Information Extraction

(Part 3)

Shobeir Fakhraei
University of Southern California

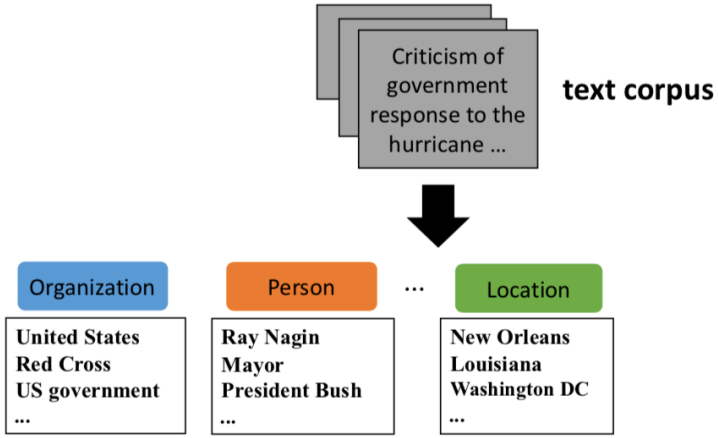
Slides based on material provided by many people, including Dan Jurafsky, Rion Snow, Jim Martin, Chris Manning, William Cohen, Michele Banko, Mike Mintz, Steven Bills, Luke Zettlemoyer, and others.

Where are we?

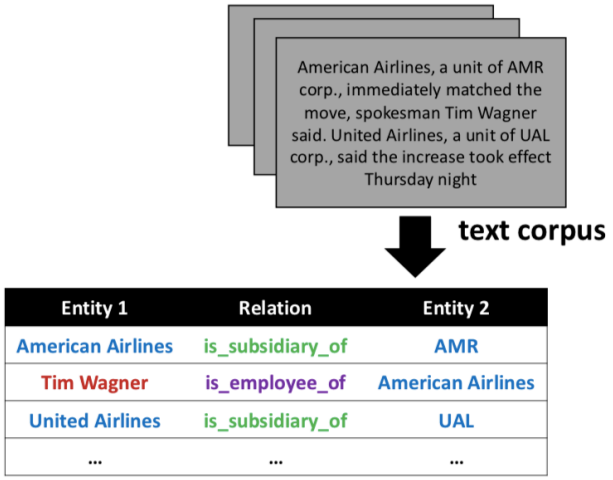
- Information Extraction
  - Semi-structured: Web, HTML
  - **Unstructured: Text, Adds, Tweets, …**
- Entity Linkage, Data Cleaning, Normalization
- Logical Data Integration
  - Mediators, Query Rewriting
  - Warehouse, Logical Data Exchange
- Automatic Source Modeling/Learning Schema Mappings
- Semantic Web
  - RDF, SPARQL, OWL, Linked Data
- Advanced Topics
  - Geospatial Data Integration, Knowledge Graphs

Landscape of IE Tasks (1): Types of Extractions

Named Entity Recognition

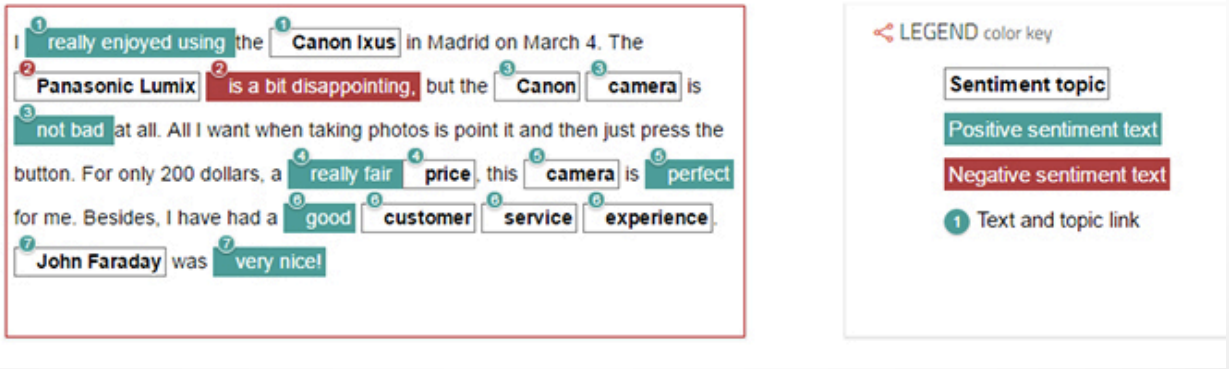


Relationship Extraction



Entity 1	Relation	Entity 2
American Airlines	is_subsidiary_of	AMR
Tim Wagner	is_employee_of	American Airlines
United Airlines	is_subsidiary_of	UAL
...	...	...

Adjective Extractions -> Sentiment Analysis



Relationship Extraction

Machine-readable summaries

IE

Subject	Relation	Object
p53	**is_a**	protein
Bax	**is_a**	protein
p53	has_function	apoptosis
Bax	has_function	induction
apoptosis	involved_in	cell_death
Bax	is_in	mitochondrial outer membrane
Bax	is_in	cytoplasm
apoptosis	related_to	caspase activation
...	...	...

textual abstract: summary for human

structured knowledge extraction: summary for machine

Relation extraction example

CHICAGO (AP) — Citing high fuel prices, United Airlines said Friday it has increased fares by $6 per round trip on flights to some cities also served by lower-cost carriers. American Airlines, a unit of AMR, immediately matched the move, spokesman Tim Wagner said. United, a unit of UAL, said the increase took effect Thursday night and applies to most routes where it competes against discount carriers, such as Chicago to Dallas and Atlanta and Denver to San Francisco, Los Angeles and New York.

Question: What relations should we extract?

example from Jim Martin

Relation extraction example

CHICAGO (AP) — Citing high fuel prices, United Airlines said Friday it has increased fares by $6 per round trip on flights to some cities also served by lower-cost carriers. American Airlines, a unit of AMR, immediately matched the move, spokesman Tim Wagner said. United, a unit of UAL, said the increase took effect Thursday night and applies to most routes where it competes against discount carriers, such as Chicago to Dallas and Atlanta and Denver to San Francisco, Los Angeles and New York.

Subject	Relation	Object
American Airlines	subsidiary	AMR
Tim Wagner	employee	American Airlines
United Airlines	subsidiary	UAL

example from Jim Martin

Relation types

For generic news texts ...

Relations		Examples	Types
Affiliations			
	Personal	*married to, mother of*	PER → PER
	Organizational	*spokesman for, president of*	PER → ORG
	Artifactual	*owns, invented, produces*	(PER | ORG) → ART
Geospatial			
	Proximity	*near, on outskirts*	LOC → LOC
	Directional	*southeast of*	LOC → LOC
Part-Of			
	Organizational	*a unit of, parent of*	ORG → ORG
	Political	*annexed, acquired*	GPE → GPE

slide adapted from Jim Martin

Relation types from ACE 2003

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**

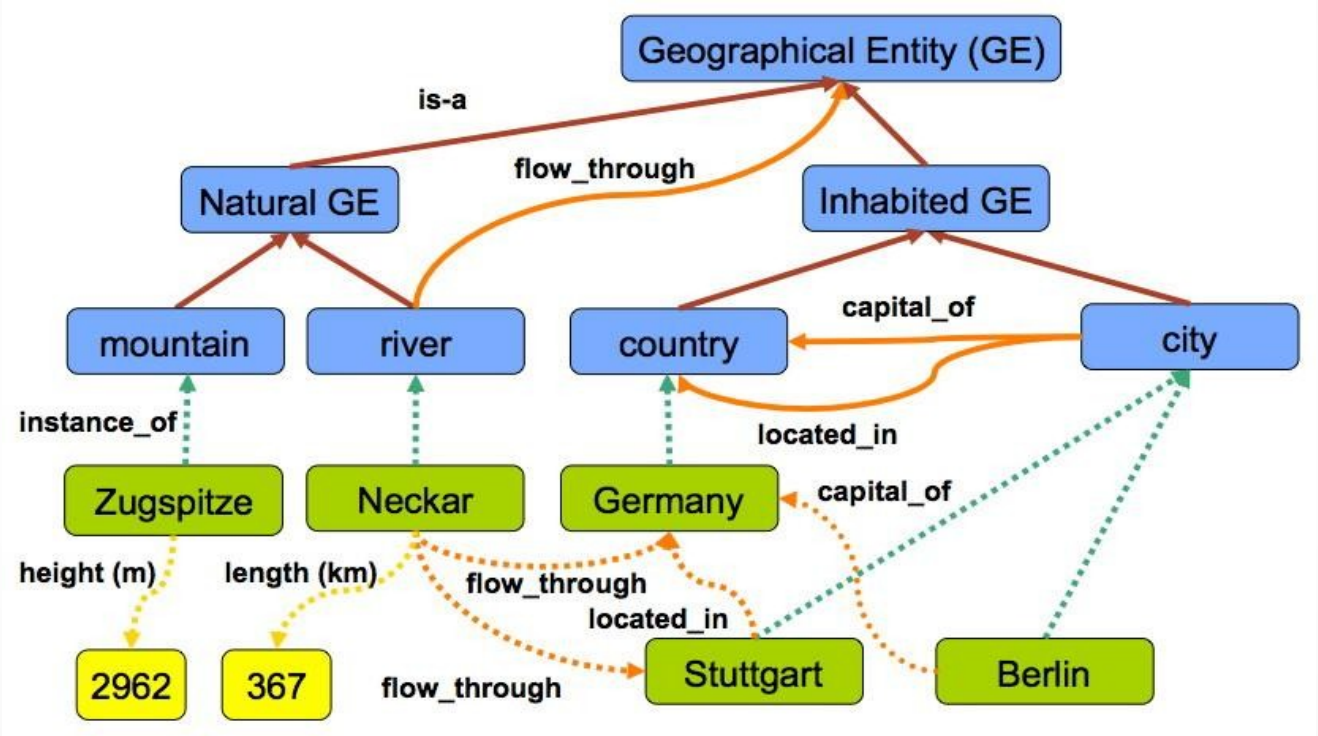
slide adapted from Doug Appelt

Relation types: Freebase

23 Million Entities, thousands of relations

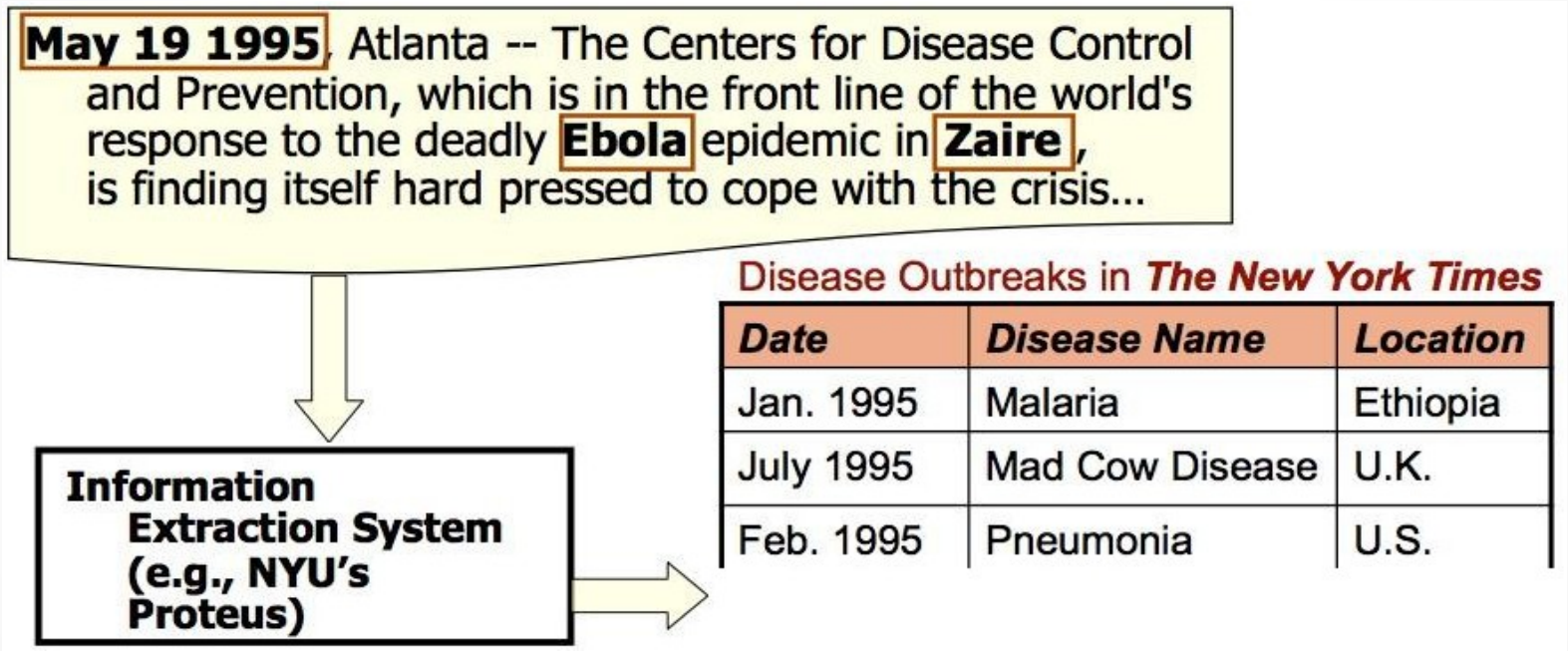
Relation name	Size	Example
/people/person/nationality	281,107	John Dugard, South Africa
/location/location/contains	253,223	Belgium, Nijlen
/people/person/profession	208,888	Dusa McDuff, Mathematician
/people/person/place_of_birth	105,799	Edwin Hubble, Marshfield
/dining/restaurant/cuisine	86,213	MacAyo's Mexican Kitchen, Mexican
/business/business_chain/location	66,529	Apple Inc., Apple Inc., South Park, NC
/biology/organism_classification_rank	42,806	Scorpaeniformes, Order
/film/film/genre	40,658	Where the Sidewalk Ends, Film noir
/film/film/language	31,103	Enter the Phoenix, Cantonese
/biology/organism_higher_classification	30,052	Calopteryx, Calopterygidae
/film/film/country	27,217	Turtle Diary, United States
/film/writer/film	23,856	Irving Shulman, Rebel Without a Cause
/film/director/film	23,539	Michael Mann, Collateral
/film/producer/film	22,079	Diane Eskenazi, Aladdin
/people/deceased_person/place_of_death	18,814	John W. Kern, Asheville
/music/artist/origin	18,619	The Octopus Project, Austin
/people/person/religion	17,582	Joseph Chartrand, Catholicism
/book/author/works_written	17,278	Paul Auster, Travels in the Scriptorium
/soccer/football_position/players	17,244	Midfielder, Chen Tao
/people/deceased_person/cause_of_death	16,709	Richard Daintree, Tuberculosis
/book/book/genre	16,431	Pony Soldiers, Science fiction
/film/film/music	14,070	Stavisky, Stephen Sondheim
/business/company/industry	13,805	ATS Medical, Health care

Relation types: geographical



slide adapted from Paul Buitelaar

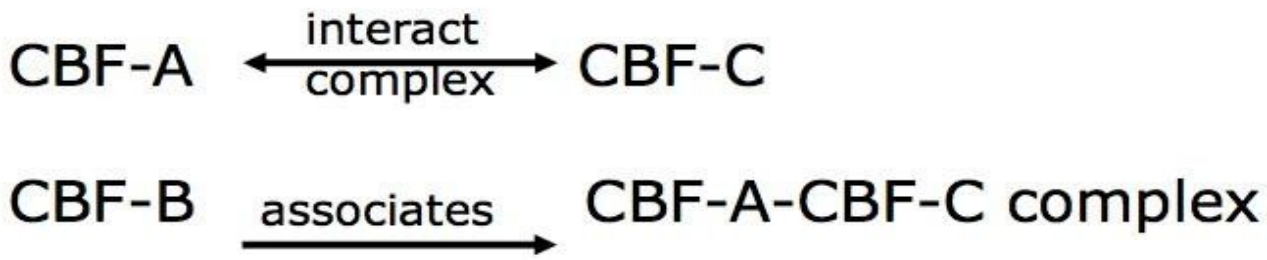
More relations: disease outbreaks



slide adapted from Eugene Agichtein

More relations: protein interactions

„We show that CBF-A and CBF-C interact with each other to form a CBF-A-CBF-C complex and that CBF-B does not interact with CBF-A or CBF-C individually but that it associates with the CBF-A-CBF-C complex."



slide adapted from Rosario & Hearst

Relations between word senses

- **NLP applications need word meaning!**
  - **Question answering**
  - **Conversational agents**
  - **Summarization**
- **One key meaning component: word relations**
  - **Hyponymy: San Francisco is an instance of a city**
  - **Antonymy: acidic is the opposite of basic**
  - **Meronymy: an alternator is a part of a car**

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 (Open IE)

Hand-written patterns

Patterns for learning hyponyms

- **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?**

Patterns for learning hyponyms

- **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 lexico-syntactic patterns

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…

Examples of the Hearst patterns

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 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

Supervised Methods

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

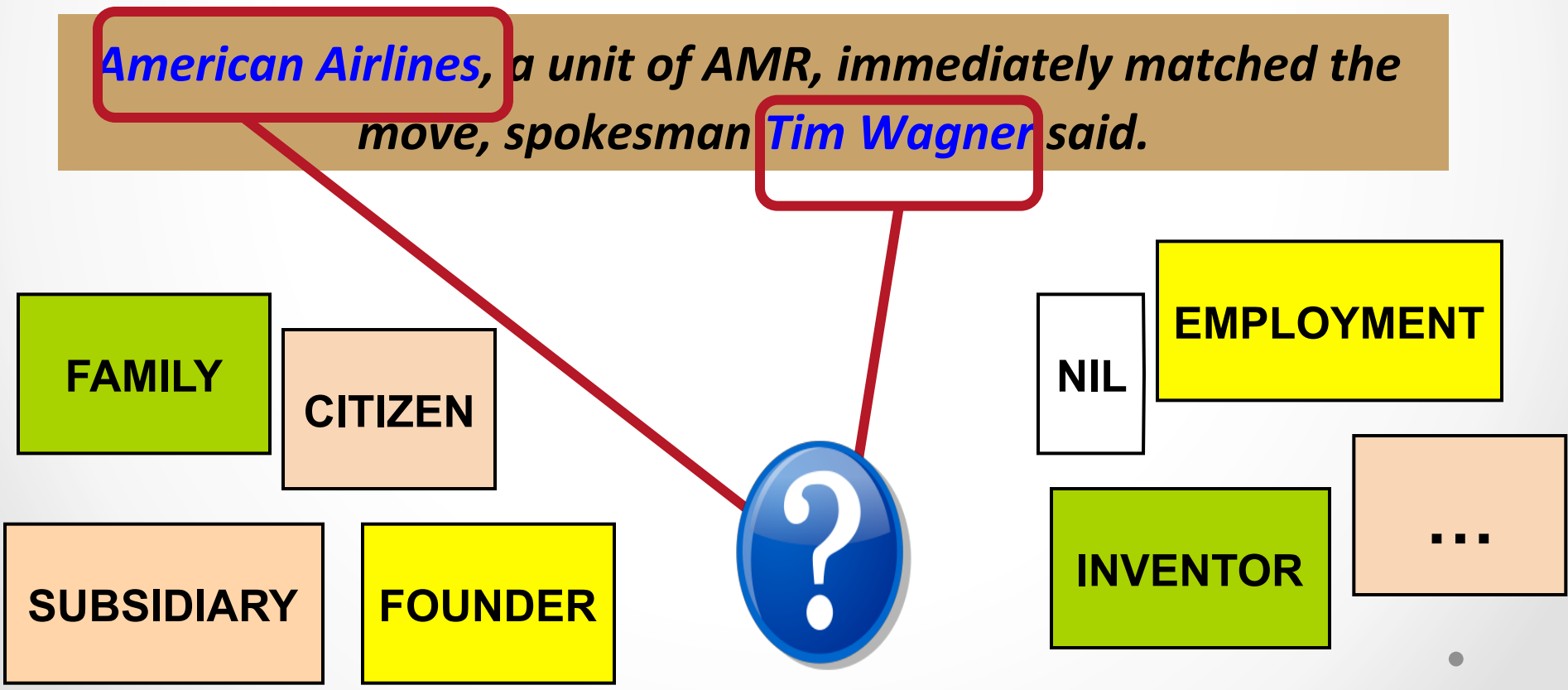
1. Find all pairs of named entities (usually in same sentence)
2. Decide if 2 entities are related
3. If yes, classify the relation

- Why the extra step?
  - Faster classification training by eliminating most pairs
  - Can use distinct feature-sets appropriate for each task.

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 **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*

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]

Features: syntactic features

Features of mention dependencies

ET1DW1 = ORG:Airlines

H1DW1 = matched:Airlines

ET2DW2 = PER:Wagner

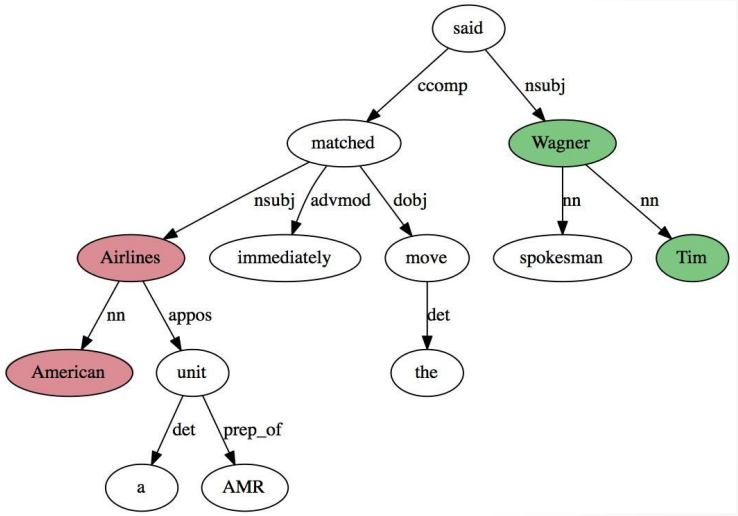
H2DW2 = said:Wagner

Features describing entity types and dependency tree

ET12SameNP = ORG-PER-false

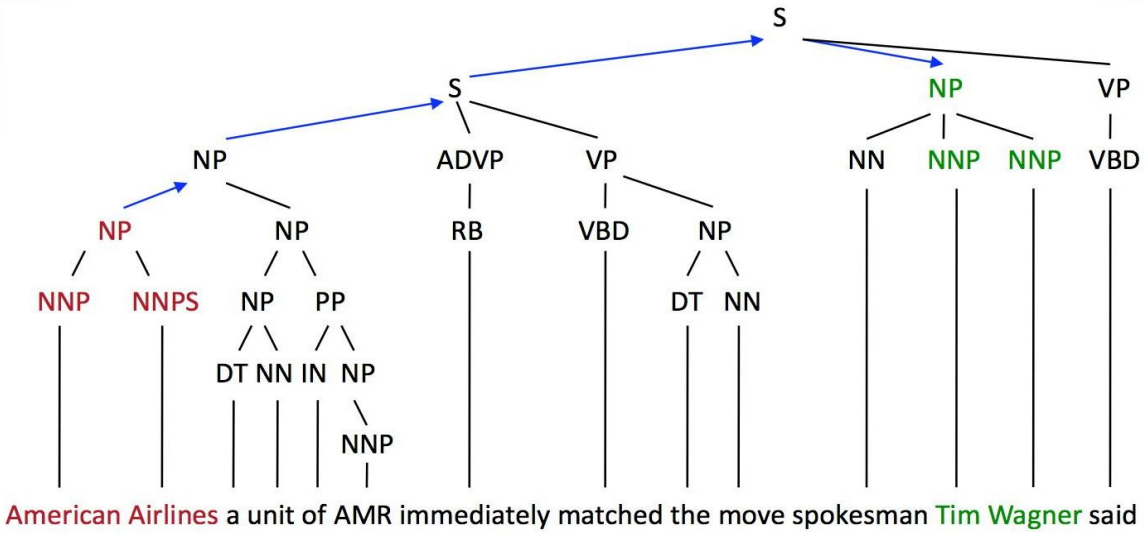
ET12SamePP = ORG-PER-false

ET12SameVP = ORG-PER-false



These features may have a small impact!

Features: syntactic features



Phrase label paths

PTP = [NP, S, NP]

PTPH = [NP:Airlines, S:matched, NP:Wagner]

These features may have a small impact!

Parse Features for Relation Extraction

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

Mention 1 **Mention 2**

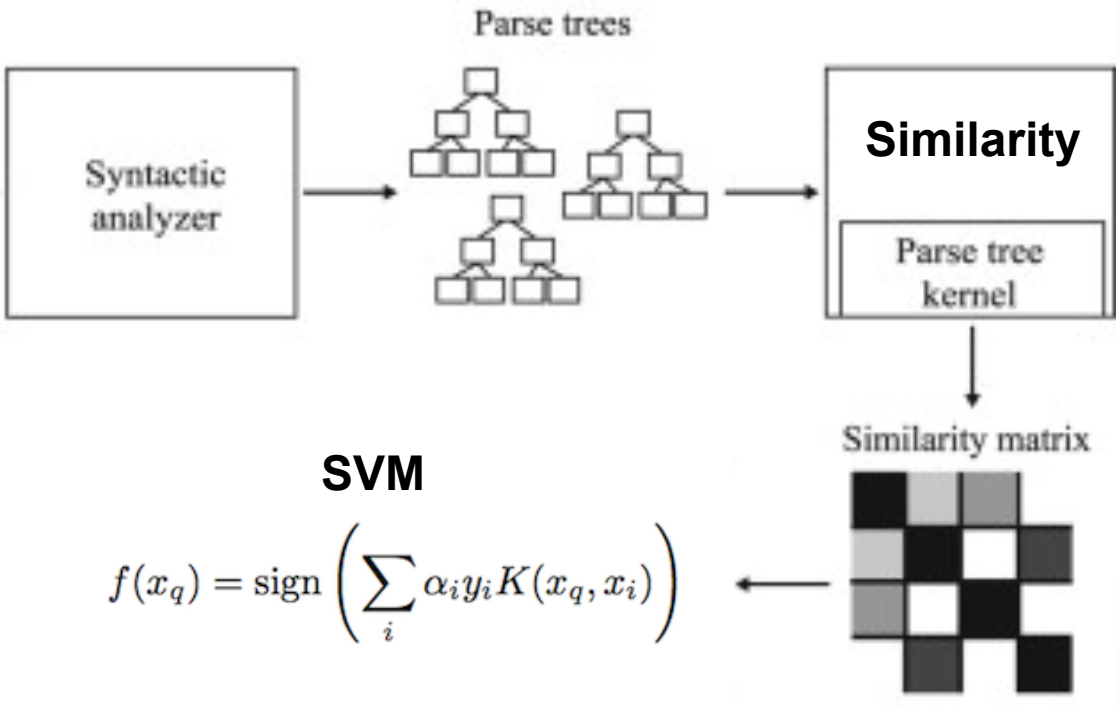
- Base syntactic chunk sequence from one to the other

  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

Side note: Graph/Tree Kernels



Gazeteer and trigger word features for relation extraction

- Trigger list for family: kinship terms
  - parent, wife, husband, grandparent, etc. [from WordNet]
- Gazeteer:
  - 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_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
  - 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 don't generalize well to different genres

Semi-supervised and Unsupervised Methods

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 (Brin 1998)

Extract (author, book) pairs

Start with these 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

Snowball

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

- 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
  - And compute a confidence for each pattern

.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

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

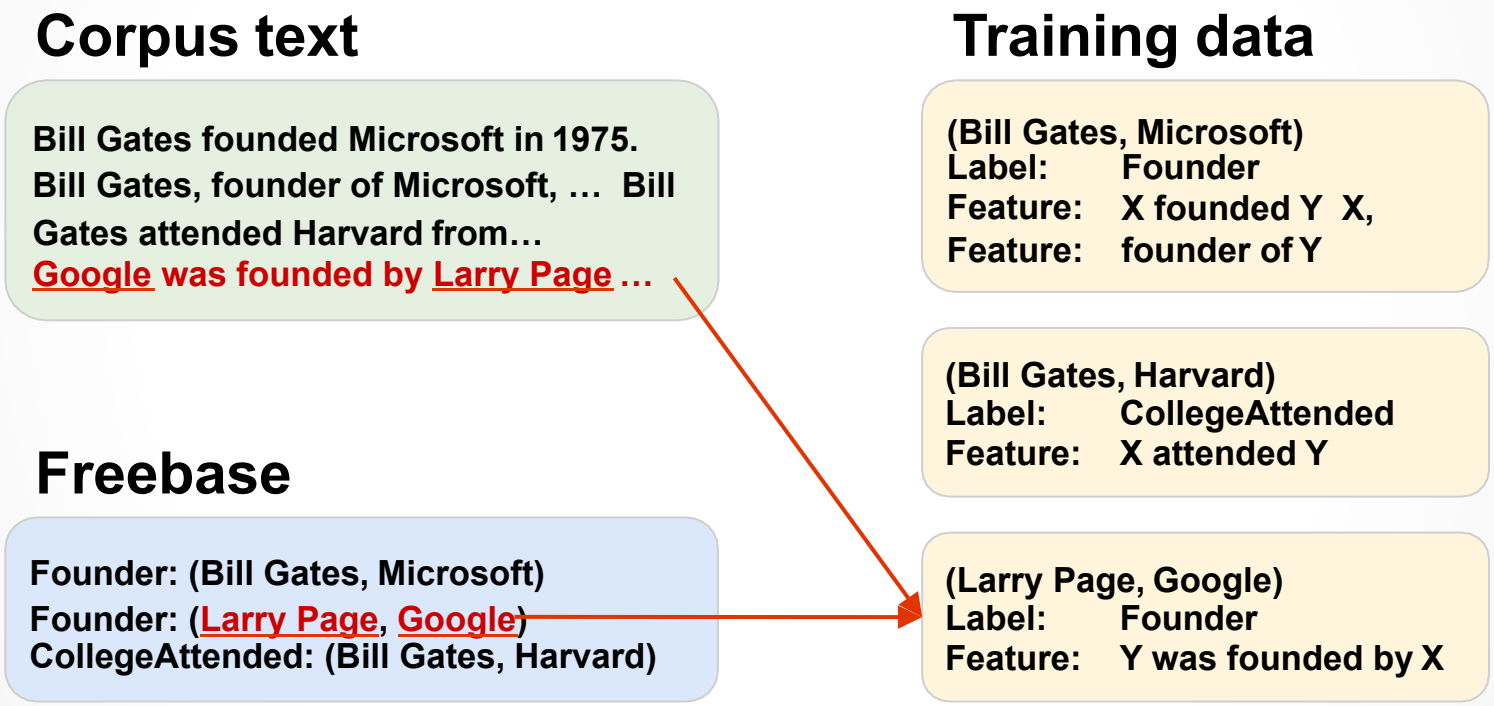
4. Extract frequent features (parse, words, etc)

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

5. Train supervised classifier using thousands of patterns

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

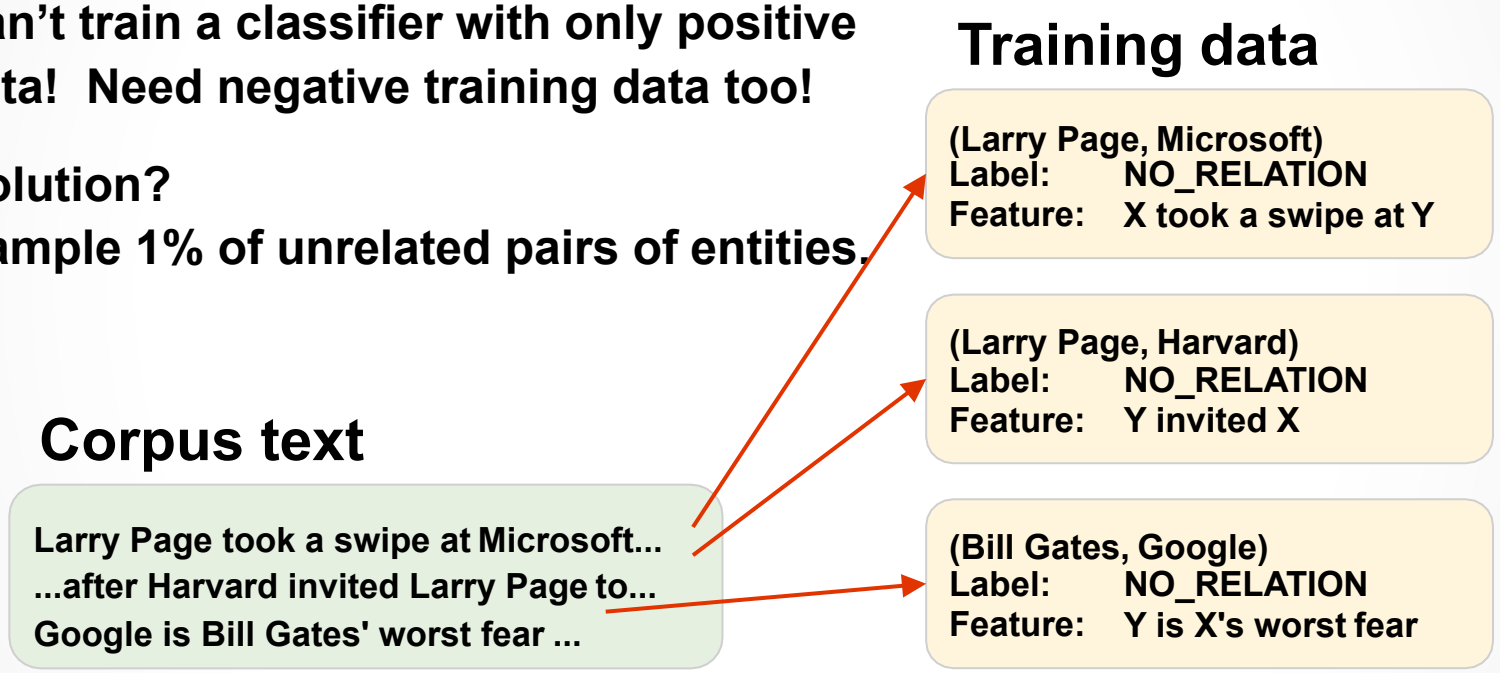
Collecting training data



Negative training data

Can't train a classifier with only positive data! Need negative training data too!

Solution?
Sample 1% of unrelated pairs of entities.



Unsupervised relation extraction (Open IE)

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)

Step 1: Self-supervised learner

- **Run a parser over 2000 sentences**
  - **Parsing is relatively expensive, so can't run on whole web**
  - **For each pair of base noun phrases NP_i and NP_j**
  - **Extract all tuples t = (NP_i, $relation_{i,j}$, NP_j)**

- **Label each tuple based on features of parse:**
  - **Positive iff the dependency path between the NPs is short, and doesn't cross a clause boundary, and neither NP is a pronoun**

- **Now train a Naïve Bayes classifier on the labeled tuples**
  - **Using *lightweight* features like POS tags nearby, stop words, etc.**

Step 2: Single-pass extractor

- **Over a huge (web-sized) corpus:**
  - **Run a dumb POS tagger**
  - **Run a dumb Base Noun Phrase chunker**
  - **Extract all text strings between base NPs**
  - **Run heuristic rules to simplify text strings**

    Scientists** from many universities **are** intently **studying** **stars

    → ⟨***scientists**, **are studying**, **stars***⟩

- **Pass candidate tuples to Naïve Bayes classifier**

- **Save only those predicted to be "trustworthy"**

Step 3: Redundancy-based assessor

- **Collect counts for each simplified tuple**
  ⟨*scientists*, *are studying*, *stars*⟩ → 17

- **Compute likelihood of each tuple**
  - **given the counts for each relation**
  - **and the number of sentences**
  - **and a combinatoric balls-and-urns model** [Downey et al. 05]

$$P(x \in C | x \text{ appears } k \text{ times in } n \text{ draws}) \approx \frac{1}{1 + \frac{|E|}{|C|}\left(\frac{p_E}{p_C}\right)^k e^{n(p_C - p_E)}}$$

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