsupervised Learning
Rules for IS-A Relation

Early intuition from Hearst (1992)

“Agar is a substance prepared from a mixture of red algae, such as Gelidium, for laboratory or industrial use”

What does Gelidium mean?

How do you know?
Hearst’s Patterns for IS-A relations

“\( Y \) such as \( X ((, X)^* (, \text{ and/or} X) \)”

“such \( Y \) as \( X \)”

“\( X \) or other \( Y \)”

“\( X \) and other \( Y \)”

“\( Y \) including \( X \)”

“\( Y \), especially \( X \)”
# Hearst’s Patterns for IS-A relations

<table>
<thead>
<tr>
<th>Hearst pattern</th>
<th>Example occurrences</th>
</tr>
</thead>
<tbody>
<tr>
<td>X and other Y</td>
<td>...temples, treasuries, and other important civic buildings.</td>
</tr>
<tr>
<td>X or other Y</td>
<td>Bruises, wounds, broken bones or other injuries...</td>
</tr>
<tr>
<td>Y such as X</td>
<td>The bow lute, such as the Bambara ndang...</td>
</tr>
<tr>
<td>Such Y as X</td>
<td>...such authors as Herrick, Goldsmith, and Shakespeare.</td>
</tr>
<tr>
<td>Y including X</td>
<td>...common-law countries, including Canada and England...</td>
</tr>
<tr>
<td>Y, especially X</td>
<td>European countries, especially France, England, and Spain...</td>
</tr>
</tbody>
</table>
Extracting Richer Relations

Intuition:

Relations often hold between specific types of entities

- located-in (ORGANIZATION, LOCATION)
- founded (PERSON, ORGANIZATION)
- cures (DRUG, DISEASE)

Start with Named Entity tags to extract relation!
Entity Types aren’t enough

Which relations hold between 2 entities?

Drug  Cure?  Prevent?  Cause?  Disease
Which relations hold between two entities?

PERSON

Founder?
Investor?
Member?
Employee?
President?

ORGANIZATION
Extracting Richer Relations Using Rules and Named Entities

Who holds what office in what organization?

**PERSON**, **POSITION** of **ORG**

- George Marshall, Secretary of State of the United States

**PERSON**(named|appointed|chose|etc.) **PERSON** Prep? **POSITION**

- Truman appointed Marshall Secretary of State

**PERSON** [be]? (named|appointed|etc.) Prep? **ORG** **POSITION**

- George Marshall was named US Secretary of State
Complex 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

Use a collection of rules as the system itself

Source:
- Manually specified
- Learned from Data

Multiple Rules:
- Attach priorities/precedence
- Attach probabilities (more later)
Hand-built patterns for relations

Pluses
- Human patterns tend to be high-precision
- Can be tailored to specific domains
- Easy to debug: why a prediction was made, how to fix?

Minuses
- Human patterns are often low-recall
- A lot of work to think of all possible patterns!
- Don’t want to have to do this for every relation!
- We’d like better accuracy (*generalization*)
Outline

- Introduction to Relation Extraction
- Hand-written Patterns
- Supervised Machine Learning
- Semi and Unsupervised Learning
Supervised Machine Learning

Choose a set of relations we’d like to extract

Choose a set of relevant named entities

Find and label data

- Choose a representative corpus
- Label the named entities in the corpus
- Hand-label the relations between these entities
- Break into training, development, and test

Train a classifier on the training set
Automated Content Extraction

- PERSON-SOCIAL
  - Family
  - Lasting Personal
  - Business

- PHYSICAL
  - Near
  - Located

- GENERAL AFFILIATION
  - Citizen-Resident-Ethnicity-Religion
  - Org-Location-Origin

- PART-WHOLE
  - Subsidiary
  - Geographical

- ORG AFFILIATION
  - Founder
  - Ownership
  - Membership
  - Investor
  - Student-Alum
  - Employment
  - Sports-Affiliation

- ARTIFACT
  - User-Owner-Inventor-Manufacturer

ACE 2008 “Relation Extraction Task”
CS 295: STATISTICAL NLP (WINTER 2017)
Relation Extraction

Classify the relation between two entities in a sentence

American Airlines, a unit of AMR, immediately matched the move, spokesman Tim Wagner said.
Word Features for Relation Extraction

**American Airlines**, a unit of AMR, immediately matched the move, spokesman **Tim Wagner** said

Mention 1

Headwords of M1 and M2, and combination

- Airlines
- Wagner
- Airlines-Wagner

Bag of words and bigrams in M1 and M2

- {American, Airlines, Tim, Wagner, American Airlines, Tim Wagner}

Words or bigrams in particular positions left and right of M1/M2

- **M2: -1 spokesman**
- **M2: +1 said**

Bag of words or bigrams between the two entities

- {a, AMR, of, immediately, matched, move, spokesman, the, unit}
Named Entity Type and Mention Level Features

American Airlines, a unit of AMR, immediately matched the move, spokesman Tim Wagner said

Mention 1

Mention 2

Named-entity types
- M1: ORG
- M2: PERSON

Concatenation of the two named-entity types
- ORG-PERSON

Entity Level of M1 and M2 (NAME, NOMINAL, PRONOUN)
- M1: NAME [it or he would be PRONOUN]
- M2: NAME [the company would be NOMINAL]
Dependency Parse Features for Relation Extraction

**American Airlines**, a unit of AMR, immediately matched the move, spokesman **Tim Wagner** said

Mention 1

Mention 2

Base syntactic chunk sequence from one to the other

NP  NP  PP  VP  NP  NP

Constituent path through the tree from one to the other

NP  ↑  NP  ↑  S  ↑  S  ↓  NP

Dependency path

Airlines  matched  Wagner  said
Gazetteer and Trigger word features for relation extraction

Trigger list for family: kinship terms
- parent, wife, husband, grandparent, etc. [from WordNet]

Gazetteer:
- Lists of useful geo or geopolitical words
  - Country name list
  - Other sub-entities
American Airlines, a unit of AMR, immediately matched the move, spokesman Tim Wagner said.

Entity-based features
- Entity\textsubscript{1} type: ORG
- Entity\textsubscript{1} head: airlines
- Entity\textsubscript{2} type: PERS
- Entity\textsubscript{2} head: Wagner
- Concatenated types: ORGPERS

Word-based features
- Between-entity bag of words: \{ a, unit, of, AMR, Inc., immediately, matched, the, move, spokesman \}
- Word(s) before Entity\textsubscript{1}: NONE
- Word(s) after Entity\textsubscript{2}: said

Syntactic features
- Constituent path: $NP \rhd NP \rhd S \rhd S \rhd NP$
- Base syntactic chunk path: $NP \rightarrow NP \rightarrow PP \rightarrow NP \rightarrow VP \rightarrow NP \rightarrow NP$
- Typed-dependency path: $Airlines \leftarrow_{subj} matched \leftarrow_{comp} said \rightarrow_{subj} Wagner$
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.
Supervised Relation Extraction

Pluses

• Can get high accuracies if enough training data
• If test similar enough to training
• Can utilize a number of NLP tasks

Minuses

• Labeling a large training set is expensive
• Supervised models are brittle, don’t generalize well to different genres
Outline

Introduction to Relation Extraction

Hand-written Patterns

Supervised Machine Learning

Semi and Unsupervised Learning
Seed-based or bootstrapping approaches to relation extraction

No training set? Maybe you have:
- A few seed tuples or
- A few high-precision patterns

Can you use those seeds to do something useful?
- **Bootstrapping**: use the seeds to directly learn a relation
Relation Bootstrapping

Gather a set of seed pairs that have the relation

1. Find sentences with these pairs
2. Look at the context between or around the pair and generalize the context to create patterns
3. Use the patterns to gather more pairs
4. Repeat
Bootstrapping Example

<Mark Twain, Elmira> Seed tuple od “died in”

Look for the environments of the seed tuple

“Mark Twain is buried in Elmira, NY.”
X is buried in Y

“The grave of Mark Twain is in Elmira”
The grave of X is in Y

“Elmira is Mark Twain’s final resting place”
Y is X’s final resting place.

Use those patterns to find new tuples

Repeat
**Dipre: Extract <author,book> pairs**

Start with 5 seeds:

<table>
<thead>
<tr>
<th>Author</th>
<th>Book</th>
</tr>
</thead>
<tbody>
<tr>
<td>Isaac Asimov</td>
<td>The Robots of Dawn</td>
</tr>
<tr>
<td>David Brin</td>
<td>Startide Rising</td>
</tr>
<tr>
<td>James Gleick</td>
<td>Chaos: Making a New Science</td>
</tr>
<tr>
<td>Charles Dickens</td>
<td>Great Expectations</td>
</tr>
<tr>
<td>William Shakespeare</td>
<td>The Comedy of Errors</td>
</tr>
</tbody>
</table>

Find Instances on the Web:

- The Comedy of Errors, by William Shakespeare, was
- The Comedy of Errors, by William Shakespeare, is
- The Comedy of Errors, one of William Shakespeare’s earliest attempts
- The Comedy of Errors, one of William Shakespeare’s most

Extract patterns (group by middle, take longest common prefix/suffix)

- \(?x\), by \(?y\),
- \(?x\), one of \(?y\)’s

Now iterate, finding new seeds that match the pattern
Snowball

Similar iterative algorithm

Group instances w/similar prefix, middle, suffix, extract patterns

- But require that X and Y be named entities
- And compute a confidence for each pattern

<table>
<thead>
<tr>
<th>Organization</th>
<th>Location of Headquarters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Microsoft</td>
<td>Redmond</td>
</tr>
<tr>
<td>Exxon</td>
<td>Irving</td>
</tr>
<tr>
<td>IBM</td>
<td>Armonk</td>
</tr>
</tbody>
</table>

.69 Organization {‘s, in, headquarters} Location

.75 Location {in, based} Organization
Distant Supervision

Combine bootstrapping with supervised learning
- Instead of 5 (or just a few) seeds,
  - Use a large database to get huge # of seed examples
  - Create lots of features from all these examples
  - Combine in a supervised classifier
Distantly Supervised learning of relation extraction patterns

1. For each relation

2. For each tuple in big database

3. Find sentences in large corpus with both entities

4. Extract frequent features (parse, words, etc)

5. Train supervised classifier using these patterns

Born-In

<Edwin Hubble, Marshfield>
<Albert Einstein, Ulm>

Hubble was born in Marshfield
Einstein, born (1879), Ulm
Hubble’s birthplace in Marshfield

PER was born in LOC
PER, born (XXXX), LOC
PER’s birthplace in LOC

P(born-in | f_1,f_2,f_3,\ldots,f_{70000})
Distant Supervision Paradigm

Like supervised classification:
- Uses a classifier with lots of features
- Supervised by detailed hand-created knowledge
- Doesn’t require iteratively expanding patterns

Like unsupervised classification:
- Uses very large amounts of unlabeled data
- Not sensitive to genre issues in training corpus
Unsupervised Relation Extraction

Open Information Extraction:
° extract relations from the web with no training data, no list of relations

1. Use parsed data to train a “trustworthy tuple” classifier
2. Single-pass extract all relations between NPs, keep if trustworthy
3. Assessor ranks relations based on text redundancy
   (FCI, specializes in, software development)
   (Tesla, invented, coil transformer)
   (Tesla, inventor of, transformer)
Evaluation of Semi-supervised and Unsupervised Relation Extraction

Since it extracts totally new relations from the web
  ◦ There is no gold set of correct instances of relations!
    ◦ Can’t compute precision (don’t know which ones are correct)
    ◦ Can’t compute recall (don’t know which ones were missed)

Instead, we can approximate precision (only)
  ◦ Draw a random sample of relations from output, check precision manually

\[ \hat{p} = \frac{\text{# of correctly extracted relations in the sample}}{\text{Total # of extracted relations in the sample}} \]

Can also compute precision at different levels of recall.
  ◦ Precision for top 1000 new relations, top 10,000 new relations, top 100,000
  ◦ In each case taking a random sample of that set

But no way to evaluate recall
Outline

Introduction to Relation Extraction

Hand-written Patterns

Supervised Machine Learning

Semi and Unsupervised Learning
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