ANLP

Week 4

(D. Jurafsky and C. Manning)

Information Extraction

- Finding structured information from unstructured (or lightly structured) text.

Information Extraction

- Information extraction (IE) systems
  - Find and understand limited relevant parts of texts
  - Gather information from many pieces of text
  - Produce a structured representation of relevant information:
    - *relations* (in the database sense), a.k.a.,
    - a *knowledge base*
  - Goals:
    1. Organize information so that it is useful to people
    2. Put information in a semantically precise form that allows further inferences to be made by computer algorithms

Information Extraction (IE)

- IE systems extract clear, factual information
  - Roughly: *Who did what to whom when?*
- E.g.,
  - Gathering earnings, profits, board members, headquarters, etc. from company reports
    - The headquarters of BHP Billiton Limited, and the global headquarters of the combined BHP Billiton Group, are located in Melbourne, Australia.
    - headquarters("BHP Biliton Limited", "Melbourne, Australia")
  - Learn drug-gene product interactions from medical research literature

Example

WASHINGTON — Senate Republicans on Sunday kept up the drumbeat of blame against President Obama for what they say is his failure to negotiate with them on the fiscal crisis that will come to a head on Thursday, when the government will run out of money to pay its bills. As the Republicans pointed fingers at the White House, Senators Harry Reid and Mitch McConnell were set to sit down again on Sunday in an effort to come up with some sort of agreement — even one that will kick the most pressing problems down the road for a few weeks or months.

- Names: "Senate Republicans", "President Obama", "the Republicans", "the White House", "Senators Harry Reid", "Mitch McConnell"
- Entity Linking: e1={"Senate Republicans", "the Republicans"}, e2={"President Obama", "his", "the White House"}
- Title: title("President", "Obama"), title("Senator", "Harry Reid"), title("Senator", "Mitch McConnell")
- "Blame" Event: "X kept up the drumbeat of blame against Y", "X pointed fingers at Y".

http://www.nytimes.com/2013/10/14/us/politics/budget-and-debt-limit-debate.html?hp

Factoid Questions as Google Queries

Overview

- Name Tagging
  - sequence models for Name tagging
  - pattern learning
- co-reference resolution
- Slot Filling
- Bootstrapping
- Distant supervision

Name Tagging

- Identify the “Named Entities” in text.
  - Named Entity Recognition (NER)
- People
- Organizations including companies, teams, etc.
- Locations and/or Geo-political Entities (GPE)

Named Entity Recognition (NER)

- A very important sub-task: find and classify names in text, for example:

  - The decision by the independent MP Andrew Wilkie to withdraw his support for the minority Labor government sounded dramatic but it should not further threaten its stability. When, after the 2010 election, Wilkie, Rob Oakeshott, Tony Windsor and the Greens agreed to support Labor, they gave just two guarantees: confidence and supply.

Named Entity Recognition (NER)

- A very important sub-task: find and classify names in text, for example:

  - The decision by the independent MP Andrew Wilkie to withdraw his support for the minority Labor government sounded dramatic but it should not further threaten its stability. When, after the 2010 election, Wilkie, Rob Oakeshott, Tony Windsor and the Greens agreed to support Labor, they gave just two guarantees: confidence and supply.

Named Entity Recognition (NER)

- A very important sub-task: find and classify names in text, for example:

  - The decision by the independent MP Andrew Wilkie to withdraw his support for the minority Labor government sounded dramatic but it should not further threaten its stability. When, after the 2010 election, Wilkie, Rob Oakeshott, Tony Windsor and the Greens agreed to support Labor, they gave just two guarantees: confidence and supply.

Person
Date
Location
Organi-
zation

Named Entity Recognition (NER)

- The uses:
  - Named entities can be indexed, linked, etc.
  - Sentiment can be attributed to companies or products
  - A lot of IE relations are associations between named entities
  - For question answering, answers are often named entities.

- Concretely:
  - Many web pages tag various entities, with links to bio or topic pages, etc.
    - Reuters' OpenCalais, Evri, AlchemyAPI, Yahoo's Term Extraction, …
  - Apple/Google/Microsoft/… smart recognizers for document content

Named Entity Recognition (NER)

- Initially, NAMED entities: People names, Locations, Company names, etc.
- What it is the first, most straight-forward feature?

- Now, more general: protein names, drug names, etc.

The Named Entity Recognition Task ~ sequence labeling

Task: Predict entities in a text

Foreign	ORG
Ministry	ORG
spokesman	O
Shen	PER
Guofang	PER
told	O
Reuters	ORG
:	:

Standard evaluation is per entity, *not* per token

Relation Extraction

What is relation extraction?

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

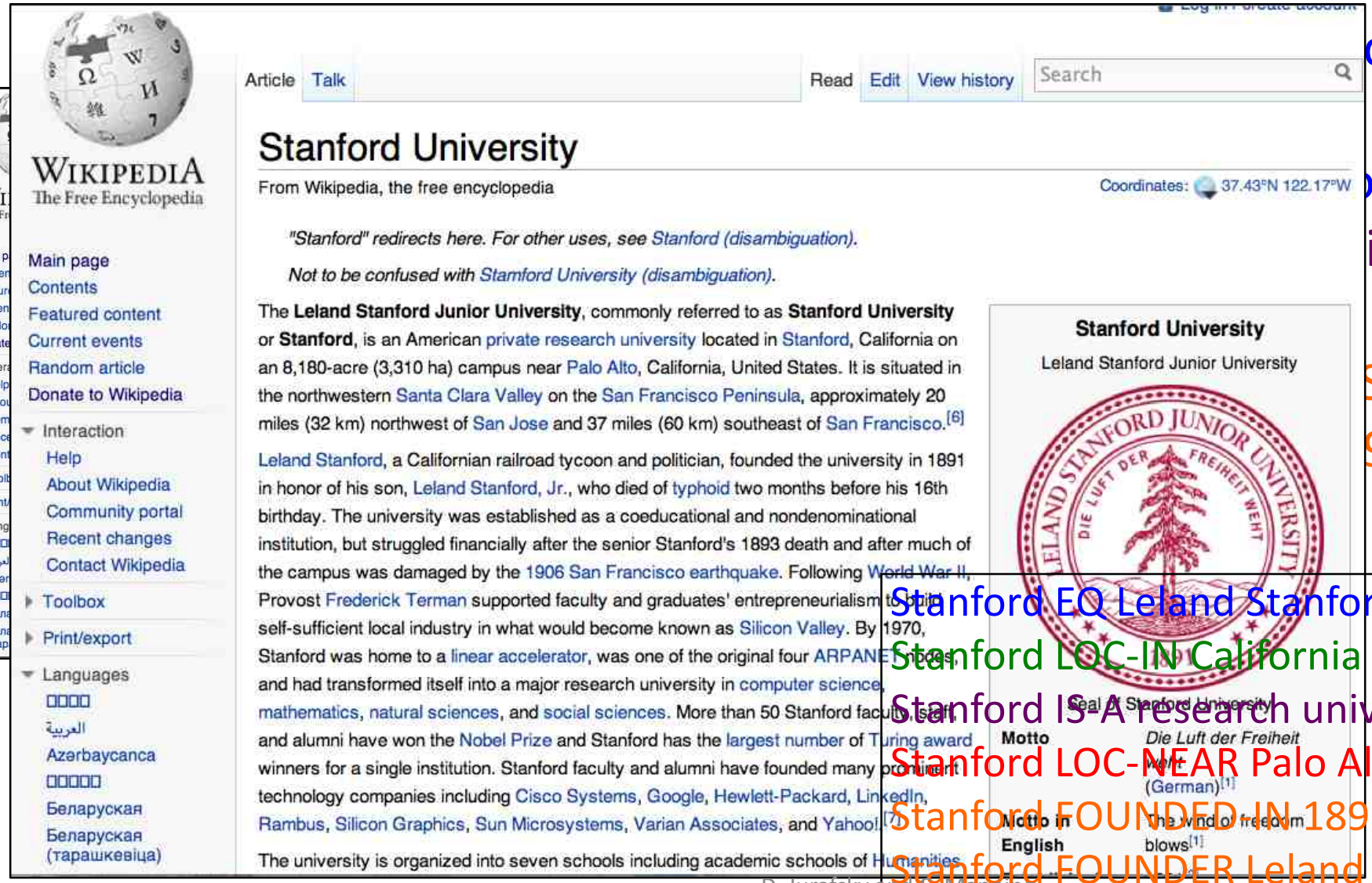
Company	IBM
Location	New York
Date	June 16, 1911
Original-Name	Computing-Tabulating-Recording Co.

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

Founding-year(IBM,1911)

Founding-location(IBM,New York)

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

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.”?

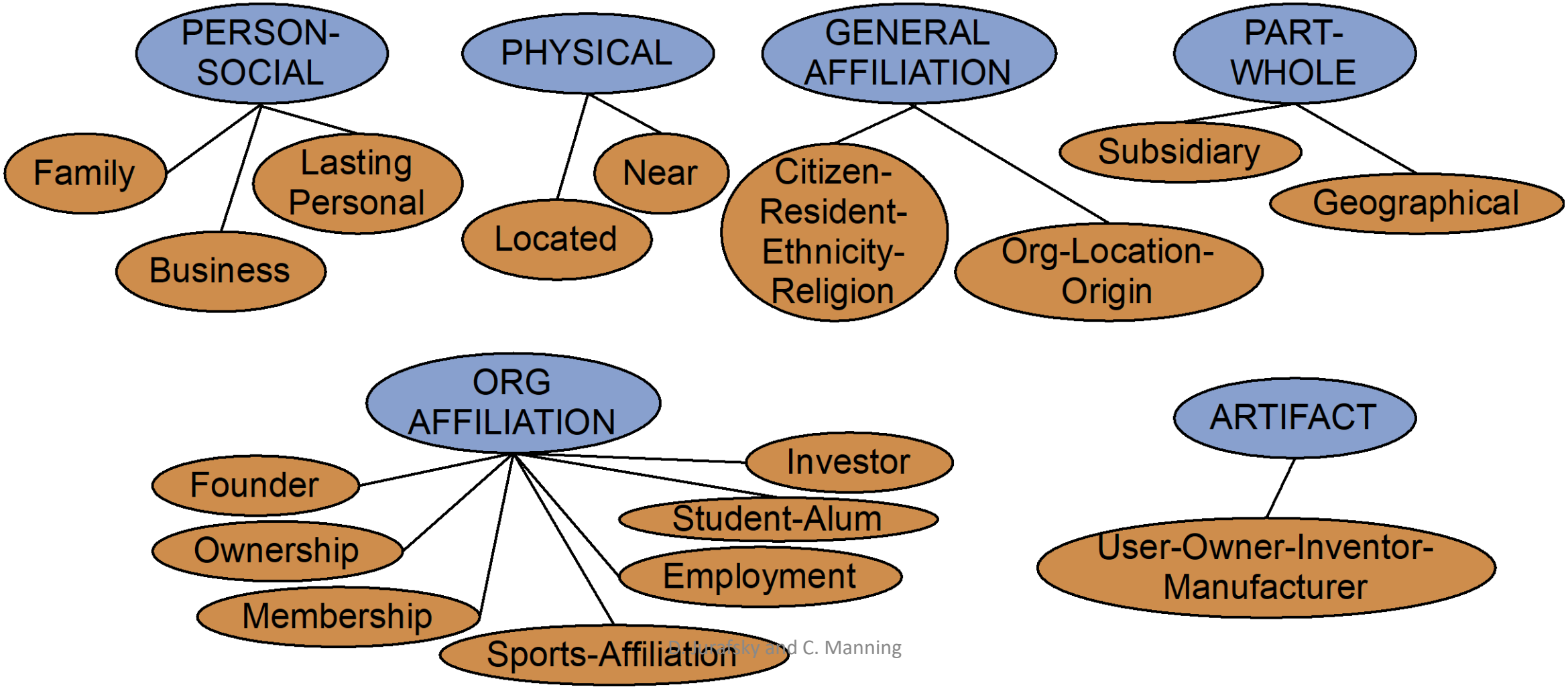
```
(acted-in ?x “E.T.”)(is-a ?y actor)(granddaughter-of ?x ?y)
```

- But which relations should we extract?

Automated Content Extraction (ACE)

17 relations from 2008 “Relation Extraction Task”

Does it remind anything to the CS people?



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

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

Extracting UMLS relations from a sentence

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

```
{{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 v
name="casper">{{cite speech|title=Die Lu
Casper|first=Gerhard|last=Casper|author
05|url=http://www.stanford.edu/dept/pr
|mottoeng = The wind of freedom blows<
|established = 1891<ref>{{cite web |
url=http://www.stanford.edu/home/stan
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"

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,
  people/person/profession,
  biology/organism_higher_classification

  location/location/contains
  people/person/place-of-birth
  film/film/genre

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`

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

What is relation extraction?

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

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

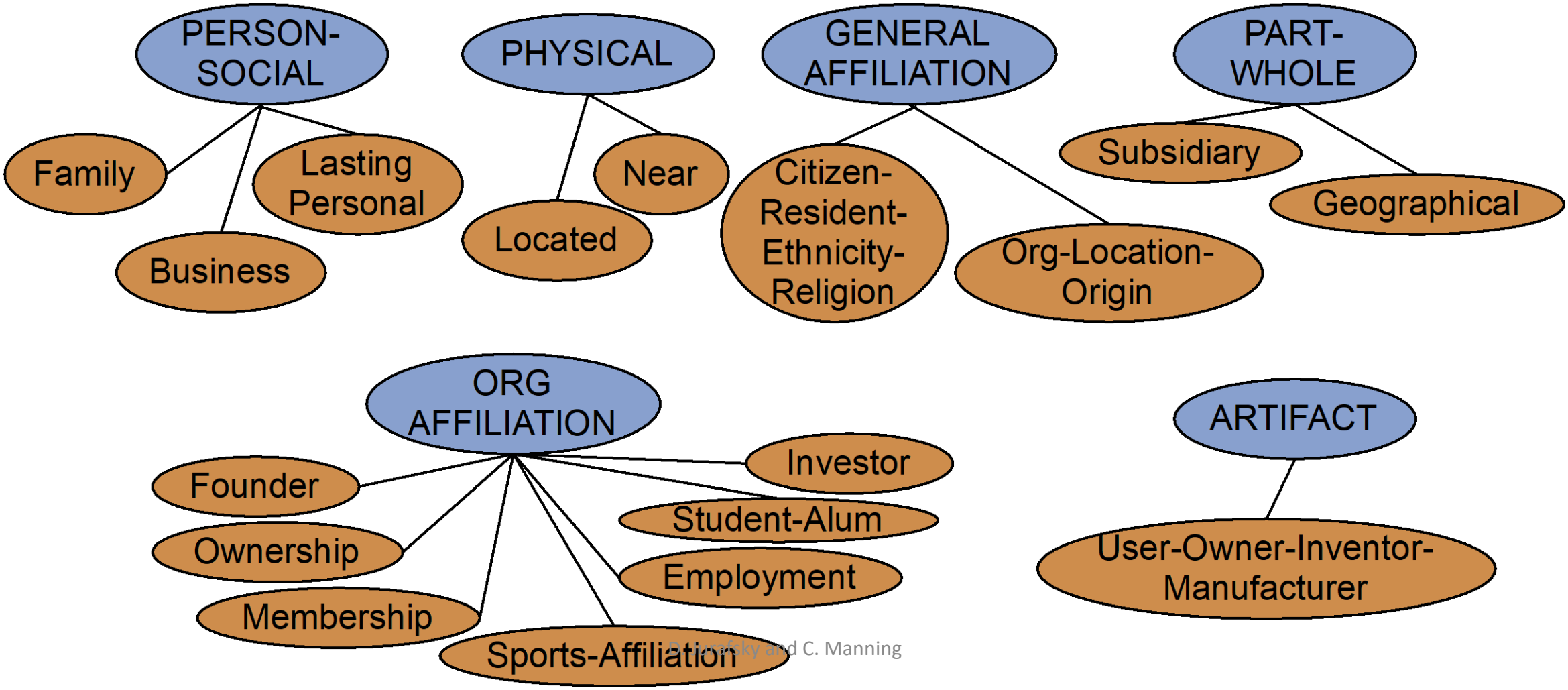
- George Marshall was named US Secretary of State

Hand-built patterns 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

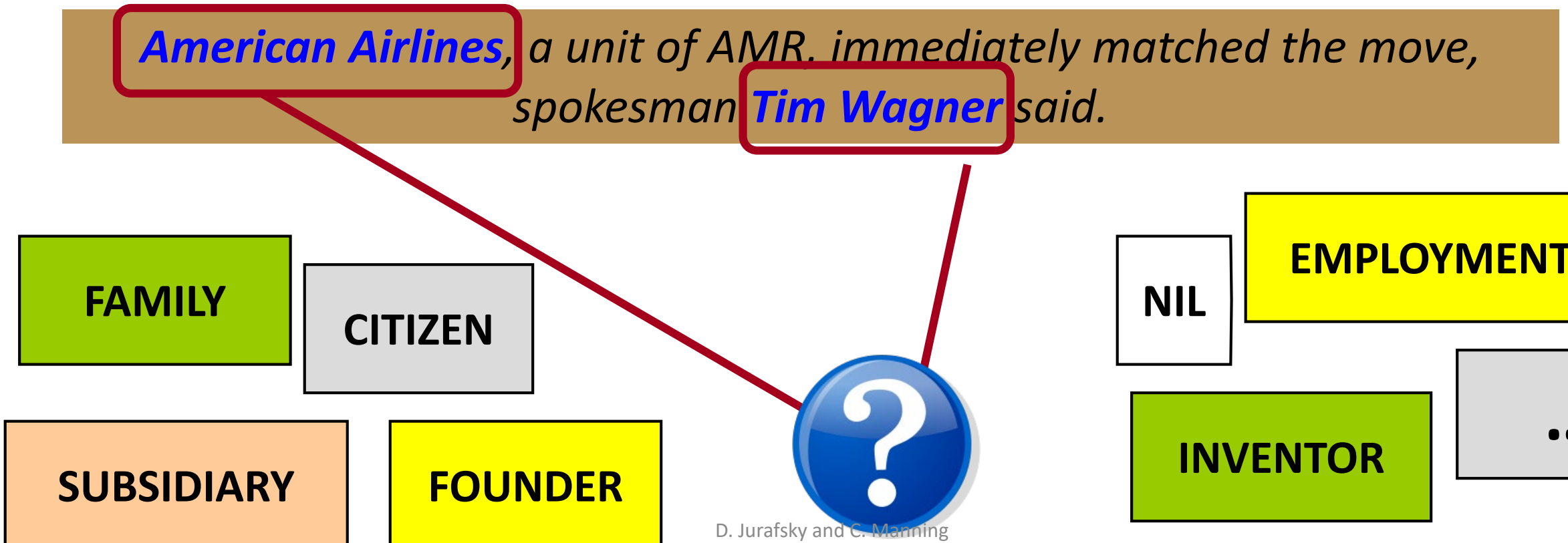
Automated Content Extraction (ACE)

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



Relation Extraction

Classify the relation between two entities in a sentence



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}

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

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

Brin, Sergei. 1998. Extracting Patterns and Relations from the World Wide Web.

- 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

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

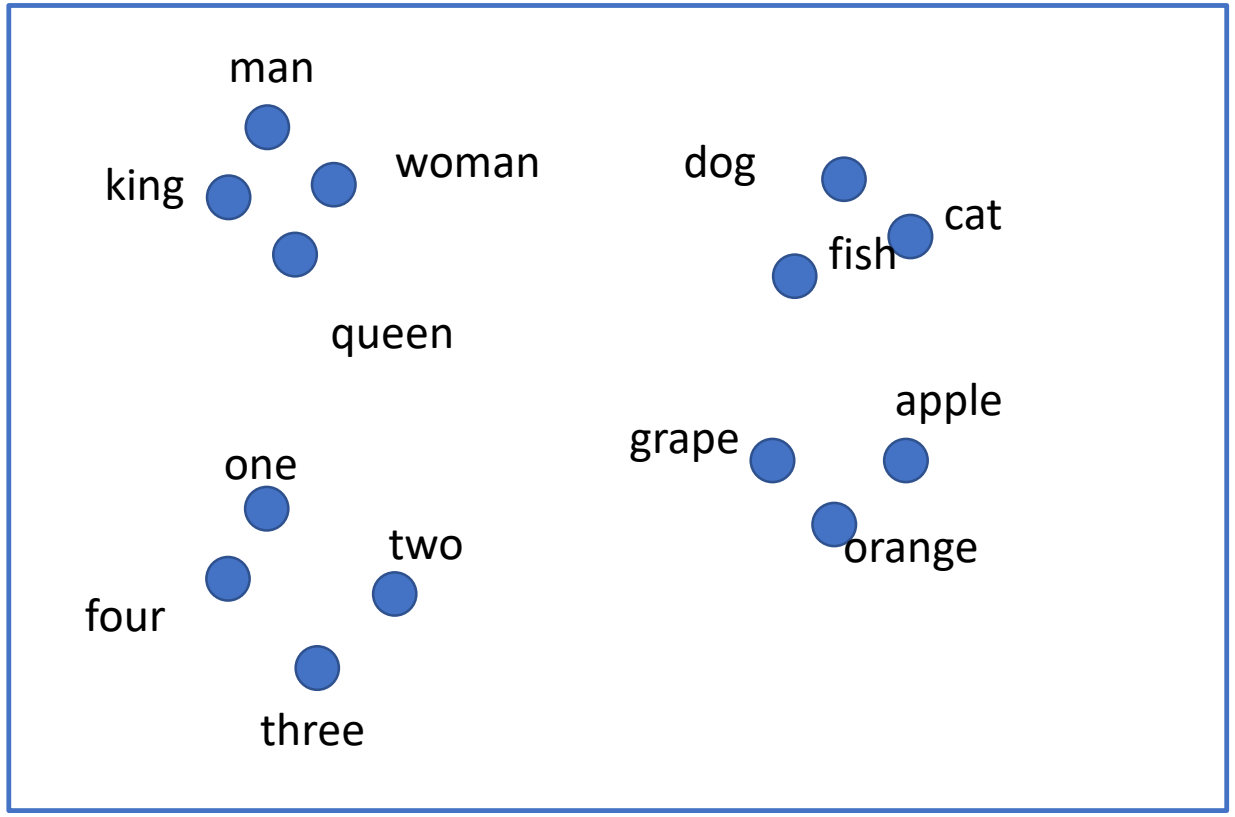
Featurized Representation: Word Embeddings (Andrew Ng)

300 features

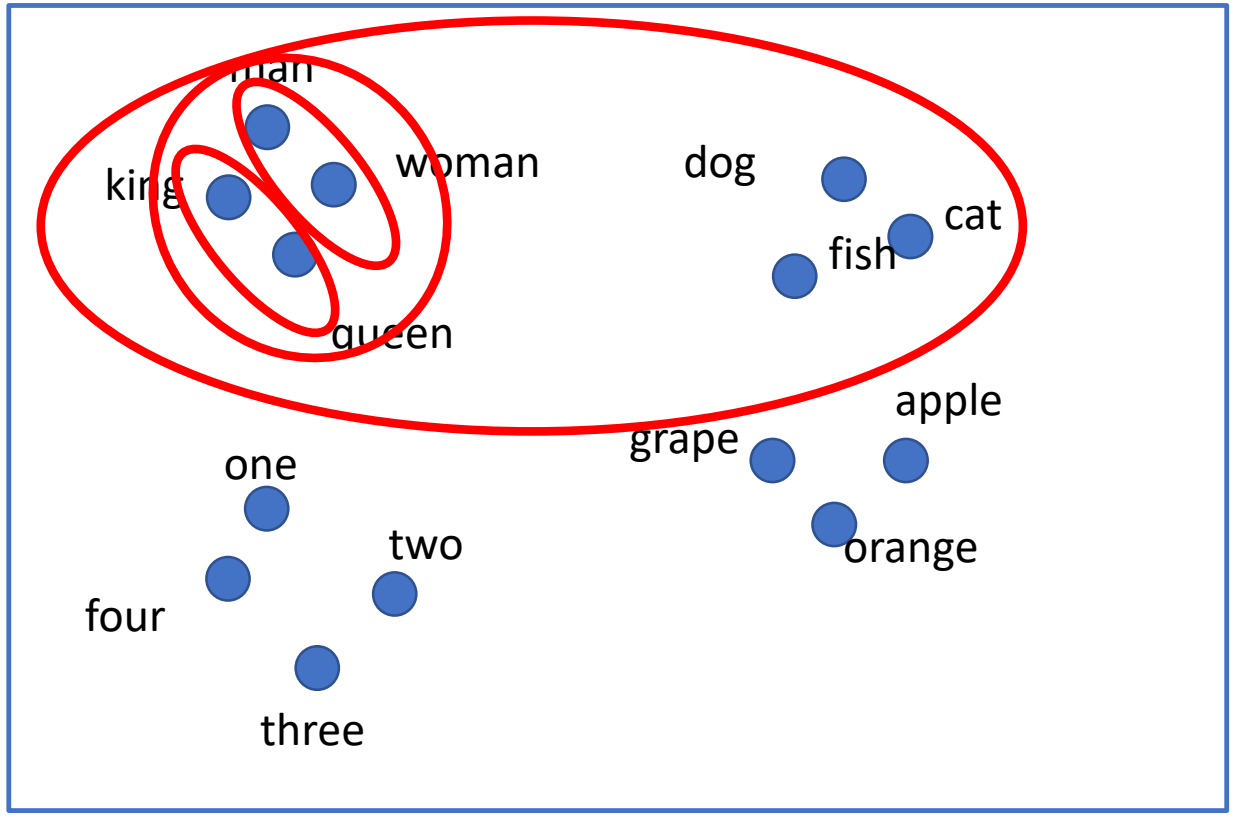
	Man	Woman	King	Queen	Apple	Orange
Gender	-1	1	-0.95	0.97	0.00	0.01
Royal	0.01	0.02	0.93	0.95	-0.01	0.00
Age	0.03	0.02	0.7	0.69	0.03	-0.02
Food	0.04	0.01	0.02	0.01	0.95	0.97
…						
….						

I want a glass of orange _____
I want a glass of apple _____

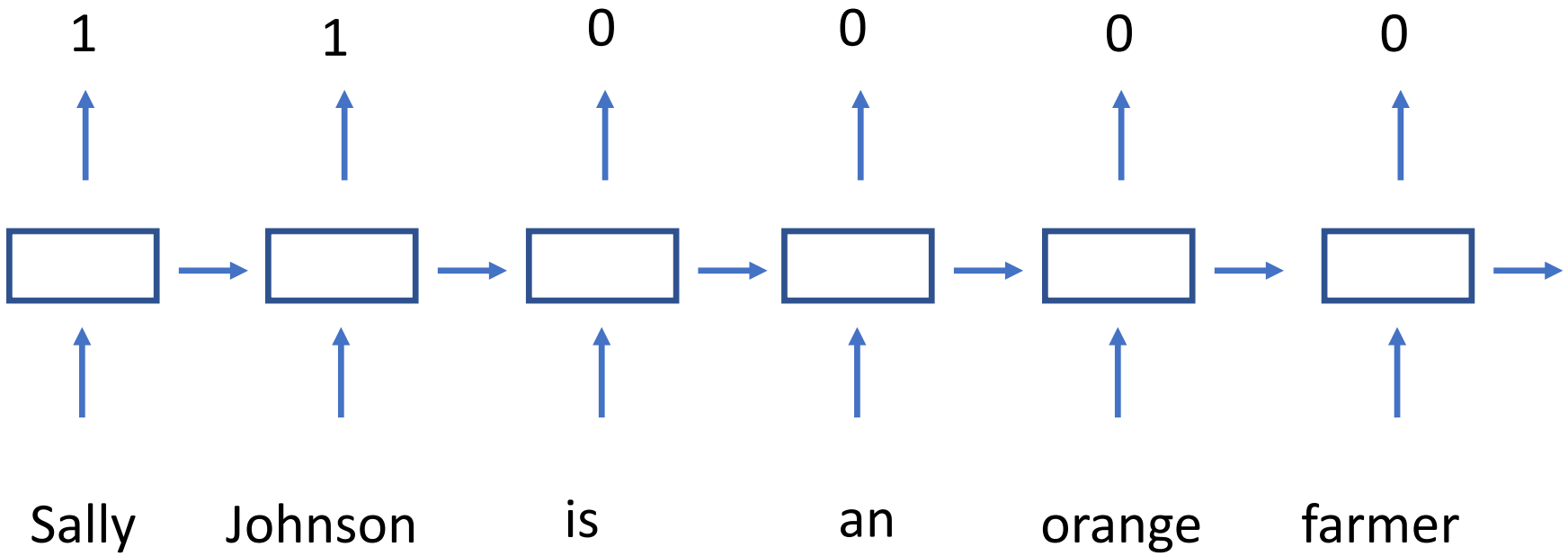
Visualizing word embeddings (Andrew Ng)



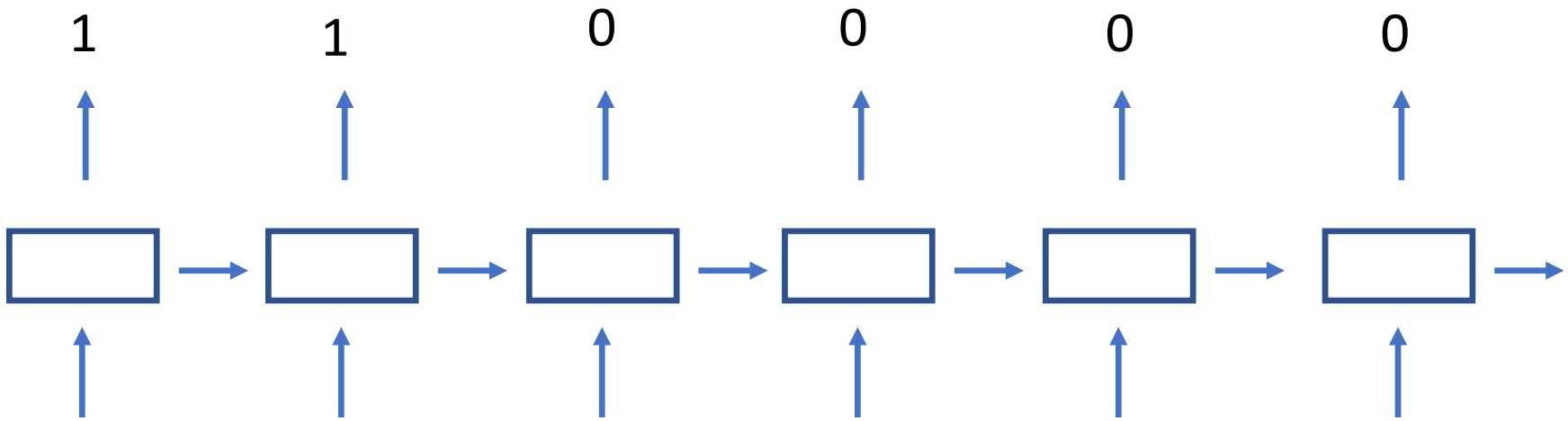
Visualizing word embeddings (Andrew Ng)



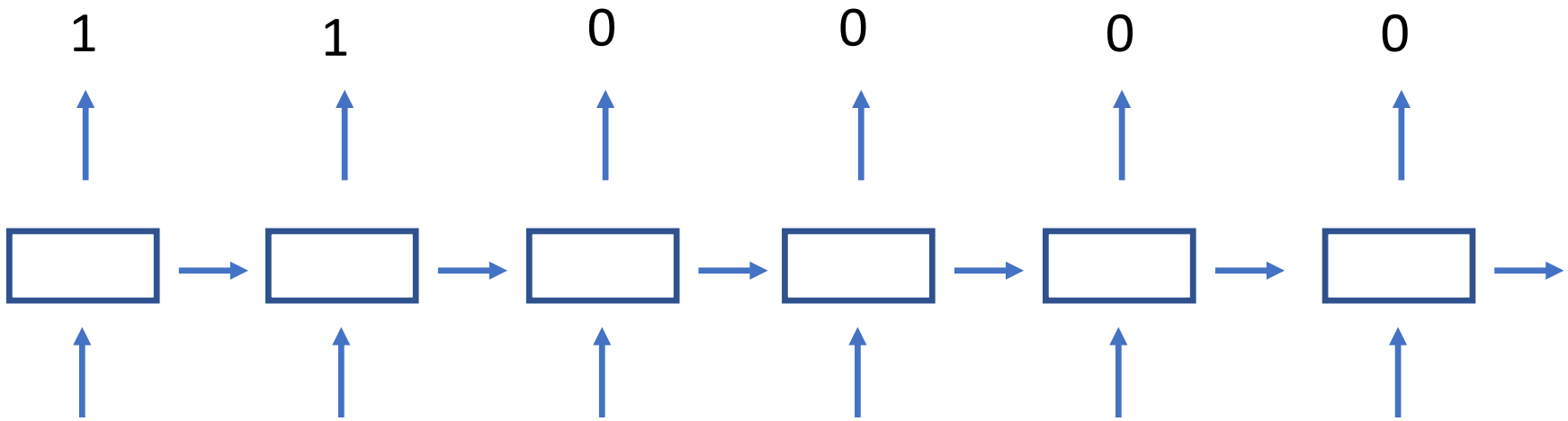
NE recognition (Andrew Ng)



NE recognition (Andrew Ng)

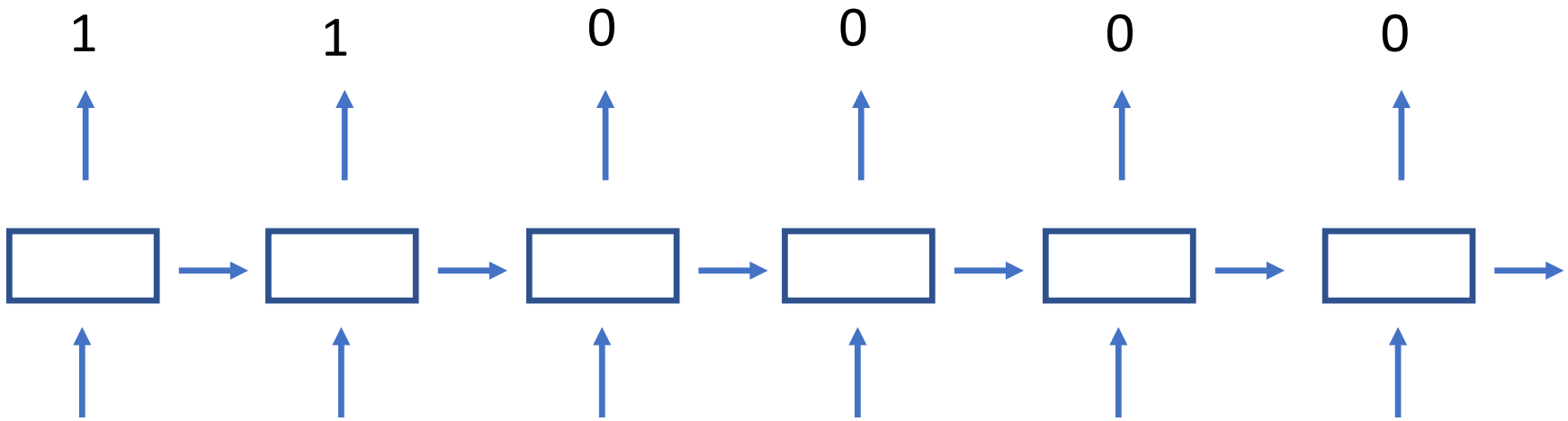


NE recognition (Andrew Ng)



Sally	Johnson	is	an	orange	farmer
Robert	Lin	is	~~an~~	~~apple~~	~~farmer~~
			a	durian	cultivator

NE recognition (Andrew Ng)



Sally Johnson is an orange farmer

Robert Lin is an apple farmer

a durian cultivator

fruits

Transfer Learning and word embeddings (Andrew Ng)

- Learn word embeddings from large corpora (1 -100B words)
  - Download pre-trained embeddings online
- Transfer embeddings to new task with smaller training set
- Optional: continue to finetune the word embeddings with new data (only if the new training set is big enough)

Analogies (Andrew Ng)

300 features

	Man	Woman	King	Queen	Apple	Orange
Gender	-1	1	-0.95	0.97	0.00	0.01
Royal	0.01	0.02	0.93	0.95	-0.01	0.00
Age	0.03	0.02	0.7	0.69	0.03	-0.02
Food	0.04	0.01	0.02	0.01	0.95	0.97

Man → Woman as King → ????

$e_{Man} - e_{Woman}$

$e_{king} - e_{?}$

$e_{Man} - e_{Woman} \sim \begin{vmatrix} -2 \\ 0 \\ 0 \\ 0 \end{vmatrix}$ $e_{king} - e_{queen} \sim \begin{vmatrix} -2 \\ 0 \\ 0 \\ 0 \end{vmatrix}$