

# Machine Learning for NLP:

—  
Advanced

Tasks & Algorithms

+Slides credit: Dan Jurafsky

# — Text Classification

# Positive or negative movie review?



- unbelievably disappointing



- Full of zany characters and richly applied satire, and some great plot twists



- this is the greatest screwball comedy ever filmed

- It was pathetic. The worst part about it was the boxing scenes.



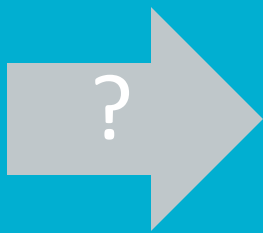
# What is the subject of this article?

## MEDLINE Article



## MeSH Subject Category Hierarchy

- Antagonists and Inhibitors
- Blood Supply
- Chemistry
- Drug Therapy
- Embryology
- Epidemiology
- ...



# Text Classification

- Assigning subject categories, topics, or genres
- Spam detection
- Authorship identification
- Age/gender identification
- Language Identification
- Sentiment analysis
-

# Text Classification: definition

- *Input:*
  - a document  $d$
  - a fixed set of classes  $C = \{c_1, c_2, \dots, c_j\}$
- *Output:* a predicted class  $c \in C$

# Classification Methods: Hand-coded rules

- Rules based on combinations of words or other features
  - spam: black-list-address OR (“dollars” AND “have been selected”)
- Accuracy can be high
  - If rules carefully refined by expert
- But building and maintaining these rules is expensive

# Classification Methods: Supervised Machine Learning

- *Input:*

- a document  $d$
- a fixed set of classes  $C = \{c_1, c_2, \dots, c_j\}$
- A training set of  $m$  hand-labeled documents  $(d_1, c_1), \dots, (d_m, c_m)$

- *Output:*

- a learned classifier  $\gamma: d \mapsto c$



# Classification Methods: Supervised Machine Learning

- Any kind of classifier
  - Naïve Bayes
  - Logistic regression
  - Support-vector machines
  - k-Nearest Neighbors
  - Neural networks
  - Pre-trained transformers
  - ...

# The bag of words representation

Y(

I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet.

) = C



# The bag of words representation

Y(

I love this movie! It's sweet,  
but with satirical humor. The  
dialogue is great and the  
adventure scenes are fun... It  
manages to be whimsical and  
romantic while laughing at the  
conventions of the fairy tale  
genre. I would recommend it to  
just about anyone. I've seen  
it several times, and I'm  
always happy to see it again  
whenever I have a friend who  
hasn't seen it yet.

) = C



# The bag of words representation: using a subset of words

Y(

```
x love xxxxxxxxxxxxxxxxxxxx sweet
xxxxxxxx satirical xxxxxxxxxxxx
xxxxxxxxxxxx great xxxxxxxx
xxxxxxxxxxxxxxxxxxxxxxxxxxxx fun xxxx
xxxxxxxxxxxxxxxx whimsical xxxx
romantic xxxx laughing
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
xxxxxxxxxxxxxxxx recommend xxxxxx
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
xx several xxxxxxxxxxxxxxxxxxxx
xxxxx happy xxxxxxxx again
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
xxxxxxxxxxxxxxxxxxxxxxxxxxxx
```

) = C



# The bag of words representation

Y(

great	2
love	2
recommend	1
laugh	1
happy	1
...	...

) = C



# Bag of words for document classification

Test document

parser  
language  
label  
translation  
...

?

Machine Learning

learning  
training  
algorithm  
shrinkage  
network...

NLP

parser  
tag  
training  
translation  
language...

Garbage Collection

garbage  
collection  
memory  
optimization  
region...

Planning

planning  
temporal  
reasoning  
plan  
language...

GUI

...

# Example classifier: Naïve Bayes

## Bayes' Rule Applied to Documents and Classes

- For a document  $d$  and a class  $c$

$$P(c | d) = \frac{P(d | c)P(c)}{P(d)}$$

# Naïve Bayes Classifier

$$c_{MAP} = \operatorname{argmax}_{c \in C} P(c \mid d)$$

MAP is “maximum a posteriori” = most likely class

$$= \operatorname{argmax}_{c \in C} \frac{P(d \mid c)P(c)}{P(d)}$$

Bayes Rule

$$= \operatorname{argmax}_{c \in C} P(d \mid c)P(c)$$

Dropping the denominator



# Naïve Bayes Classifier

$$c_{MAP} = \operatorname{argmax}_{c \in C} P(d \mid c)P(c)$$

$$= \operatorname{argmax}_{c \in C} P(x_1, x_2, \dots, x_n \mid c)P(c)$$

Document  $d$   
represented as  
features  $x_1 \dots x_n$

# Multinomial Naïve Bayes Classifier

$$c_{MAP} = \operatorname{argmax}_{c \in C} P(x_1, x_2, \dots, x_n | c) P(c)$$

$$c_{NB} = \operatorname{argmax}_{c \in C} P(c_j) \prod_{x \in X} P(x | c)$$

# Multinomial Naïve Bayes: Learning

- From training corpus, extract *Vocabulary*
- Calculate  $P(c_j)$  terms
  - For each  $c_j$  in  $C$  do
    - $docs_j \leftarrow$  all docs with class  $= c_j$
$$P(c_j) \leftarrow \frac{|docs_j|}{|\text{total \# documents}|}$$
- Calculate  $P(w_k | c_j)$  terms
  - $Text_j \leftarrow$  single doc containing all  $docs_j$
  - For each word  $w_k$  in *Vocabulary*
    - $n_k \leftarrow$  # of occurrences of  $w_k$  in  $Text_j$
$$P(w_k | c_j) \leftarrow \frac{n_k + \alpha}{n + \alpha |Vocabulary|}$$

# Text Classification Evaluation

Precision, Recall, and the F measure

# The 2-by-2 contingency table

	correct	not correct
selected	tp	fp
not selected	fn	tn

# Precision and recall

- **Precision:** % of selected items that are correct  
**Recall:** % of correct items that are selected

	correct	not correct
selected	tp	fp
not selected	fn	tn

# A combined measure: F

- A combined measure that assesses the P/R tradeoff is F measure (weighted harmonic mean):

$$F = \frac{1}{\alpha \frac{1}{P} + (1-\alpha) \frac{1}{R}} = \frac{(\beta^2 + 1)PR}{\beta^2 P + R}$$

- The harmonic mean is a very conservative average; see IIR § 8.3
- People usually use balanced F1 measure
  - i.e., with  $\beta = 1$  (that is,  $\alpha = \frac{1}{2}$ ):  $F = 2PR/(P+R)$

# More Than Two Classes: Sets of binary classifiers

- Dealing with **any-of** or **multivalued** classification
  - A document can belong to 0, 1, or >1 classes.
- For each class  $c \in C$ 
  - Build a classifier  $\gamma_c$  to distinguish  $c$  from all other classes  $c' \in C$
- Given test doc  $d$ ,
  - Evaluate it for membership in each class using each  $\gamma_c$
  - $d$  belongs to **any** class for which  $\gamma_c$  returns true



# More Than Two Classes: Sets of binary classifiers

- One-of or multinomial classification
  - Classes are mutually exclusive: each document in exactly one class
- For each class  $c \in C$ 
  - Build a classifier  $\gamma_c$  to distinguish  $c$  from all other classes  $c' \in C$
- Given test doc  $d$ ,
  - Evaluate it for membership in each class using each  $\gamma_c$
  - $d$  belongs to the one class with maximum score

# Evaluation:

## Classic Reuters-21578 Data Set

- Most (over)used data set, 21,578 docs (each 90 types, 200 tokens)
- 9603 training, 3299 test articles (ModApte/Lewis split)
- 118 categories
  - An article can be in more than one category
  - Learn 118 binary category distinctions
- Average document (with at least one category) has 1.24 classes
- Only about 10 out of 118 categories are large

Common categories  
(#train, #test)

- |                            |                       |
|----------------------------|-----------------------|
| • Earn (2877, 1087)        | • Trade (369, 119)    |
| • Acquisitions (1650, 179) | • Interest (347, 131) |
| • Money-fx (538, 179)      | • Ship (197, 89)      |
| • Grain (433, 149)         | • Wheat (212, 71)     |
| • Crude (389, 189)         | • Corn (182, 56)      |

# Reuters Text Categorization data set (Reuters-21578) document

```
<REUTERS TOPICS="YES" LEWISSPLIT="TRAIN" CGISPLIT="TRAINING-SET" OLDID="12981"
NEWID="798">
```

```
<DATE> 2-MAR-1987 16:51:43.42</DATE>
```

```
<TOPICS><D>livestock</D><D>hog</D></TOPICS>
```

```
<TITLE>AMERICAN PORK CONGRESS KICKS OFF TOMORROW</TITLE>
```

```
<DATELINE> CHICAGO, March 2 - </DATELINE><BODY>The American Pork Congress kicks off tomorrow,
March 3, in Indianapolis with 160 of the nations pork producers from 44 member states determining industry positions
on a number of issues, according to the National Pork Producers Council, NPPC.
```

Delegates to the three day Congress will be considering 26 resolutions concerning various issues, including the future direction of farm policy and the tax law as it applies to the agriculture sector. The delegates will also debate whether to endorse concepts of a national PRV (pseudorabies virus) control and eradication program, the NPPC said.

A large trade show, in conjunction with the congress, will feature the latest in technology in all areas of the industry, the NPPC added. Reuter

```
&#3;</BODY></TEXT></REUTERS>
```

# Confusion matrix c

- For each pair of classes  $\langle c_1, c_2 \rangle$  how many documents from  $c_1$  were incorrectly assigned to  $c_2$ ?
  - $c_{3,2}$ : 90 wheat documents incorrectly assigned to poultry

Docs in test set	Assigned UK	Assigned poultry	Assigned wheat	Assigned coffee	Assigned interest	Assigned trade
True UK	95	1	13	0	1	0
True poultry	0	1	0	0	0	0
True wheat	10	90	0	1	0	0
True coffee	0	0	0	34	3	7
True interest	-	1	2	13	26	5
True trade	0	0	2	14	5	10

# Per class evaluation measures

## Recall:

Fraction of docs in class  $i$  classified correctly:

$$\frac{c_{ii}}{\sum_j c_{ij}}$$

## Precision:

Fraction of docs assigned class  $i$  that are actually about class  $i$ :

$$\frac{c_{ii}}{\sum_j c_{ji}}$$

## Accuracy: (1 - error rate)

Fraction of docs classified correctly:

$$\frac{\sum_i c_{ii}}{\sum_j \sum_i c_{ij}}$$

## Micro- vs. Macro-Averaging

- If we have more than one class, how do we combine multiple performance measures into one quantity?
- **Macroaveraging:** Compute performance for each class, then average.
- **Microaveraging:** Collect decisions for all classes, compute contingency table, evaluate.

# Micro- vs. Macro-Averaging: Example

Class 1

	Truth: yes	Truth: no
Classifier: yes	10	10
Classifier: no	10	970

Class 2

	Truth: yes	Truth: no
Classifier: yes	90	10
Classifier: no	10	890

Micro Ave. Table

	Truth: yes	Truth: no
Classifier: yes	100	20
Classifier: no	20	1860

- Macroaveraged precision:  $(0.5 + 0.9)/2 = 0.7$
- Microaveraged precision:  $100/120 = .83$
- Microaveraged score is dominated by score on common classes

# Development Test Sets and Cross-validation

Training set

Development Test Set

Test Set

- **Metric: P/R/F1 or Accuracy**
- Unseen test set
  - avoid overfitting ('tuning to the test set')
  - more conservative estimate of performance
- **Cross-validation over multiple splits**
  - Handle sampling errors from different datasets
  - Pool results over each split
  - Compute pooled dev set performance





# The Real World

- Gee, I'm building a text classifier for real, now!
- What should I do?

# No training data?

## Manually written rules

If (wheat or grain) and not (whole or bread) then  
Categorize as grain

- Need careful crafting
  - Human tuning on development data
  - Time-consuming: 2 days per class

# Very little data?

- Use Naïve Bayes
  - Naïve Bayes is a “high-bias” algorithm (Ng and Jordan 2002 NIPS)
- Get more labeled data
  - Find clever ways to get humans to label data for you
- Try semi-supervised training methods:
  - Bootstrapping, EM over unlabeled documents, ...

## A reasonable amount of data?

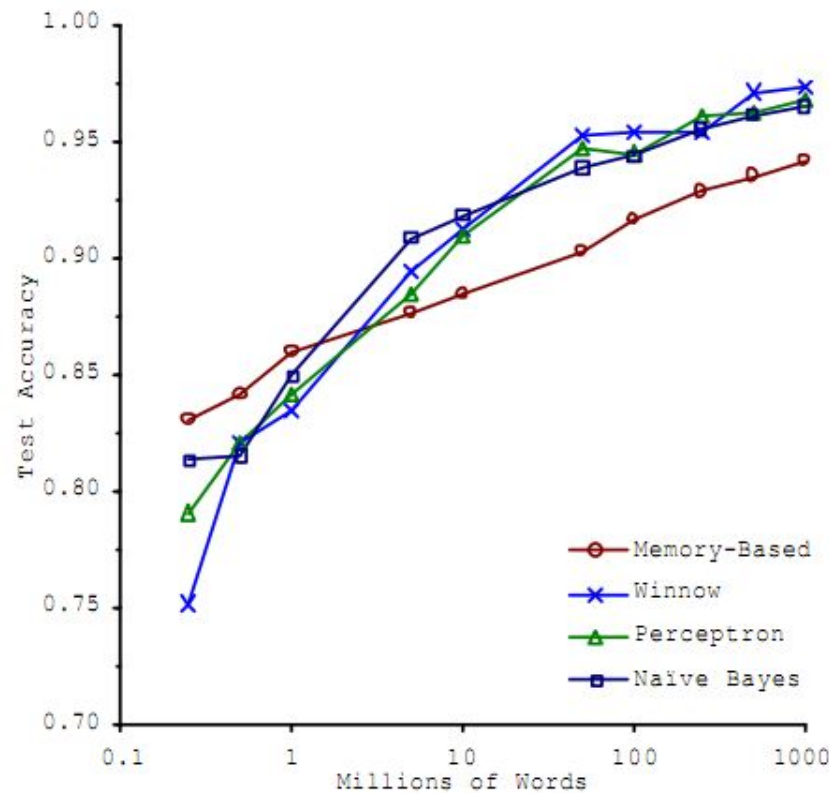
- Perfect for all the clever classifiers
  - SVM
  - Regularized Logistic Regression
- You can even use user-interpretable decision trees
  - Users like to hack
  - Management likes quick fixes

# A huge amount of data?

- Can achieve high accuracy!
- Deep learning (recurrent neural networks, word embedding representations, ...)
- **Pretrained language models:** depending on suitability of pretrained model to target task (same data distribution), can be fine-tuned for virtually any NLP task and achieve high performance - even with little fine-tuning (see huggingface for models) - also applies to tasks in next slides

# Accuracy as a function of data size

- With enough data
  - Classifier may not matter



Brill and Banko on spelling correction

# How to tweak performance

- Domain-specific features and weights: *very* important in real performance
- Sometimes need to collapse terms:
  - Part numbers, chemical formulas, ...
  - Stemming (?) etc (see text preprocessing)

# Information Extraction and Named Entity Recognition

---

Introducing the tasks:

Getting simple structured information out of text



# Information Extraction

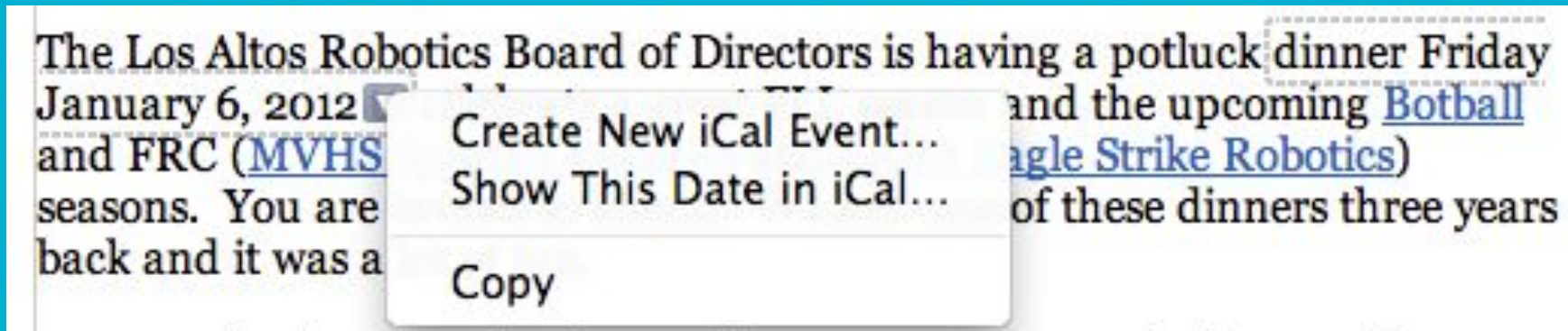
- Information extraction (IE) systems
  - Find and understand limited relevant parts of texts
  - Gather information from many pieces of text
  - Produce a structured representation of relevant information:
    - *relations* (in the database sense), a.k.a.,
    - *a knowledge base*
  - Goals:
    1. 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 Billiton Limited", "Melbourne, Australia")`
  - Learn drug-gene product interactions from medical research literature

# Low-level information extraction

- Is now available – and I think popular – in applications like Apple or Google mail, and web indexing



- Often seems to be based on regular expressions and name lists

# Low-level information extraction

Google

**Search** About 123,000 results (0.23 seconds)

Everything	Best guess for BHP Billiton Ltd. Headquarters is <b>Melbourne, London</b> Mentioned on at least 9 websites including <a href="#">wikipedia.org</a> , <a href="#">bhpbilliton.com</a> and <a href="#">bhpbilliton.com</a> - <a href="#">Feedback</a>
Images	
Maps	
Videos	<a href="#">BHP Billiton - Wikipedia, the free encyclopedia</a> <a href="#">en.wikipedia.org/wiki/BHP_Billiton</a>
News	Merger of BHP & Billiton 2001 (creation of a DLC). <b>Headquarters, Melbourne, Australia (BHP Billiton Limited and BHP Billiton Group) London, United Kingdom ...</b>
Shopping	<a href="#">History</a> - <a href="#">Corporate affairs</a> - <a href="#">Operations</a> - <a href="#">Accidents</a>



# 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 off, 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
- Finding relations between medical entities, automatically indexing documents



# Evaluation of Named Entity Recognition

The extension of Precision, Recall, and the F measure to sequences

# The Named Entity Recognition Task

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

# Precision/Recall/F1 for IE/NER

- Recall and precision are straightforward for tasks like IR and text categorization, where there is only one grain size (documents)
- The measure behaves a bit funnily for IE/NER when there are *boundary errors* (which are *common*):
  - First Bank of Chicago announced earnings ...
- This counts as both a fp and a fn
- Selecting *nothing* would have been better
- Some other metrics (e.g., MUC scorer) give partial credit (according to complex rules)

# Sequence Models for Named Entity Recognition

# The ML sequence model approach to NER

## Training

1. Collect a set of representative training documents
2. Label each token for its entity class or other (O)
3. Design feature extractors appropriate to the text and classes
4. Train a sequence classifier to predict the labels from the data

## Testing

1. Receive a set of testing documents
2. Run sequence model inference to label each token
3. Appropriately output the recognized entities

# Encoding classes for sequence labeling

IO encoding IOB encoding

Fred	PER	B-PER
showed	O	O
Sue	PER	B-PER
Mengqiu	PER	B-PER
Huang	PER	I-PER
's	O	O
new	O	O
painting	O	O

# Features for sequence labeling

- Words
  - Current word (essentially like a learned dictionary)
  - Previous/next word (context)
- Other kinds of inferred linguistic classification
  - Part-of-speech tags
- Label context
  - Previous (and perhaps next) label

# Features: Word substrings

oxa

:

field

Cotrimoxazole

Wethersfield

Alien Fury: Countdown to Invasion



# Features: Word shapes

- Word Shapes
  - Map words to simplified representation that encodes attributes such as length, capitalization, numerals, Greek letters, internal punctuation, etc.

Varicella-zoster	Xx-xxx
mRNA	xXXX
CPA1	XXXd

# Sequence problems

- Many problems in NLP have data which is a sequence of characters, words, phrases, lines, or sentences ...
- We can think of our task as one of labeling each item

VBG	NN	IN	DT	NN	IN	NN
Chasing	opportunity	in	an	age	of	upheaval

POS tagging

PERS	O	O	O	ORG	ORG
Murdoch	discusses	future	of	News	Corp.

Named entity recognition

B	B	I	I	B	I	B	I	B	B
而	相	对	于	这	些	品	牌	的	价

Word segmentation



Text  
segmentation

# 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

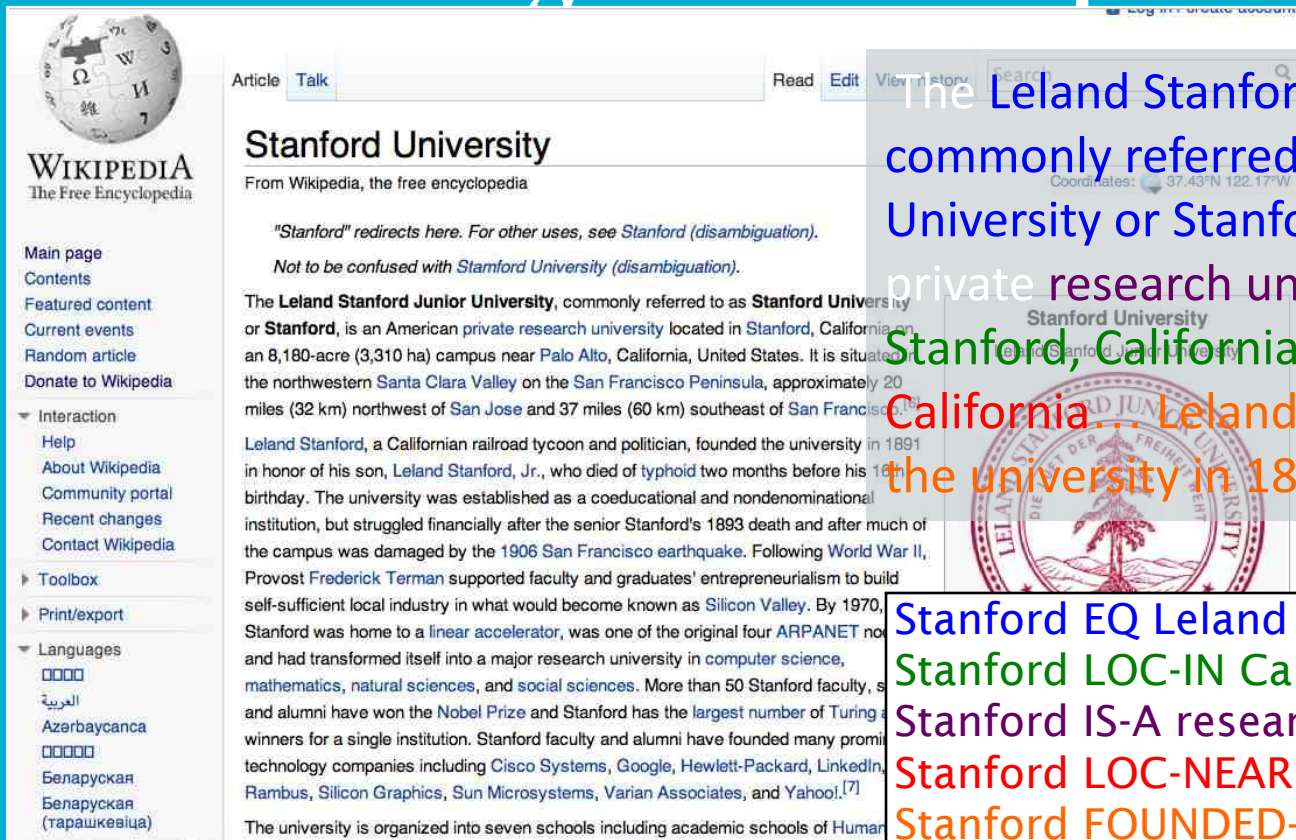
Original-Name Computing-Tabulating-Recording Co.

- 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



Article Talk Read Edit View history

## Stanford University

From Wikipedia, the free encyclopedia

*"Stanford" redirects here. For other uses, see Stanford (disambiguation).*

*Not to be confused with Stamford University (disambiguation).*

The **Leland Stanford Junior University**, commonly referred to as **Stanford University** or **Stanford**, is an American private research university located in Stanford, California, on an 8,180-acre (3,310 ha) campus near Palo Alto, California, United States. It is situated in the northwestern Santa Clara Valley on the San Francisco Peninsula, approximately 20 miles (32 km) northwest of San Jose and 37 miles (60 km) southeast of San Francisco.

Leland Stanford, a Californian railroad tycoon and politician, founded the university in 1891 in honor of his son, Leland Stanford, Jr., who died of typhoid two months before his 16th birthday. The university was established as a coeducational and nondenominational institution, but struggled financially after the senior Stanford's 1893 death and after much of the campus was damaged by the 1906 San Francisco earthquake. Following World War II, Provost Frederick Terman supported faculty and graduates' entrepreneurialism to build self-sufficient local industry in what would become known as Silicon Valley. By 1970, Stanford was home to a linear accelerator, was one of the original four ARPANET nodes and had transformed itself into a major research university in computer science, mathematics, natural sciences, and social sciences. More than 50 Stanford faculty, students, and alumni have won the Nobel Prize and Stanford has the largest number of Turing Award winners for a single institution. Stanford faculty and alumni have founded many prominent technology companies including Cisco Systems, Google, Hewlett-Packard, LinkedIn, Rambus, Silicon Graphics, Sun Microsystems, Varian Associates, and Yahoo!.[7]

The university is organized into seven schools including academic schools of Human

The Leland Stanford Junior University, commonly referred to as Stanford University or Stanford, 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 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)
- which relations should we extract?

# 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

# 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

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, treasures, <b>and other</b> important civic buildings.
X or other Y	Bruises, wounds, broken bones <b>or other</b> injuries...
Y such as X	The bow lute, <b>such as</b> the Bambara ndang...
Such Y as X	... <b>such</b> authors <b>as</b> Herrick, Goldsmith, and Shakespeare.
Y including X	...common-law countries, <b>including</b> Canada and England...
Y , especially X	European countries, <b>especially</b> 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!

Named Entities aren't quite enough.  
Which relations hold between 2 entities?



Drug

Cure?  
Prevent?  
Cause?



Disease

# What relations hold between 2 entities?



PERSON

Founder?

Investor?

Member?

Employee?

President?



ORGANIZATION

# 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

# Relation Extraction

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.



# 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.

FAMILY

CITIZEN

SUBSIDIARY

FOUNDER



NIL

EMPLOYMENT

INVENTOR

...

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

### Entity-based features

Entity <sub>1</sub> type	ORG
Entity <sub>1</sub> head	<i>airlines</i>
Entity <sub>2</sub> type	PERS
Entity <sub>2</sub> head	<i>Wagner</i>
Concatenated types	ORGPERS

### Word-based features

Between-entity bag of words	{ <i>a, unit, of, AMR, Inc., immediately, matched, the, move, spokesman</i> }
Word(s) before Entity <sub>1</sub>	NONE
Word(s) after Entity <sub>2</sub>	<i>said</i>

### 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
  - MaxEnt
  - Naïve Bayes
  - 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

# 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

# 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

# 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

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

# Unsupervised relation extraction

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)

# 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