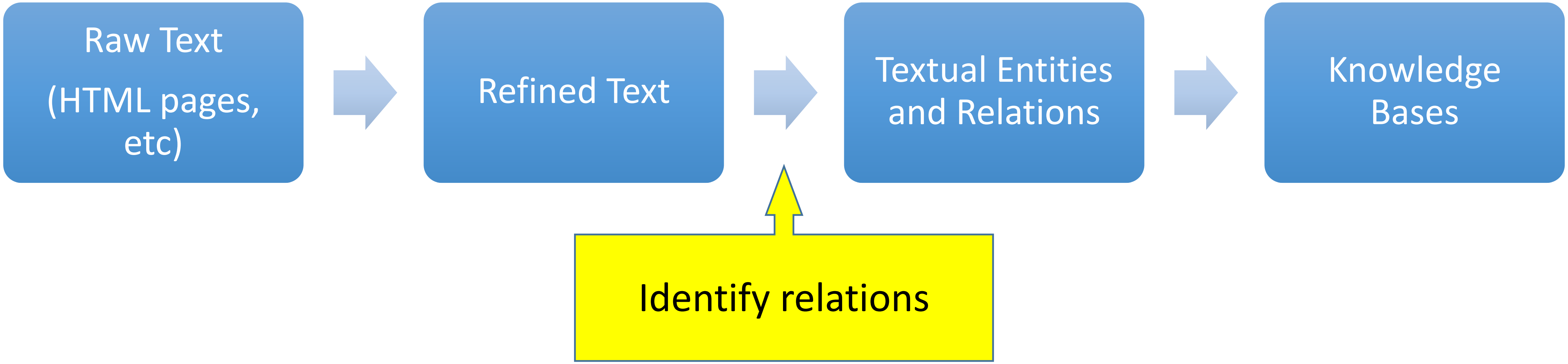
Relation Extraction

Slides from http://web.stanford.edu/~jurafsky/slp3/ Some of this material is also available on coursera

What is knowledge acquisition?

Knowledge acquisition: process to extract knowledge (to be integrated into knowledge bases) from unstructured text or other data



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

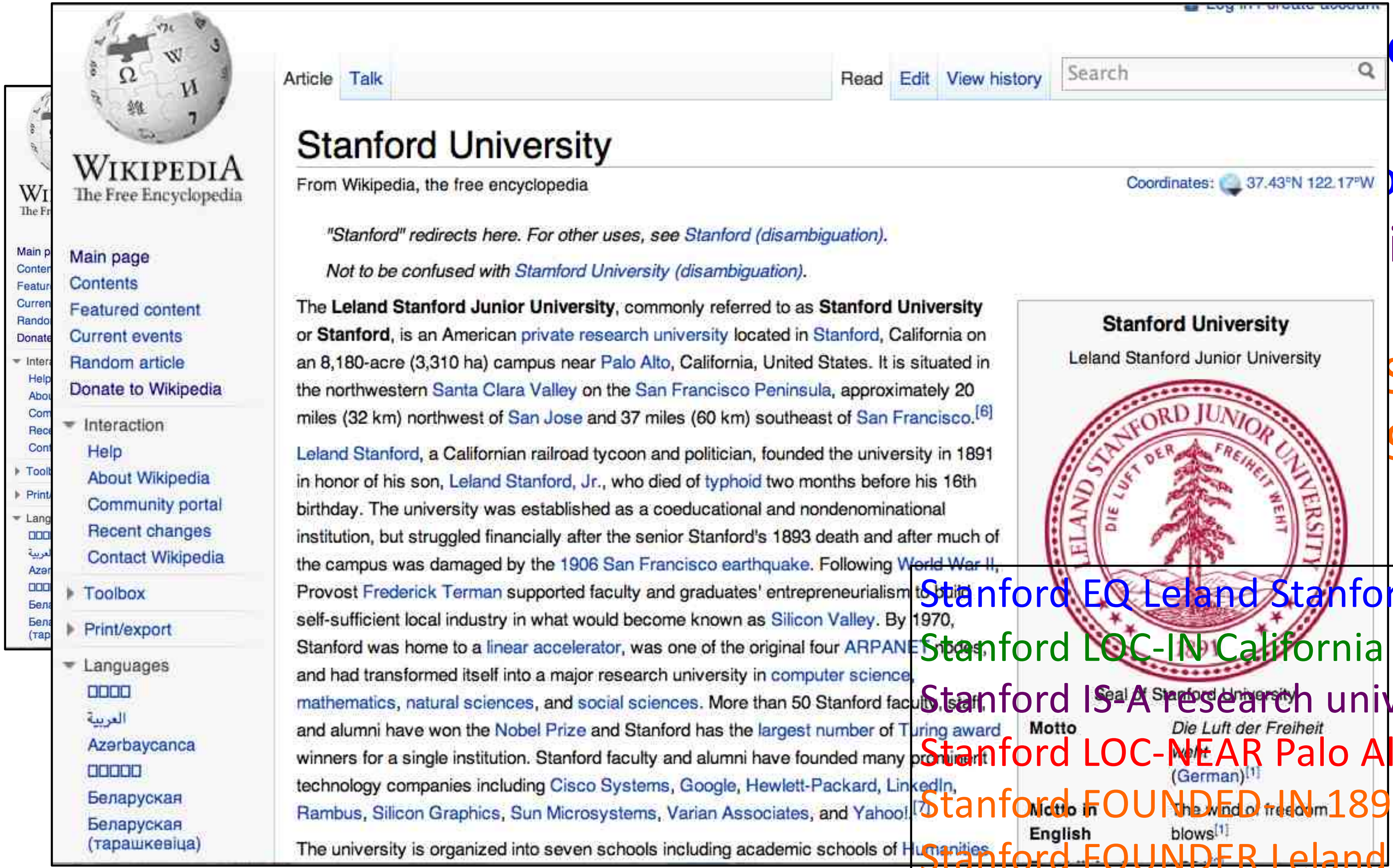
Company	IBM
Location	New York
Date	June 16, 1911

- But 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



d Junior University,
to as Stanford
rd, is an American
iversity located in
… near Palo Alto,
Stanford…founded
91

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

Types of Relation Extraction

- It's hard to enrich knowledge bases manually. **How can we do it automatically?**

- *<u>Traditional Extraction:</u>* start from a set of known relations, and annotated input
- *<u>Open Extraction:</u>* extract relations without any prior information

[1] M. Banko, O. Etzioni, and T. Center, "The Tradeoffs Between Open and Traditional Relation Extraction.," in *ACL*, 2008, vol. 8, pp. 28–36.

Traditional Extraction

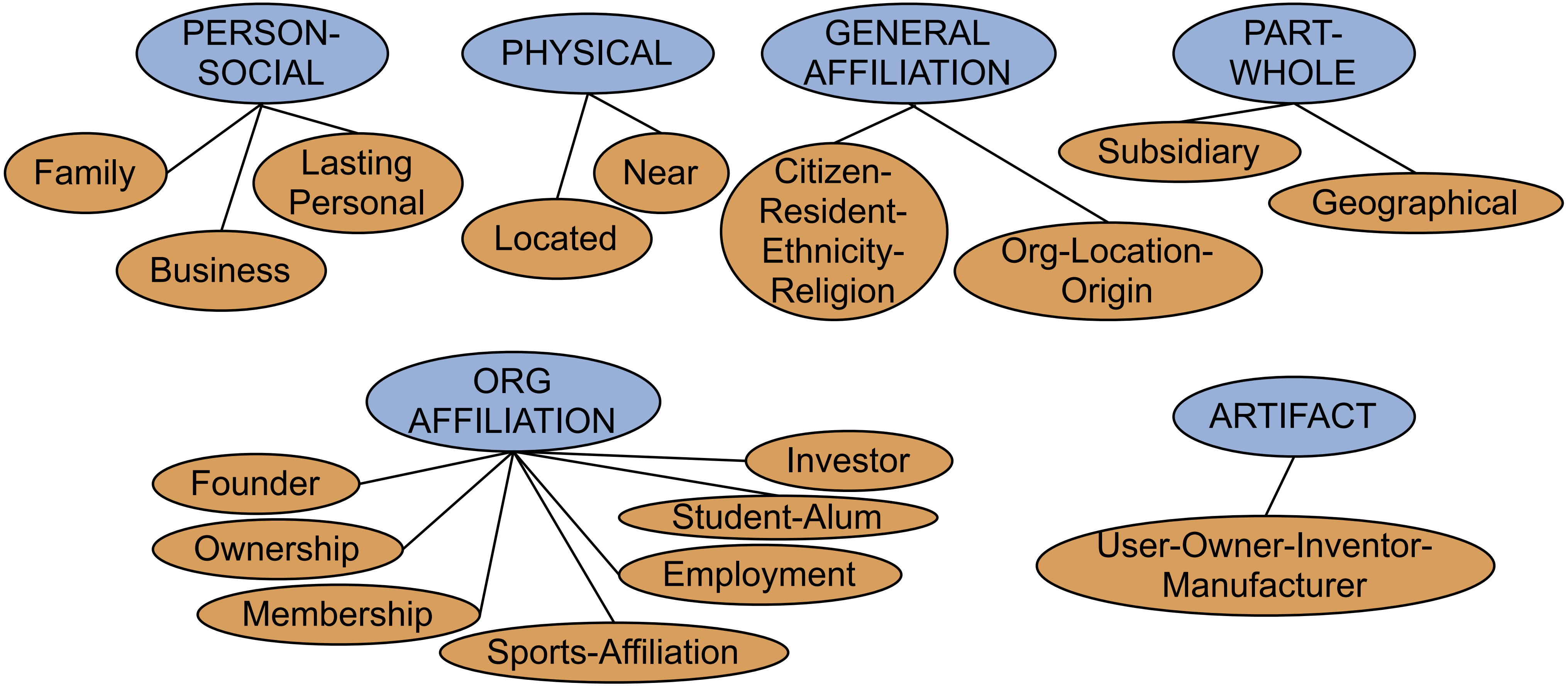
Automated Content Extraction (ACE)

- **Automatic Content Extraction (ACE)** is a research program for developing advanced information extraction convened by the NIST from 1999 to 2008 [1]
- Challenge of the program was to detect
  - **Entities** mentioned in the text, such as: persons, organizations, locations, etc.
  - **Relations** between entities
  - **Events** such as interactions, etc.
- The ACE corpus is one of the standard benchmarks for testing new information extraction algorithms

[1] G. R. Doddington, A. Mitchell, M. A. Przybocki, L. A. Ramshaw, S. Strassel, and R. M. Weischedel, "The Automatic Content Extraction (ACE) Program-Tasks, Data, and Evaluation.," in *LREC*, 2004, vol. 2, p. 1.

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…`

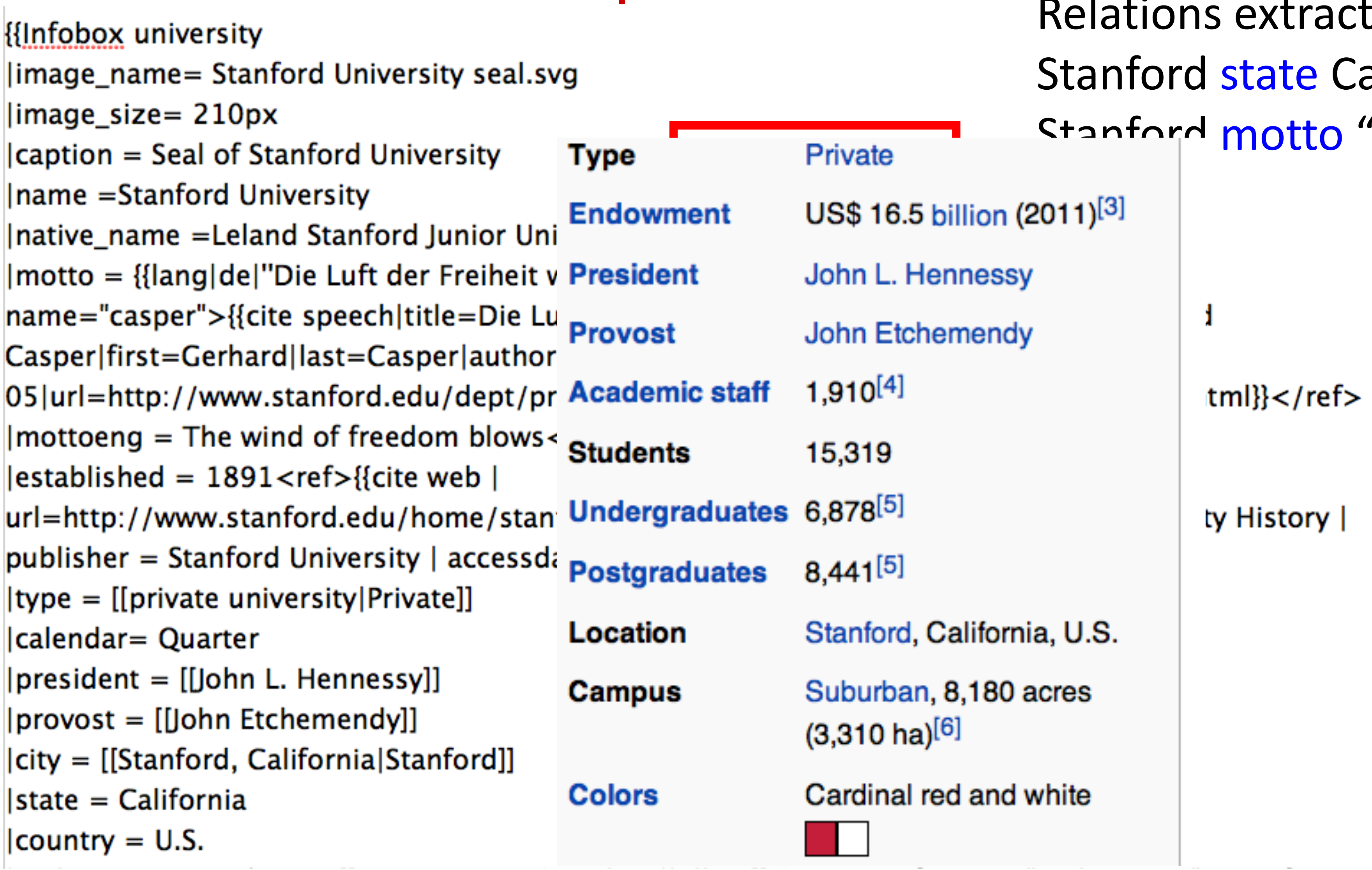
UMLS: Unified Medical Language System

- Specific to the medical domain, defines 134 entity types, 54 relations

Injury	disrupts	Physiological Function
Bodily Location	location-of	Biologic Function
Anatomical Structure	part-of	Organism
Pharmacologic Substance	causes	Pathological Function
Pharmacologic Substance	treats	Pathologic Function

Databases of Wikipedia Relations

Wikipedia Infobox


{{Infobox university
|image_name= Stanford University seal.svg
|image_size= 210px
|caption = Seal of Stanford University
|name =Stanford University
|native_name =Leland Stanford Junior Uni
|motto = {{lang|de|"Die Luft der Freiheit w
name="casper">{{cite speech|title=Die Lu
Casper|first=Gerhard|last=Casper|author
05|url=http://www.stanford.edu/dept/pr tml}}</ref>
|mottoeng = The wind of freedom blows<
|established = 1891<ref>{{cite web |
url=http://www.stanford.edu/home/stan ty History |
publisher = Stanford University | accessda
|type = [[private university|Private]]
|calendar= Quarter
|president = [[John L. Hennessy]]
|provost = [[John Etchemendy]]
|city = [[Stanford, California|Stanford]]
|state = California
|country = U.S.

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)[6]
Colors	Cardinal red and white

Relations extracted from Infobox

Stanford state California

Stanford motto “Die Luft der Freiheit weht”

Ontological relations

Examples from the WordNet Thesaurus

- 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`

Types of traditional relational extraction methods

1. Hand-written patterns
2. Supervised machine learning
3. Semi-supervised
   - Bootstrapping (using seeds)
   - Distant supervision

Types of traditional relational extraction methods

1. **Hand-written patterns**
2. Supervised machine learning
3. Semi-supervised
   - Bootstrapping (using seeds)
   - Distant supervision

Rules for extracting IS-A relation

Early intuition from **Hearst (1992) [1]**

- "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?

[1] M. A. Hearst, "Automatic acquisition of hyponyms from large text corpora," in *Proceedings of the 14th conference on Computational linguistics-Volume 2,* 1992, pp. 539–545.

Rules for extracting IS-A relation

Early intuition from **Hearst (1992) [1]**

- "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?`

[1] M. A. Hearst, "Automatic acquisition of hyponyms from large text corpora," in *Proceedings of the 14th conference on Computational linguistics-Volume 2,* 1992, pp. 539–545.

Hearst's Patterns for extracting IS-A relations

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

Hearst pattern	Example occurrences
X and other Y	...temples, treasuries, and other important civic buildings.
X or other Y	Bruises, wounds, broken bones or other injuries...
Y such as X	The bow lute, such as the Bambara ndang...
Such Y as X	...such authors as Herrick, Goldsmith, and Shakespeare.
Y including X	...common-law countries, including Canada and England...
Y , especially X	European countries, especially France, England, and Spain...

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!

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

- 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

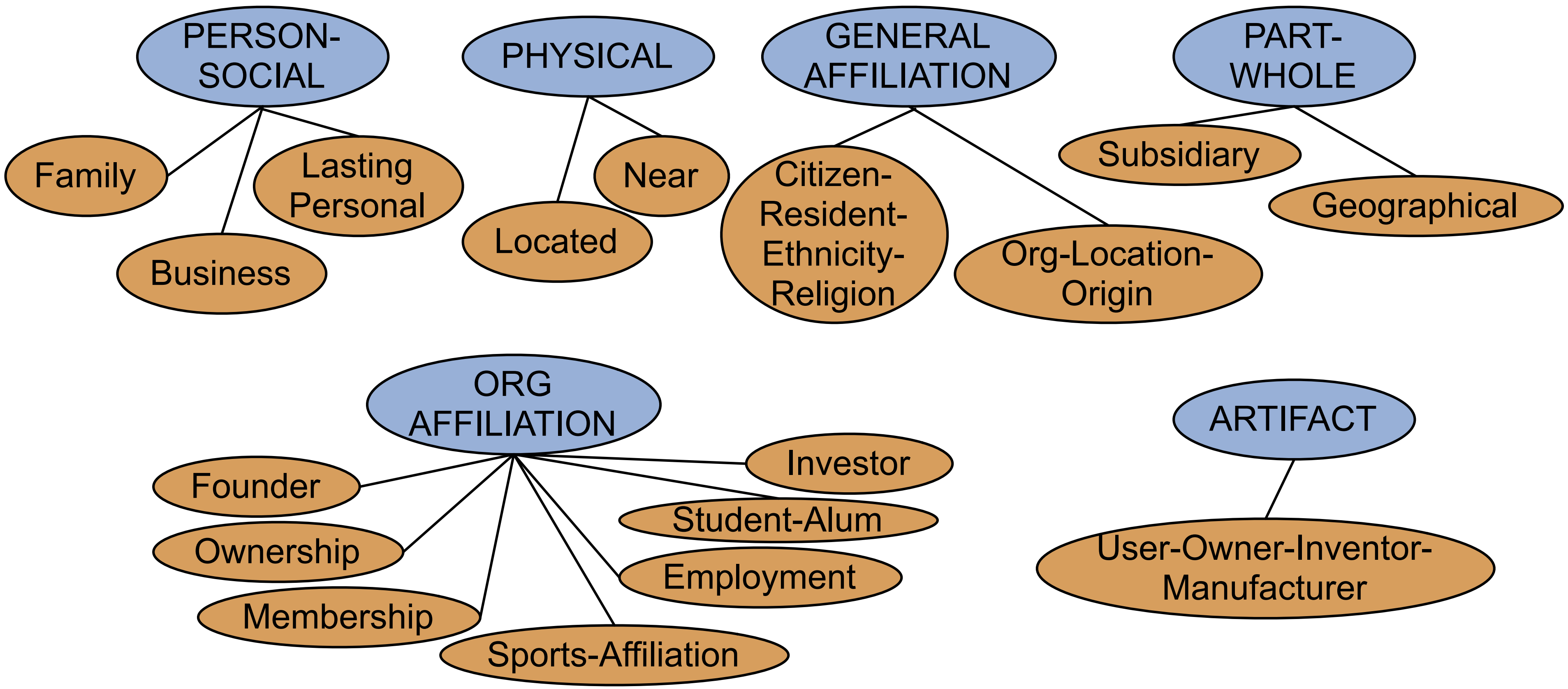
Types of traditional relational extraction methods

1. Hand-written patterns
2. **Supervised machine learning**
3. Semi-supervised
   - Bootstrapping (using seeds)
   - Distant supervision

Supervised machine learning for relations

- Choose a set of relations we'd like to extract (e.g. ACE, UMLS)
- 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

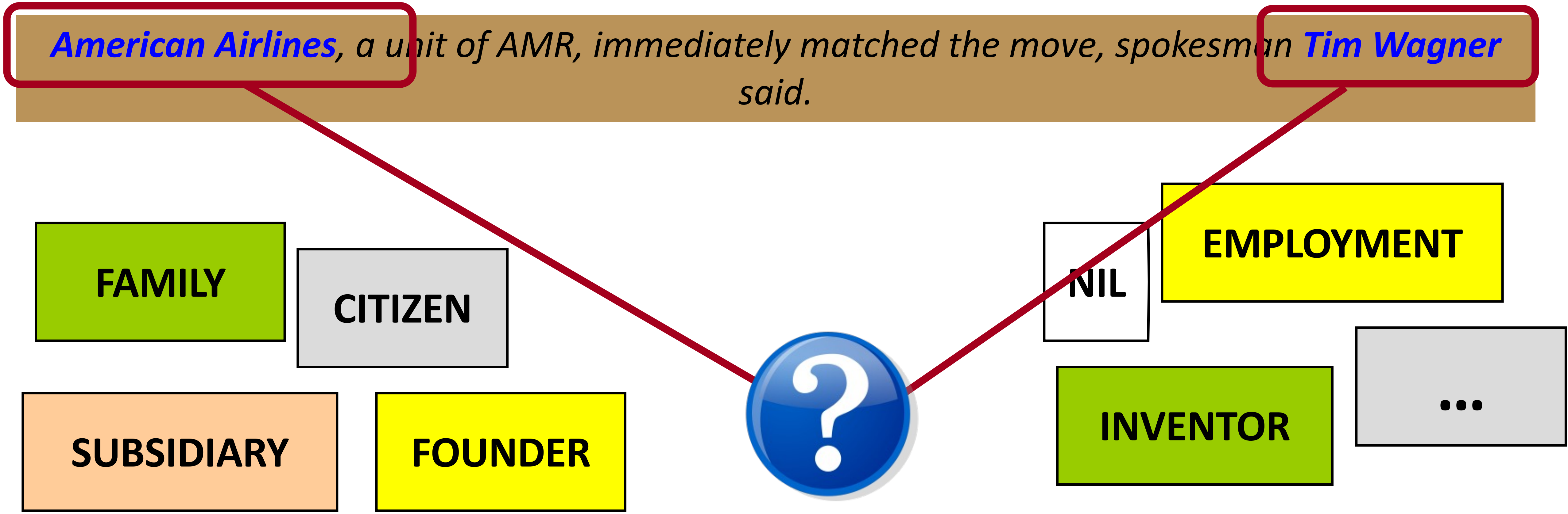
Choose a set of relations we'd like to extract



17 sub-relations of 6 relations from 2008 “Relation Extraction Task”

Find and Label Data

Classify the relation between two entities in a sentence



Common Word Features for Classifier

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}

Common Word Features for Classifier

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*

Mention 1 Mention 2

- Constituent path through the tree from one to the other

  NP ↑ NP ↑ S ↓ NP

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

Entity-based features		
	Entity$_1$ type	ORG
	Entity$_1$ head	*airlines*
	Entity$_2$ type	PERS
	Entity$_2$ 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$_1$	NONE
	Word(s) after Entity$_2$	*said*
Syntactic features		
	Constituent path	$NP \uparrow NP \uparrow S \uparrow S \downarrow NP$
	Base syntactic chunk path	$NP \rightarrow NP \rightarrow PP \rightarrow NP \rightarrow VP \rightarrow NP \rightarrow NP$
	Typed-dependency path	$Airlines \leftarrow_{subj} matched \leftarrow_{comp} said \rightarrow_{subj} Wagner$

Classifiers for supervised methods

- Now you can use any classifier you like
  - Max Entropy
  - Naïve Bayes
  - SVM
  - ...
- Train it on the *training set*, tune on the *dev set*, test on the *test set*

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

Types of traditional relational extraction methods

1. Hand-written patterns
2. Supervised machine learning
3. **Semi-supervised**
   - Bootstrapping (using seeds)
   - Distant supervision

Relation Bootstrapping

- Supervised methods assume you have a (large) training set that is available
- 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

- 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 [1]: Extract <author,book> pairs

- Start 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
William Shakespeare	The Comedy of Errors

- 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

[1] S. Brin, "Extracting patterns and relations from the world wide web," in *International Workshop on The World Wide Web and Databases*, 1998, pp. 172–183.

Snowball [1]

- Inspired by DIPRE.
  Similar iterative algorithm

Organization	Location of Headquarters
Microsoft	Redmond
Exxon	Irving
IBM	Armonk

- Group instances w/similar prefix, middle, suffix, extract patterns
  - But require that X and Y be named entities (DIPRE did not do this)
  - And compute a confidence for each pattern

.69 ORGANIZATION `{'s, in, headquarters}` LOCATION

.75 LOCATION `{in, based}` ORGANIZATION

[1] E. Agichtein and L. Gravano, "Snowball: Extracting relations from large plain-text collections," in *Proceedings of the fifth ACM conference on Digital libraries*, 2000, pp. 85–94.

Snowball

Example of calculation of a pattern’s confidence

$$Conf(P) = \frac{P.positive}{(P.positive + P.negative)}$$

P=<{},ORGANIZATION, <“,”,1>, LOCATION,{}>
P.positive = “Exxon, Invine said”; “Intel, Santa Clara cut prices”
P.negative = “invest in Microsoft, New York-based analyst Jane Smith said”

Distant Supervision [1]

- Combine bootstrapping with supervised learning
  - Instead of 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

[1] M. Mintz, S. Bills, R. Snow, and D. Jurafsky, "Distant supervision for relation extraction without labeled data," in *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, 2009, pp. 1003–1011.

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

2. For each tuple in big database

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

3. Find sentences in large corpus with both entities

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

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

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

5. Train supervised classifier using thousands of patterns

   $P(\text{born-in} \mid f_1,f_2,f_3,\ldots,f_{70000})$