Relation Extraction

What is relation extraction?

(Reading: J+M 22)
Extracting relations from text

- **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**

<table>
<thead>
<tr>
<th>Relation</th>
<th>IBM, 1911</th>
</tr>
</thead>
<tbody>
<tr>
<td>Founding-year</td>
<td></td>
</tr>
<tr>
<td>Founding-location</td>
<td></td>
</tr>
</tbody>
</table>
Extracting Relation Triples from Text

Stanford University

From Wikipedia, the free encyclopedia.

"Stanford" redirects here. For other uses, see Stanford (disambiguation).

Not to be confused with Stanford University (disambiguation).

The Leland Stanford Junior University, commonly referred to as Stanford University or Stanford, is an American private research university located in Stanford, California on an 8,180-acre (3,310 ha) campus near Palo Alto, California, United States. It is situated in the northwestern Santa Clara Valley on the San Francisco Peninsula, approximately 20 miles (32 km) northwest of San Jose and 37 miles (60 km) southeast of San Francisco.[6]

Leland Stanford, a Californian railroad tycoon and politician, founded the university in 1881 in honor of his son, Leland Stanford, Jr., who died of typhoid two months before his 18th birthday. The university was established as a coeducational and nondenominational institution, but struggled financially after the senior Stanford's 1893 death and after much of the campus was damaged by the 1906 San Francisco earthquake. Following World War II, Provost Frederick Terman supported faculty and graduates' entrepreneurialism to build self-sufficient local industry in what would become known as Silicon Valley. By 1970, Stanford was home to a linear accelerator, was one of the original four ARPANET nodes, and had transformed itself into a major research university in computer science, mathematics, natural sciences, and social sciences. More than 50 Stanford faculty, staff, and alumni have won the Nobel Prize and Stanford has the largest number of Turing award winners for a single institution. Stanford faculty and alumni have founded many prominent technology companies including Cisco Systems, Google, Hewlett-Packard, LinkedIn, Rambus, Silicon Graphics, Sun Microsystems, Varian Associates, and Yahoo.[7]

The university is organized into seven schools including academic schools of Humanities...
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.
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

Stanford EQ Leland Stanford Junior University
Stanford LOC–IN California
Stanford IS–A research university
Stanford LOC–NEAR Palo Alto
Stanford FOUNDED–IN 1891
Stanford FOUNDER Leland Stanford
Why Relation Extraction?

• Create new structured knowledge bases, useful for any app
• Augment current knowledge bases
  • Adding words to WordNet thesaurus, facts to FreeBase or DBPedia
Why Relation Extraction?

• Create new structured knowledge bases, useful for any app
• Augment current knowledge bases
  • Adding words to WordNet thesaurus, facts to FreeBase or DBPedia
• Support question answering
  • The granddaughter of which actor starred in the movie “E.T.”?
    (acted-in ?x “E.T.”) (is-a ?y actor) (granddaughter-of ?x ?y)
Why Relation Extraction?

- Create new structured knowledge bases, useful for any app
- Augment current knowledge bases
  - Adding words to WordNet thesaurus, facts to FreeBase or DBPedia
- Support question answering
  - The granddaughter of which actor starred in the movie “E.T.”?
    (acted-in ?x “E.T.”) (is-a ?y actor) (granddaughter-of ?x ?y)
- But which relations should we extract?
Automated Content Extraction (ACE)

17 relations from 2008 “Relation Extraction Task”
Automated Content Extraction (ACE)

- 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...
UMLS: Unified Medical Language System

- 134 entity types, 54 relations

<table>
<thead>
<tr>
<th>Entity Type</th>
<th>Relation</th>
<th>Function Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Injury</td>
<td>disrupts</td>
<td>Physiological Function</td>
</tr>
<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>
Doppler echocardiography can be used to diagnose left anterior descending artery stenosis in patients with type 2 diabetes.

Echocardiography, Doppler DIAGNOSES Acquired stenosis
Databases of Wikipedia Relations

Wikipedia Infobox

The Leland Stanford Junior University, commonly referred to as Stanford University or Stanford, is an American private research university located in Stanford, California on an 8,180-acre (3,310 ha) campus near Palo Alto, California, United States. It is situated in the northwestern Santa Clara Valley on the San Francisco Peninsula, approximately 20 miles (32 km) northwest of San Jose and 57 miles (90 km) southeast of San Francisco.

Leland Stanford, a Californian railroad tycoon and politician, founded the university in 1891 in honor of his son, Leland Stanford, Jr., who died of typhoid two months before his 16th birthday. The university was established as a coeducational and non-sectarian institution, but struggled financially after the senior Stanford's 1893 death and after much of the campus was damaged by the 1906 San Francisco earthquake. Following World War II, Provost Frederick Terman supported faculty and graduate entrepreneurship to build self-sufficient local industry in what would become known as Silicon Valley. By 1979, Stanford was home to a linear accelerator, was one of the original four ARPANET nodes, and had transformed itself into a major research university in computer science, mathematics, natural sciences, and social sciences. More than 50 Stanford faculty, staff, and alumni have won the Nobel Prize and Stanford has the largest number of Turing award winners for a single institution. Stanford faculty and alumni have founded many prominent technology companies including Cisco Systems, Google, Hewlett-Packard, LinkedIn, Rambus, Silicon Graphics, Sun Microsystems, Varian Associates, and Yahoo.

The university is organized into seven schools including academic schools of Humanities and Sciences and Earth Sciences as well as professional schools of Business, Education, Engineering, Law, and Medicine. Stanford has a student body of approximately 9,888 undergraduate and 8,400 graduate students. Stanford is a founding member of the Association of American Universities. For the 2011-2012 year, the university has a budget of US$4.1 billion, US$3.2 billion in research expenditures, and manages a US$16.5 billion endowment, with $35.1 billion in consolidated net assets.

Stanford competes in 34 varsity sports and is one of two private universities in the Division I FBS Pac-12 Conference. Stanford's athletic program has won the NCAA Directors' Cup every year since 1996. In the 2008 Summer Olympics in Beijing, Stanford athletes won 24 medals, including eight gold medals, more than any other university in the United States.

Contents
1 History
  1.1 Origins
Databases of Wikipedia Relations

Wikipedia Infobox

Stanford University

Motto: Die Luft der Freiheit weht (German)

Motto in English: The wind of freedom

Established: 1891

Type: Private

Endowment: US$16.5 billion (2011)

President: John L. Hennessy

Vice President: John V. Hennessy

Academic staff: 1,910

Students: 15,810

Undergraduates: 6,890

Postgraduates: 8,400

Location: Stanford, California, U.S.

Campus: Shriram, 8,180 acres (33.3 km²)

Stanford University

Leland Stanford Junior University

Motto: Die Luft der Freiheit weht (German)

Motto in English: The wind of freedom

Established: 1891

Type: Private

Endowment: US$16.5 billion (2011)

President: John L. Hennessy

Vice President: John V. Hennessy

Academic staff: 1,910

Students: 15,810

Undergraduates: 6,890

Postgraduates: 8,400

Location: Stanford, California, U.S.

Campus: Shriram, 8,180 acres (33.3 km²)
# Databases of Wikipedia Relations

<table>
<thead>
<tr>
<th>Type</th>
<th>Private</th>
</tr>
</thead>
<tbody>
<tr>
<td>Endowment</td>
<td>US$ 16.5 billion (2011)(^3)</td>
</tr>
<tr>
<td>President</td>
<td>John L. Hennessy</td>
</tr>
<tr>
<td>Provost</td>
<td>John Etchemendy</td>
</tr>
<tr>
<td>Academic staff</td>
<td>1,910(^4)</td>
</tr>
<tr>
<td>Students</td>
<td>15,319</td>
</tr>
<tr>
<td>Undergraduates</td>
<td>6,878(^5)</td>
</tr>
<tr>
<td>Postgraduates</td>
<td>8,441(^5)</td>
</tr>
<tr>
<td>Location</td>
<td>Stanford, California, U.S.</td>
</tr>
<tr>
<td>Campus</td>
<td>Suburban, 8,180 acres (3,310 ha)(^6)</td>
</tr>
<tr>
<td>Colors</td>
<td>Cardinal red and white</td>
</tr>
</tbody>
</table>
Databases of Wikipedia Relations

{{Infobox university
|image_name= Stanford University seal.svg
|image_size= 210px
|caption = Seal of Stanford University
|name = Stanford University
|native_name = Leland Stanford Junior University
|motto = {{lang|de|"Die Luft der Freiheit weht"}}<br /> ({{German language|German}})<ref
name="casper"><cite speech|title=Die Luft der Freiheit weht—On and Off|author=Gerhard Casper|first=Gerhard|last=Casper|authorlink=Gerhard Casper|date=1995-10-05|url=http://www.stanford.edu/dept/pres-provost/president/speeches/951005dieluft.html}}</ref>
mottoenç = The wind of freedom blows<ref name="casper" />
established = 1891<ref>{{cite web |
url=http://www.stanford.edu/home/stanford/history/begin.html | title=Stanford University History |
publisher = Stanford University | accessdate = 2007-04-26}}</ref>
type = [[private university|Private]]
calendar= Quarter
|president = [[John L. Hennessy]]
|provost = [[John Etchemendy]]
city = [[Stanford, California|Stanford]]
|state = California
country = U.S.
Databases of Wikipedia Relations

Relations extracted from Infobox
Stanford state California
Stanford motto “Die Luft der Freiheit weht”
...
Relation databases that draw from Wikipedia

- Resource Description Framework (RDF) triples
  subject predicate object
  Golden Gate Park location San Francisco
  dbpedia:Golden_Gate_Park dbpedia-owl:location dbpedia:San_Francisco

- DBPedia: 1 billion RDF triples, 385 from English Wikipedia

- Frequent Freebase relations:
  people/person/nationality, location/location/contains,
  people/person/profession, people/person/place-of-birth,
  film/film/genre,
Ontological relations

Examples from the WordNet Thesaurus

• **IS-A (hyponym):** 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
How to build relation extractors

1. Hand-written patterns
2. Supervised machine learning
3. Semi-supervised and unsupervised
   • Bootstrapping (using seeds)
   • Distant supervision
   • Unsupervised learning from the web
Relation Extraction

Using patterns to extract relations
Rules for extracting 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?`
Rules for extracting 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 extracting IS-A relations

(Hearst, 1992): Automatic Acquisition of Hyponyms

“Y such as X ((, X)* (, 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 extracting 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 Using Rules

- Intuition: relations often hold between specific entities
  - located-in (ORGANIZATION, LOCATION)
  - founded (PERSON, ORGANIZATION)
  - cures (DRUG, DISEASE)
- Start with Named Entity tags to help extract relation!
Named Entities aren’t quite enough. Which relations hold between 2 entities?

Drug

Cure?
Prevent?
Cause?

Disease
What relations hold between 2 entities?

PERSON

Founder?
Investor?
Member?
Employee?
President?

ORGANIZATION
Extracting Richer Relations Using Rules and Named Entities

Who holds what office in what 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
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|chosen|etc.) PERSON Prep? POSITION

- Truman appointed Marshall Secretary of State
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|chosen|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
Hand-built patterns for relations

• Plus:
  • Human patterns tend to be high-precision
  • Can be tailored to specific domains
Hand-built patterns for relations

• Plus:
  • Human patterns tend to be high-precision
  • Can be tailored to specific domains

• Minus
  • 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
Relation Extraction

Supervised relation extraction
Supervised machine learning for relations

- Choose a set of relations we’d like to extract
Supervised machine learning for relations

- Choose a set of relations we’d like to extract
- Choose a set of relevant named entities
Supervised machine learning for relations

- 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
Supervised machine learning for relations

- 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
How to do classification in supervised relation extraction

1. Find all pairs of named entities (usually in same sentence)
How to do classification in supervised relation extraction

1. Find all pairs of named entities (usually in same sentence)
2. Decide if 2 entities are related
How to do classification in supervised relation extraction

1. Find all pairs of named entities (usually in same sentence)
2. Decide if 2 entities are related
3. If yes, classify the relation
   • Why the extra step?
     • Faster classification training by eliminating most pairs
     • Can use distinct feature-sets appropriate for each task.
Automated Content Extraction (ACE)

17 sub-relations of 6 relations from 2008 “Relation Extraction Task”
American Airlines, a unit of AMR, immediately matched the move, spokesman Tim Wagner said.
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.
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.
American Airlines, a unit of AMR, immediately matched the move, spokesman Tim Wagner said.
American Airlines, a unit of AMR, immediately matched the move, spokesman Tim Wagner said.

- Headwords of M1 and M2, and combination

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

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

- 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
Word Features for Relation Extraction

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

- **Mention 1**
- **Mention 2**

- **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 for Relation Extraction

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

- Named-entity types
  - M1: ORG
  - M2: PERSON
American Airlines, a unit of AMR, immediately matched the move, spokesman Tim Wagner said.

- Named-entity types
  - M1: ORG
  - M2: PERSON
- Concatenation of the two named-entity types
  - ORG-PERSON
American Airlines, a unit of AMR, immediately matched the move, spokesman Tim Wagner said.

- 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]
Parse Features for Relation Extraction

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

- Base syntactic chunk sequence from one to the other
  
  NP  NP  PP  VP  NP  NP
American Airlines, a unit of AMR, immediately matched the move, spokesman Tim Wagner said.

- 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
Parse Features for Relation Extraction

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

- Base syntactic chunk sequence from one to the other
  \[\text{NP} \quad \text{NP} \quad \text{PP} \quad \text{VP} \quad \text{NP} \quad \text{NP}\]

- Constituent path through the tree from one to the other
  \[\text{NP} \quad \rightarrow \quad \text{NP} \quad \rightarrow \quad \text{S} \quad \rightarrow \quad \text{S} \quad \downarrow \quad \text{NP}\]

- Dependency path
  \[\text{Airlines} \quad \text{matched} \quad \text{Wagner} \quad \text{said}\]
Gazetteer and trigger word features for relation extraction

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

- Now you can use any classifier you like
  - MaxEnt
  - Naïve Bayes
  - SVM
  - ...

Classifiers for supervised methods

• Now you can use any classifier you like
  • MaxEnt
  • Naïve Bayes
  • SVM
  • ...

• Train it on the training set, tune on the dev set, test on the test set
Evaluation of Supervised Relation Extraction

- Compute P/R/F$_1$ for each relation

\[
P = \frac{\text{# of correctly extracted relations}}{\text{Total # of extracted relations}}
\]

\[
R = \frac{\text{# of correctly extracted relations}}{\text{Total # of gold relations}}
\]

\[
F_1 = \frac{2PR}{P + R}
\]
Summary: Supervised Relation Extraction

+ Can get high accuracies with enough hand-labeled training data, if test similar enough to training
- Labeling a large training set is expensive
- Supervised models are brittle, don’t generalize well to different genres
Relation Extraction

Semi-supervised and unsupervised relation extraction
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 to populate a relation
Relation Bootstrapping (Hearst 1992)

- Gather a set of seed pairs that have relation R
- Iterate:
  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 for grep for more pairs
Bootstrapping

• <Mark Twain, Elmira> **Seed tuple**
  • Grep (google) 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 grep for new tuples
  • Iterate
**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:**
  
  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

E. Agichtein and L. Gravano 2000. Snowball: Extracting Relations from Large Plain-Text Collections. ICDL

• 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

Snow, Jurafsky, Ng. 2005. Learning syntactic patterns for automatic hypernym discovery. NIPS 17
Fei Wu and Daniel S. Weld. 2007. Autonomously Semantifying Wikipeida. CIKM 2007
Mintz, Bills, Snow, Jurafsky. 2009. Distant supervision for relation extraction without labeled data. ACL09

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

1. For each relation Born-In
Distantly supervised learning of relation extraction patterns

1. For each relation

2. For each tuple in big database

Born-In

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

- Born-In
  - <Edwin Hubble, Marshfield>
  - <Albert Einstein, Ulm>

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

\[ P(\text{born-in} \mid f_1, f_2, f_3, \ldots, f_{70000}) \]
Unsupervised relation extraction

M. Banko, M. Cararella, S. Soderland, M. Broadhead, and O. Etzioni. 2007. Open information extraction from the web. IJCAI

• Open Information Extraction:
  • extract relations from the web with no training data, no list of relations
Unsupervised relation extraction

M. Banko, M. Cararella, S. Soderland, M. Broadhead, and O. Etzioni. 2007. Open information extraction from the web. IJCAI

• 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
Unsupervised relation extraction

M. Banko, M. Cararella, S. Soderland, M. Broadhead, and O. Etzioni. 2007. Open information extraction from the web. IJCAI

• 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)
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)
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
    \[ \hat{P} = \frac{\text{# of correctly extracted relations in the sample}}{\text{Total # of extracted relations in the sample}} \]
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, eg. 100, samples of that set
- But no way to evaluate recall