Learning Thesauruses and Knowledge Bases

Thesaurus induction and relation extraction
What is thesaurus induction?

Relation extraction

- Lexico-syntactic patterns (Hearst, 1992),
- LRA (Turney, 2005),
- Espresso (Pantel & Pennacchiotti, 2006),
- Distributional similarity…

A structured, consistent thesaurus of sense-disambiguated synsets

And hundreds of thousands more…
Thesaurus induction is a special case of relation extraction

- **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

- **Co-ordinate term (co-hyponym)**
  
  Chicago, Boston, Austin, Los Angeles

- **Meronym**
  
  Bumper is-part-of car
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:**
  - **Company-Founding**
    - Company: IBM
    - Location: New York
    - Date: June 16, 1911
    - Original-Name: Computing-Tabulating-Recording Co.

- **But we will focus on the simpler task of extracting relation triples**
  - Founding-year(IBM, 1911)
  - Founding-location(IBM, New York)
Dan Jurafsky

Extracting Relation Triples from Text

Stanford University, commonly referred to as Stanford University, 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 LOG NEAR Palo Alto
Stanford FOUNDED IN 1891
Stanford FOUNDER Leland Stanford
Why Relation Extraction?

• Create new structured knowledge bases
• Augment current knowledge bases
  • Lexical resources: Add words to WordNet thesaurus
  • Fact bases: Add facts to FreeBase or DBPedia
• Sample application: question answering
  • The granddaughter of which actor starred in the movie “E.T.”?
• But which relations should we extract?
Automated Content Extraction (ACE)

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

- **Physical-Located**  
  He was in Tennessee

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

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

- **Org-AFF-Founder**  
  Steve Jobs, co-founder of Apple...
Databases of Wikipedia Relations

Wikipedia Infobox

Relations extracted from Infobox

Stanford state California

motto “Die Luft der Freiheit weht”

Type Private
Endowment US$ 16.5 billion (2011)[3]
President John L. Hennessy
Provost John Etchemendy
Academic staff 1,910[4]
Students 15,319
Undergraduates 6,878[5]
Postgraduates 8,441[5]
Location Stanford, California, U.S.
Campus Suburban, 8,180 acres (3,310 ha)[8]
Colors Cardinal red and white
Thesaurus relations

• **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

• **Co-ordinate term (co-hyponym)**
  Chicago, Boston, Austin, Los Angeles

• **Meronym**
  bumper is part-of car
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
  biology/organism_higher_classification film/film/genre
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
Learning Thesauruses and Knowledge Bases

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?

- Founder?
- Investor?
- Member?
- Employee?
- President?
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
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
Learning Thesauruses and Knowledge Bases

Using patterns to extract relations
Learning Thesauruses and Knowledge Bases

Supervised relation extraction
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)
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”
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 for Relation Extraction

**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**]
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

- Constituent path through the tree from one to the other
  
  NP  ↑  NP  ↑  S  ↑  S  ↓  NP

- Dependency path

```
Airlines  matched  Wagner  said
```

subj  comp  subj
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
  - Naive Bayes
  - Logistic Regression (MaxEnt)
  - 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
Learning Thesauruses and Knowledge Bases

Supervised relation extraction
Learning Thesauruses and Knowledge Bases

Semi-supervised 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 Wikipedia. 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
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(born-in | f₁, f₂, f₃, ..., f₇₀₀₀₀)
Distantly supervised learning of IS-A extraction patterns

1. For each X IS-A Y in WordNet

2. Find sentences in large corpus with X and Y

3. Extract parse path between X and Y

4. Represent each noun pair as a 70,000d vector with counts for each of 70,000 parse patterns

5. Train supervised classifier

<sarcoma, cancer>
<deuterium, atom>

an uncommon bone cancer called osteogenic sarcoma in the doubly heavy hydrogen atom called deuterium.

P(IS-A,X,Y | f_1,f_2,f_3,...,f_{7000})

Snow, Jurafsky, Ng 2005
Using Discovered Patterns to Find Novel Hyponym/Hypernym Pairs

\[
<\text{hypernym}> \textit{called} <\text{hyponym}>
\]

Learned from:

“sarcoma / cancer”: ...an uncommon bone cancer \textit{called} osteogenic sarcoma and to...
“deuterium / atom” ....heavy water rich in the doubly heavy hydrogen atom \textit{called} deuterium.

Discovers new hypernym pairs:

“efflorescence / condition”: ...and a \textit{condition} \textit{called} efflorescence are other reasons for...
“\textit{hat\_creek\_outfit} / \textit{ranch}” ...run a small \textit{ranch} \textit{called} the Hat Creek Outfit.
“tardive\_dyskinesia / problem”: ...irreversible \textit{problem} \textit{called} tardive dyskinesia...
“\textit{bateau\_mouche} / \textit{attraction}” ...local sightseeing \textit{attraction} \textit{called} the Bateau Mouche...
Precision / Recall for each of the 70,000 parse patterns considered as a single classifier

Snow, Jurafsky, Ng 2005
Can even combine multiple relations

**IS-A** (hyponym):

Learn by distant supervision

San Francisco IS-A city IS-A municipality IS-A populated area IS-A geographic region...

**Co-ordinate term** (co-hyponym)

Learn by distributional similarity

Chicago, Boston, Austin, Los Angeles, San Diego
Overcoming Hypernym Sparsity with Distributional Information

Snow, Jurafsky, Ng (2006)

What is the hypernym of San Diego?

Coordinate Classifier: “is similar to”

San Diego
San Francisco
Denver
Seattle
Cincinnati
Pittsburgh
New York city
Detroit
Boston
Chicago

Hypernym Classifier: “is a kind of”

-------
city
-------
-------
-------
place, city
-------
city
Learning Thesauruses and Knowledge Bases

Semi-supervised relation extraction
Summary

• Thesaurus Induction
  • Hypernymy, meronymy

• Mostly modeled as relation extraction
  • Pattern-based
  • Supervised
  • Semisupervised and bootstrapping

• Then combined with synonymy from distributional semantics
Computing Relations between Word Meaning: Summary from Lectures 1-6

• Word Similarity/Relatedness/Synonymy
  • Graph algorithms based on human-built Wordnet (Lec 2)
  • Learn from distributional/vector semantics (Lec 3,4)

• Word Connotation (Lec 5)
  • Human hand-labeled
  • Supervised from Reviews
  • Semisupervised from seed words

• Hypernymy (IS-A) (Lec 6)
  • Modeled as relation extraction
    • Supervised, semisupervised