

Relation Extraction

Dr. Demetrios Glinos
University of Central Florida

CAP6640 – Computer Understanding of Natural Language

Today

- The relation extraction task
- Using patterns to extract relations
- Supervised relation extraction
- Semi-supervised and unsupervised methods

Relation Extraction

- From Wikipedia:

IBM (International Business Machines Corporation) is an American [multinational technology](#) company headquartered in [Armonk, New York, United States](#), with operations in over 170 countries. The company originated in 1911 as the [Computing-Tabulating-Recording Company](#) (CTR) and was renamed "International Business Machines" in 1924.

- Extracted complex relation

Company-Founding:

Company	IBM
Location	Armonk, NY
Year	1911
Original-Name	Computing-Tabulating-Recording Co.

- We will focus on the simpler task of extracting relation **triples**

Founding-year(IBM, 1911)

Founding-location(IBM, New York)

Extracting Relation Triples from Text



The screenshot shows the Wikipedia page for the University of Central Florida. The main text states: "The **University of Central Florida**, or **UCF**, is an American public state university in Orlando, Florida. Among U.S. colleges and universities, it has the largest by enrollment at a single campus.^[3]" It also mentions it was founded in 1963 by the Florida Legislature and opened in 1968 as Florida Technological University. A sidebar on the right contains a table with details about the university.

Former names	Florida Technological University
Motto	Reach for the Stars
Type	Public state university Space-grant university ^[1]
Established	June 10, 1963

The **University of Central Florida**, or **UCF**, is an American public state university in Orlando Florida. Among U.S. colleges and universities, it has the largest by enrollment at a single campus. Founded in 1963 ...

Relations extracted:

UCF
UCF
UCF
UCF

EQ
LOC
IS-A
FOUNDED-IN

University of Central Florida
Orlando, Florida
public state university
1963

Why Extract Relations?

- Create new knowledge bases
 - structured data useful for many applications
- Augment current knowledge bases
 - e.g., adding words to existing dictionaries, thesauri, or ontologies such as WordNet
- Support question answering

Q: Which actor's granddaughter 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)

Type	Subtype
ART (artifact)	User-Owner-Inventor-Manufacturer
GEN-AFF (General affiliation)	Citizen-Resident-Religion-Ethnicity, Org-Location
METONYMY*	<i>None</i>
ORG-AFF (Org-affiliation)	Employment, Founder, Ownership, Student-Alum, Sports-Affiliation, Investor-Shareholder, Membership
PART-WHOLE (part-to-whole)	Artifact, Geographical, Subsidiary
PER-SOC* (person-social)	Business, Family, Lasting-Personal
PHYS* (physical)	Located, Near

- source: Fig. 4, Automatic Content Extraction 2008 Evaluation Plan (ACE08)

Examples: ACE relations

- Physical-Located

PER-GPE: He was in California

- Part-Whole-Subsidiary

ORG-ORG: XYZ, the parent company of ABC

- Person-Social-Family

PER-PER: John's wife Yoko

- Org-Affiliation-Founder

PER-ORG: Steve Jobs, co-founder of Apple

Type	Subtype
ART (artifact)	User-Owner-Inventor-Manufacturer
GEN-AFF (General affiliation)	Citizen-Resident-Religion-Ethnicity, Org-Location
METONYMY*	<i>None</i>
ORG-AFF (Org-affiliation)	Employment, Founder, Ownership, Student-Alum, Sports-Affiliation, Investor-Shareholder, Membership
PART-WHOLE (part-to-whole)	Artifact, Geographical, Subsidiary
PER-SOC* (person-social)	Business, Family, Lasting-Personal
PHYS* (physical)	Located, Near

Unified Medical Language System

- Published and maintained by the U.S. National Library of Medicine
- Maps many controlled vocabularies in the biomedical sciences
- Also serves as a thesaurus and ontology of biomedical concepts

- 134 entity types, 54 relations

- Examples

injury

bodily location

anatomical structure

pharmacologic substance

pharmacologic substance

disrupts

location-of

part-of

causes

treats

physiological function

biologic function

organism

pathological function

pathological function

Extracting UMLS relations

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



Echocardiography, Doppler **DIAGNOSES** Acquired stenosis




(Note how terms are mapped to the controlled vocabularies)

Wikipedia Relations

- Wikipedia infobox categories express relations that are relevant to the entity type

 <p>Theatrical release poster by John Alvin^[1]</p>	
Directed by	Steven Spielberg
Produced by	Kathleen Kennedy Steven Spielberg
Written by	Melissa Mathison
Starring	Dee Wallace Peter Coyote Henry Thomas
Music by	John Williams
Cinematography	Allen Daviau
Edited by	Carol Littleton
Production company	Universal Pictures
Distributed by	Universal Pictures
Release date	May 26, 1982 (Cannes) June 11, 1982 (United States)
Running time	114 minutes ^[2]
Country	United States
Language	English
Budget	\$10.5 million ^[3]
Box office	\$792.9 million ^[3]

 <p>Logo as of 1972</p>  <p>IBM Watson system in 2011</p>	
Type	Public
Traded as	NYSE: IBM ⓘ DJIA Component S&P 100 Component S&P 500 Component
ISIN	US4592001014
Industry	Cloud computing · Cognitive computing
Founded	June 16, 1911; 106 years ago (as Computing-Tabulating-Recording Company) Endicott, New York, U.S. ^[1]
Founder	Charles Ranlett Flint
Headquarters	Armonk, New York, U.S.
Area served	177 countries ^[2]
Key people	Ginni Rometty (Chairwoman, President and CEO)
Products	See IBM products
Revenue	▼ US\$ 79.139 billion (2017) ^[3]
Operating income	▼ US\$ 11.400 billion (2017) ^[3]
Net income	▼ US\$ 5.753 billion (2017) ^[3]
Total assets	▲ US\$ 125.35 billion (2017) ^[3]
Total equity	▼ US\$ 17.594 billion (2017) ^[3]
Number of employees	380,300 (2017) ^[4]
Website	www.ibm.com ⓘ

	
Former names	Florida Technological University
Motto	<i>Reach for the Stars</i>
Type	Public state university Space-grant university ^[1]
Established	June 10, 1963
Endowment	US\$146.9 million ^[2]
President	John C. Hitt
Provost	Dale Whittaker
Academic staff	2,686 (Fall 2016) ^[3]
Administrative staff	9,900 (Fall 2016) ^[3]
Students	66,183 (Fall 2017) ^[4]
Location	Orlando, Florida, U.S.
Campus	Suburban Main: 1,415 acres (5.73 km ²) Total: 1,893 acres (7.66 km ²) ^[5]
Colors	Black and Gold ^[6] 
Nickname	Knights
Sporting affiliations	NCAA Division I, FBS The American
Mascot	Knightro, Pegasus and the UCF Knight
Website	www.ucf.edu ⓘ
 UNIVERSITY OF CENTRAL FLORIDA	

Relation Databases

- **Resource Description Framework (RDF) triples**
 - < subject, predicate, object >
 - e.g., < Central Park, location, New York City>
- **DBPedia**
 - Published by Free University of Berlin and Leipzig University
 - project to extract RDF triples from Wikipedia
 - currently 3 billion RDF triples
 - 580 million from English Wikipedia, 2.46 billion from other languages
- **Wikidata**
 - formerly Freebase
 - working to provide centralized links between Wikipedia articles in different languages on the same topic
 - also, to provide a central place for infobox data for all Wikipedias

Ontological relations

- **Ontological relationships**
 - how one concept is related to another
 - there are many ways in which concepts may be related
 - particular concepts may or may not be related by a particular relation type
- **WordNet**
 - a lexical database for the English language; included in NLTK
 - developed by the Cognitive Sciences Laboratory at Princeton Univ.
 - groups words into "synsets" and encodes several semantic relations, including:
 - **IS-A (hypernym) relation: subsumption between classes**
 - e.g., giraffe **IS-A** ruminant **IS-A** ruminant **IS-A** ungulate **IS-A** mammal **IS-A** vertebrate **IS-A** animal ...
 - **INSTANCE-OF: relation between an individual and a class**
 - Orlando **INSTANCE-OF** city

demo: wnb

Today

- The relation extraction task
- Using patterns to extract relations
- Supervised relation extraction
- Semi-supervised and unsupervised methods

Rules for extracting IS-A relation

- Early intuition from M. Hearst, Automatic Acquisition of Hyponyms from Large Text Corpora (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 that?

Rules for extracting IS-A relation

- Early intuition from M. Hearst, Automatic Acquisition of Hyponyms from Large Text Corpora (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 that?

Hearst's patterns for IS-A

1. "NP₀ such as { NP₁, NP₂ , (and/or)} NP_n"

e.g., The bow lute, such as the Bambara ndang, is plucked ...

2. "such NP as { NP ,}* {(or|and)} NP"

e.g., ... works by such authors as Herrick, Goldsmith, and Shakespeare.

3. "NP { , NP}* { , } or other NP"

e.g., Bruises, wounds, broken bones or other injuries...

4. "NP { , NP}* { , } and other NP"

e.g., ... temples, treasuries, and other important civic buildings.

5. "NP { , } including {NP ,}* {or|and} NP"

e.g., All common-law countries, including Canada and England ...

6. "NP { , } especially {NP ,}* {or|and} NP"

e.g., ... most European countries, especially France, England, and Spain.

Extracting Richer Relations Using Rules

- Intuition: specific types of relations hold between specific types of entities
 - **located-in** (ORGANIZATION, LOCATION)
 - **founded** (PERSON, ORGANIZATION)
 - **cures** (DRUG, DISEASE)
- So, start with named entity tags help extract the relation

Named Entities are Not Sufficient

- Which relation holds between these two entities?



Pharmaceutical

Cure ?

Prevent ?

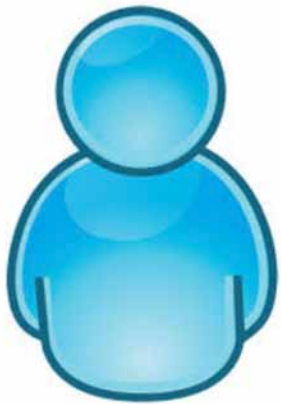
Cause ?



Disease

Named Entities are Not Sufficient

- Which relation holds between these two entities?



Person

Founder ?

Investor ?

Member ?

Employee ?

President ?



Organization

Using Rules and Named Entities

- Relation: The holder of a particular office in an organization
- Some rules to extract this relation

PER, POSITION of ORG

- George Marshall, Secretary of State of the United States

PER (named|appointed|selected|etc.) PER Prep? POSITION

- Truman appointed Marshall Secretary of State

PER (be)? (named|appointed|selected|etc.) Prep? ORG ('s)? POSITION

- George Marshall was named US Secretary of State

Effectiveness of Hand-Built Patterns

- Advantages

- human-developed patterns tend to be high-precision
- can be tailored to specific domains

- Disadvantages

- human-developed patterns also tend to be low-recall
- much effort is required to develop a good set of patterns
- we don't wish to do this for every relation
- we desire better accuracy

Today

- The relation extraction task
- Using patterns to extract relations
- Supervised relation extraction
- Semi-supervised and unsupervised methods

Supervised Machine Learning of Relations

- Choose a set of relations we wish to extract
- Choose a set of relevant named entity types for the relations
- Find and label data
 - choose a representative corpus
 - label the named entities in the corpus
 - hand-label the relations among the entities
 - divide the corpus into training, development, and test sets
- Train a classifier on the training set

Classifying Novel Instances

1. Find all pairs of named entities (usually in the same sentence)
2. Examine each pair of entities and decide if they are related
3. If yes, then classify the relation

Q: Why is step 2 needed?

A: Faster classification, since most pairs are eliminated

Also, can use different feature sets for each type of entity pair

Example: Novel Instance

- Classify the relation between the two identified entities in this sentence:

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

- For this PER-ORG combination, the relation can be

- GEN-AFF
 - citizen, resident, etc.
- ORG-AFF
 - founder, employee, etc.

Type	Subtype
ART (artifact)	User-Owner-Inventor-Manufacturer
GEN-AFF (General affiliation)	Citizen-Resident-Religion-Ethnicity, Org-Location
METONYMY*	None
ORG-AFF (Org-affiliation)	Employment, Founder, Ownership, Student-Alum, Sports-Affiliation, Investor-Shareholder, Membership
PART-WHOLE (part-to-whole)	Artifact, Geographical, Subsidiary
PER-SOC* (person-social)	Business, Family, Lasting-Personal
PHYS* (physical)	Located, Near

Word Features for Relation Extraction

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

Mention 1

Mention 2

- Head words 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 position to left and right of M1/M2
 - M2: -1 spokesman
 - M2: +1 said
- Bag of words or bigrams between the two entities
 - { a, AMR, immediately, matched, move, of, 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: **PER**
- Concatenation of the two named-entity types
 - **ORG-PER**
- Mention level of M1 and M2 (**NAME**, **NOMINAL**, **PRONOMINAL**)
 - M1: **NAME** ["the company" would be **NOMINAL**,
"it" would be **PRONOMINAL**]
 - M2: **NAME** ["the spokesman" would be **NOMINAL**,
"he" would be **PRONOMINAL**]

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 mention to the other
 - NP NP PP VP NP NP ← "spokesman Tim Wagner" is the last NP
- Constituent path through the tree from one to the other
 - NP ↑ NP ↑ S ↓ VP ↓ S ↓ NP ← last NP is "American Airlines"
- Words in dependency path from one mention to the other
 - Airlines matched Wagner said

Gazeteer and Trigger Word Features

- Trigger lists for family kinship terms
 - parent, wife, husband, grandparent, etc.
- Gazetteer lists of useful geographical or geopolitical entities
 - country name lists
 - other sub-entities

Classifiers for Supervised Methods

- Assemble all desired features into feature vectors
 - entity-based features
 - word-based features
 - syntactic features
 - parse-based features
 - etc.
- Use your favorite classifier
 - MaxEnt
 - Naïve Bayes
 - SVM
 - ...
- Train on training set, tune on dev set, test on test set

Evaluation of Supervised Methods

- Compute precision, recall, and F_1 score for each relation

$$P = \frac{\text{\textit{\# of correctly extracted relations}}}{\text{\textit{Total \# of extracted relations}}}$$

$$R = \frac{\text{\textit{\# of correctly extracted relations}}}{\text{\textit{Total \# of gold relations}}}$$

$$F_1 = \frac{2PR}{P + R}$$

Effectiveness of Supervised Methods

- Advantage
 - Can get high accuracies with sufficient hand-labeled training data, provided test data is sufficiently similar to training data
- Disadvantages
 - Labeling a large training set is expensive
 - Supervised models are brittle and don't generalize well to different genres

Today

- The relation extraction task
- Using patterns to extract relations
- Supervised relation extraction
- Semi-supervised and unsupervised methods

Seed-Based or Bootstrapping Approaches

- What if you don't have a training set?
- If you have
 - a few seed tuples
 - or, a few high-precision patterns
- Then you can use "bootstrapping" to 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 to find more pairs

Example: Bootstrapping

- Suppose we have the **seed tuple** <Mark Twain, Elmira>
 - Search (grep, Google) for the contexts in which the seed tuple occurs

"Mark Twain is buried in Elmira, NY."

PER is buried in LOC

"The grave of Mark Twain is in Elmira."

The grave of PER is in LOC

"Elmira is Mark Twain's final resting place."

LOC is PER's final resting place

- Use these new patterns to find more tuples
- Mine those tuples for even more patterns
- Iterate

Extracting <author,book> pairs

Sergey Brin, Extracting Patterns and Relations from the World Wide Web (1998)

- Started with 5 seeds:

Author	Book
Isaac Asimov	The Robots of Dawn
David Brin	Startide Rising
James Gleick	Chaos: Making a New Science
Charles Dickens	Great Expectations
Eilliam Shakespeare	The Comedy of Errors

- Found 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 ...

- Extracted patterns

?x , by ?y , ?x , one of ?y 's

- Iterated
- Ultimately found 15,257 unique book titles

Extracting <ORG,LOC> pairs

Snowball: Extracting Relations from Large Plain-Text Collections (2000)

- Started with 3 seeds:

Organization	Headquarters Location
Microsoft	Redmond
Exxon	Irving
IBM	Armonk

- Grouped instances with similar prefix, middle, suffix, extraction patterns
 - but required that X and Y be named entities
 - also computed confidence for each pattern
- Generated extraction patterns like
 - ORG 's headquarters in LOC
 - LOC –based ORG
 - ORG , LOC
- Tested on over 300,000 newspaper articles
 - results show comparable recall and improved precision over baseline

Distant Supervision

Mintz, Bills, Snow, Jurafsky, Distant Supervision for Relation Extraction Without Labeled Data (2009)

- Basic idea:
 - combine bootstrapping with supervised learning
- instead of 5 seeds,
 - mine an existing (effectively labeled) large database of relations (Freebase)
 - this produces a huge set of seed examples
- search a large unlabeled corpus (e.g., WWW) for instances of these examples
- create lots of features from all of these examples
- train a supervised classifier on these examples
- run the classifier on the unlabeled corpus to discover novel instances
- achieved precision of 67.6%

Example: Distant Supervision

- For each relation in the labeled collection **Born-in**
- For each tuple in the labeled database **<Edwin Hubble, Marshfield>**
- Find sentences in the unlabeled database with both entities Hubble was born in Marshfield
Einstein, born (1879), Ulm
Hubble's birthplace in Marshfield
- Extract frequent features (parse, words, etc.) **PER** was born in **LOC**
PER, born (XXXX), **LOC**
PER's birthplace in **LOC**
- Train supervised classifier using thousands of patterns $P(\text{born-in} | f_1, f_2, \dots, f_{70000})$
- Evaluate using held-out data from labeled database

TextRunner: Unsupervised Relation Extraction

M. Banko, M. Cararella, S. Soderland, M. Broadhead, and O. Etzioni, Open Information Extraction from the Web (2007)

- Defined the term "Open Information Extraction (OIE)"
 - extracting relations from the web with no training data, nor list of relations
- Learner component
 - uses a deep linguistic parser to extract relationships on a small subset of the large unlabeled corpus (the Web)
 - labels entities and extracts "trustworthy tuples" (relations)
 - train a Naïve Bayes classifier
- Assessor component
 - single pass over data to extract all relations between NPs, keep if trustworthy
 - relations are ranked based on text redundancy in the corpus
 - e.g., <Tesla, invented, coil transformer>
- High-throughput relation discovery, with lower error rate than competing systems

Evaluating Semi-Supervised and Unsupervised Methods

- These methods extract totally new relations from the Web
- There is no gold standard set of correct extractions
 - 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 only approximate precision
 - Draw a random sample of relations from output, check precision manually
- Can also compute approximate precision at different levels of recall
 - i.e., compute approximate precision for top 1,000, or 10,000, etc., new relations
 - in each case, take a random sample of that set
- But there is no way to evaluate recall at all!