# Machine Learning for NLP

- Until the 1990s, NLP systems primarily consisted of manually written dictionaries, grammars, and rules.
- Manually creating resources is time-consuming, prone to errors & omissions, and requires linguistic expertise.
- Manually built resources tend to be brittle when given real texts to process.
- Today, most NLP systems use statistical methods and machine learning, which are much more robust.

# Machine Learning

- **Machine learning (ML)** algorithms try to generalize from examples in order to make predictions about new instances that it sees.
- Three are many different machine learning methods!
- Too much generalization will make many errors. Too little generalization will **overfit** and fail to recognize instances that are different from the training examples.
- Each instance is typically represented as a feature vector. Each feature is an attribute that can take a range of values.

# Types of Machine Learning

Machine Learning (ML) systems generally come in three flavors:

1. **Supervised ML**: manually annotated texts are used for training.
2. **Weakly Supervised / Semi-Supervised ML**: small amounts of manually annotated text and unannotated texts are used for training.
3. **Unsupervised ML**: predictions are made based entirely on unannotated texts, usually with clustering algorithms.

# Simple Feature Vectors

XYZ Corp.$_{\textbf{ORG}}$ hired Susan Miller$_{\textbf{PER}}$ as its president .

| **Features** | **NP1** | **NP2** | **NP3** |
|---|---|---|---|
| *Head* | Corp. | Miller | president |
| *Capitalized* | yes | yes | no |
| *ContainsPunc* | yes | no | no |
| *PrevWord* | ϕ | hired | its |
| *NextWord* | hired | as | . |

# Feature Engineering

- The performance of any ML system depends crucially on the features used to represent the examples.
- The choice of features is more important than the choice of learning algorithm!
  - given inadequate features, no ML algorithm will perform well.
  - given strong features, most ML algorithms will perform well.
- Many ML algorithms exist with different strengths and weaknesses, but defining good features is *always* important!

# Bird or Mammal?

| Feature | Example 1 | Example 2 | Example 3 |
|---|---|---|---|
| *Weight* | 2 lbs | 5 lbs | 20 lbs |
| *Color* | yellow | black | brown |
| *Habitat* | forest | desert | prairie |
| *Food* | nuts | insects | mice |
| *Wings* | yes | yes | no |
| *Beak* | yes | no | no |

# Manual Annotation

We manually annotate text for several reasons:

- to understand the nature of the data

  Ex: what % of sentences in news are opinions?
- to precisely define the problem

  Ex: establishing guidelines that humans can agree on
- to establish the level of human performance on the task

  Ex: how consistently do people assign POS tags?
- to evaluate the performance of an NLP system

  Ex: how often does my tagger produce the correct labels?

# Annotated Text

From CNN:

Legislative subpoenas could be served to the aides of New Jersey Gov. Chris Christie as early as Monday sources told ABC News today.

Christie has been under intense political scrutiny after it was revealed that some of his top political aides shut down key traffic lanes on the George Washington Bridge -- the busiest bridge in the world -- in September for what appear to be politically motivated reasons.

# Named Entity Annotations

*From ABC News:*

Legislative subpoenas could be served to the aides of New Jersey$_{LOC}$ Gov. Chris Christie$_{PER}$ as early as Monday$_{DATE}$ sources told ABC News$_{ORG}$ today.

Christie$_{PER}$ has been under intense political scrutiny after it was revealed that some of his top political aides shut down key traffic lanes on the George Washington Bridge$_{SITE}$ -- the busiest bridge in the world -- in September$_{DATE}$ for what appear to be politically motivated reasons.

# Annotations for all people

*From ABC News:*

Legislative subpoenas could be served to the aides$_{PER}$ of New Jersey Gov. Chris Christie$_{PER}$ as early as Monday sources$_{PER}$ told ABC News today.

Christie$_{PER}$ has been under intense political scrutiny after it was revealed that some of his top political aides$_{PER}$ shut down key traffic lanes on the George Washington Bridge -- the busiest bridge in the world -- in September for what appear to be politically motivated reasons.

# Annotating Text is Deceptively Trick

- Which parts of a phrase are essential? Modifers? Prepositional phrases?
- Some categories can be expressed as clauses (e.g., opinions).
- Do you label all references to an entity? Only the most specific one? Are multiple specific references allowed?
- Do conjunctions get labeled as one item or two?
- Metonymy is particularly troublesome (e.g., when locations refer to governments, which refer to representatives of the government…)

# Annotating Event Information

Alleged guerrilla urban commandos launched highpower bombs against a Mark Miller Toyota in downtown Salt Lake City this morning. Three men were seen fleeing the scene.

A police report said that the attack set the car dealership on fire, but did not result any casualties. A nearby traffic light was damaged in the explosion, causing traffic jams during rush hour. Police apprehended three suspects late this evening.

# Annotation Consistency is Essential

- Inter-annotator agreement (IAA) measures the consistency of different people when annotating data.
- High IAA is important to ensure that the task is well-defined and that the annotations have high integrity.
- To achieve high IAA, you must:
  - precisely define the annotation task, which usually requires ***detailed annotation guidelines*** that include examples and discuss how to handle boundary cases.
  - train the annotators, often iteratively refining the guidelines.
  - measure IAA using an appropriate statistical measure.

# The Kappa Statistic

The Kappa ($\kappa$) statistic measures the degree to which sets of annotations agree, adjusted for agreement due to chance.

$$\kappa = \frac{P(agree) - P(expected)}{1 - P(expected)}$$

where:

P(agree) = proportion of times the annotators agree

P(expected) = proportion of times the annotators are expected to agree by chance.

# Kappa for Two Annotators

$$P(expected) = \sum_{c \in C} P(c \mid A_1) * P(c \mid A_2)$$

where:

C = the set of possible classes (labels)

$A_1$ = annotator #1's labels

$A_2$ = annotator #2's labels

# An Example

$A_1$ = Y Y N Y N Y N N Y Y

$A_2$ = Y Y N N Y Y Y N Y Y

P(agree) = 7/10 = .70

P(expected)
= (P(Y | $A_1$) * P(Y | $A_2$)) + (P(N | $A_1$) * P(N | $A_2$))
= (6/10 * 7/10) + (4/10 * 3/10)
= .54

$\kappa$ = (.70 - .54) / (1 - .54) = .348

# Data Sets

NLP experiments typically rely on three annotated data sets:

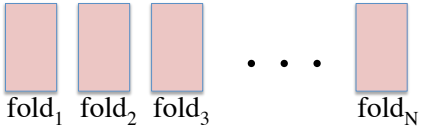
**Training/Development Data:** used for training. The developer can also inspect the data to help design the system.

**Tuning Data:** used as a pseudo-test set to evaluate the system during development and determine parameter settings.

**Test Data:** blind data used for evaluation. The developer should never look at this data.

# Cross-Validation

- An alternative experimental design is called **cross-validation** (**or jack-knifing**).
- The annotated data is divided into N partitions called ***folds***. N experiments are performed, each using a different fold for testing and the remaining folds for training. The results are then averaged.




Cross-validation allows ALL of the data to be used for both training and testing, but never at the same time.

# Baselines

- It is important to understand how difficult a problem is, and how well simple or standard techniques perform.

  Ex: choosing the most frequent POS tag yields ~90% accuracy

- Simple techniques sometimes work surprisingly well. Don't assume that a fancy technique will be better!

  Ex: a POS tagger that achieves < 90% accuracy is not very good!

- It is essential to know the underlying class distributions in your data.

# Performance Measures

Common measures to evaluate performance are:

**Accuracy:** the % of instances assigned a correct label

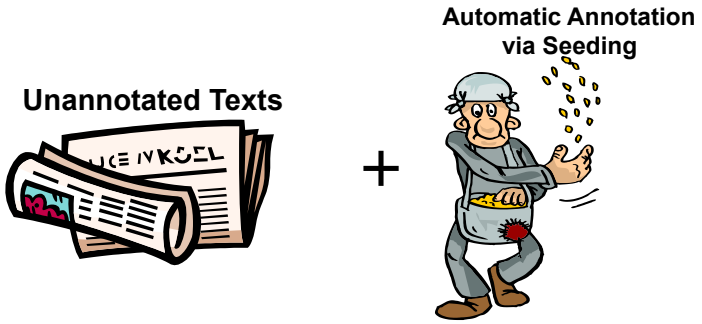
**Recall:** for a category C, the % of true instances of C that are correctly labeled: $\frac{\text{\# correctly labeled as C}}{\text{\# true instances of C}}$

**Precision:** for a category C, the % of instances assigned the label C that are correctly labeled: $\frac{\text{\# correctly labeled as C}}{\text{\# labeled as C}}$

**F Score:** the harmonic mean of Recall and Precision

$$\frac{2 * \text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}}$$

# The Bootstrapping Era




# Why Bootstrapping?

- Manually annotating data:
  - is time-consuming and expensive
  - is deceptively difficult
  - often requires linguistic expertise
- NLP systems benefit from domain-specific training
  - it is not realistic to expect manually annotated data for every domain and task.
  - domain-specific training is sometimes *essential*

# Additional Benefits of Bootstrapping

- Dramatically easier and faster system development time.
  - Allows for free-wheeling experimentation with different categories and domains.
- Encourages cross-resource experimentation.
  - Allows for more analysis across domains, corpora, genres, and languages.

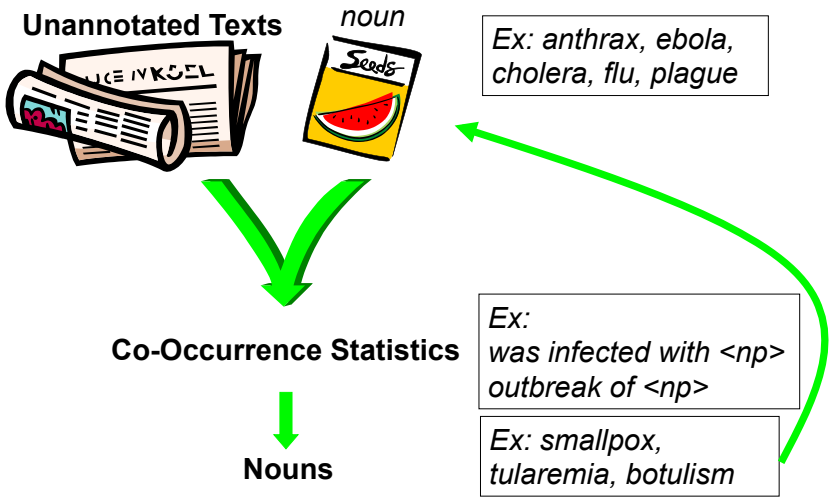
# Automatic Annotation with Seeding

- Goal: to avoid the need for manual annotation.
- The system should be trainable with a small set of seeds that can be done by anyone!
- Seeding is often is done using "stand-alone" examples or rules.
- Fast, less expertise required, but noisier!

# Many seeding strategies are possible

- seed words
- seed patterns
- seed rules
- seed heuristics
- seed classifiers

## Seed Nouns for Semantic Class Bootstrapping




## Seed Words for Word Sense Disambiguation

[Yarowsky, 1995]

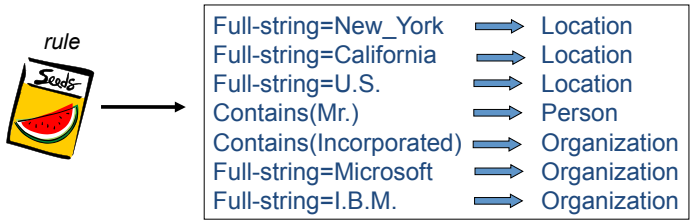
Yarowsky's best WSD performance came from a list of *top collocates*, manually assigned to the correct sense.

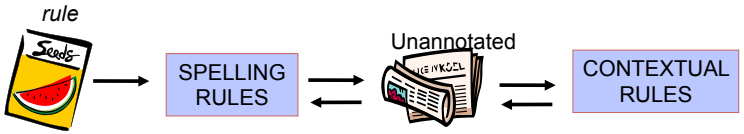
Ex: *"life"* and *"manufacturing"* for *"plant"*

| Sense | Training Example |
|---|---|
| A | …zonal distribution of plant life from the… |
| A | …many dangers to plant and animal life… |
| ? | …union responses to plant closures… |
| ? | …company said the plant still operating… |
| B | …copper manufacturing plant found that… |
| B | …keep a manufacturing plant profitable… |

Also exploited *"one sense per discourse"* heuristic.

## Seed Rules for Named Entity Recognition Bootstrapping [Collins & Singer, 1999]



## The Importance of Good Seeding

Poor seeding can lead to a variety of problems:

- Bootstrapping sputters and dies
  ⇒ nothing learned
- Bootstrapping goes astray / gets derailed
  ⇒ the wrong concept learned
- Bootstrapping thrashes
  ⇒ only subsets learned
- Bootstrapping learns low-frequency cases
  ⇒ high precision but low recall

## General Criteria for Seed Data

Seeding instances should be *FREQUENT*

Want as much coverage and contextual diversity as possible!

Bad animal seeds: *coatimundi, giraffe, terrier*

## Common Mistake #1

Assuming you know which instances are the most frequent.

*Never assume* you know what is frequent!
Seed instances must be frequent in **your** texts.

We are not very good at guessing how common terms are, especially for a specific domain.

**Trust the data:** determine which words are *actually* the most frequent.

## General Criteria for Seed Data

Seeding instances should be UNAMBIGUOUS

Ambiguous seeds create noisy training data.

Bad animal seeds: *bat, jaguar, turkey*

# Common Mistake #2

Careless inattention to ambiguity.

A word may seem like a perfect example at first, but upon further reflection you may realize that it has other common meanings as well.

If a seed instance does not *consistently* represent the desired concept (in your corpus), then bootstrapping can be derailed.

# General Criteria for Seed Data

Seeding instances should be: *REPRESENTATIVE*

You want instances that:

- cover all of the desired categories
- are not atypical category members

Why? Bootstrapping is fragile in its early stages…

Bad bird seeds: *penguin, ostrich, hummingbird*

# Common Mistake #3

Insufficient coverage of different classes or contexts.

It is easy to forget that all desired classes and types need to be adequately represented in the seed data.

Seed data for *negative* instances may need to be included as well!

# General Criteria for Seed Data

Seeding instances should be: *DIVERSE*

Want instances that cover different regions of the search space.

Bad animal seeds: *cat, cats, kitty, kitten*

Bad animal seeds: *dog, cat, goldfish, parakeet*

# Common Mistake #4

Need a balance between coverage and diversity.

Diversity is important, but need to have critical mass representing different parts of the search space.

One example from each of several wildly different classes may not provide enough traction.

# Conclusions

- Many different seeding strategies have been used successfully.
- Bootstrapping methods are sensitive to their initial seeds. You want seed cases that are:
  - frequent
  - unambiguous
  - representative
  - diverse
- But there is still a need for manually annotated data for system evaluation!