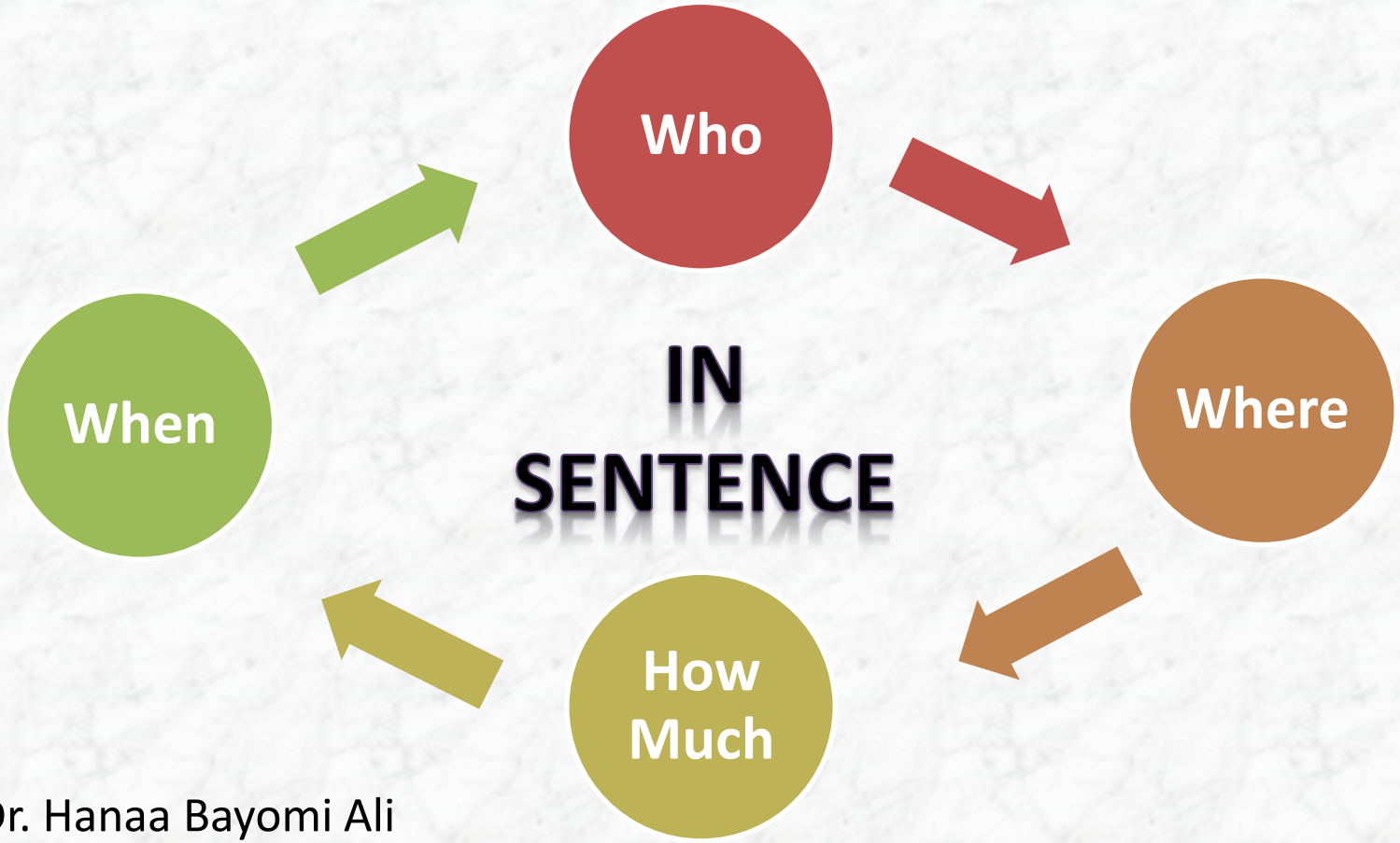
Machine Learning for Named Entity Recognition(NER)



Instructor : Dr. Hanaa Bayomi Ali
Mail : h.mobarz @ fci-cu.edu.eg

The who, where, when & how much in a sentence

- The task: identify lexical and phrasal information in text which express references to named entities NE, e.g.,

- person names
- company/organization names
- locations
- dates & time
- percentages
- monetary amounts (Currency)
- number
- Device
- Jop
- Car
- Cell Phone

- Determination of an NE

- Specific type according to some taxonomy
- Canonical representation (template structure)

Example of NE-annotated text

Delimit the named entities in a text and tag them with NE types:

Italy‘s business world was rocked by the announcement last Thursday that Mr. Verdi would leave his job as vice president of Music Masters of Milan, Inc to become operations director of Arthur Andersen

Example of NE-annotated text

Delimit the named entities in a text and tag them with NE types:

<ENAMEX TYPE=„LOCATION“>Italy</ENAMEX>‘s business world was rocked by the announcement <TIMEX TYPE=„DATE“>last Thursday</TIMEX> that Mr. <ENAMEX TYPE=„PERSON“>Verdi</ENAMEX> would leave his job as vice-president of <ENAMEX TYPE=„ORGANIZATION“>Music Masters of Milan, Inc</ENAMEX> to become operations director of <ENAMEX TYPE=„ORGANIZATION“>Arthur Andersen</ENAMEX>.

- “Milan“ is part of organization name
- “Arthur Andersen“ is a company
- “Italy“ is sentence-initial ⇒ capitalization useless

NE and Question-Answering

Often, the expected answer type of a question is a **NE**

What was the first USA president?

- Expected answer type is PERSON

Name the five most important software companies

- Expected answer type is a list of COMPANY

Where will be the INFOS2020 conferance take place?

- Expected answer type is LOCATION

When will be the next talk?

- Expected answer type is DATE

Difficulties of Automatic NER

- Potential set of NE is too numerous to include in dictionaries/Gazetteers
- Names changing constantly
- Ambiguity of NE types :
  - John Smith (company vs. person)
  - Washington (person vs. Location)
- Names appear in many variant forms.

  e.g. John Smith, Mr.Smith, John.
- Subsequent occurrences of names might be abbreviated

  ⇒ list search/matching does not perform well

  ⇒ context based pattern matching needed

Arabic NER Problems

- Lack of resources
- Arabic high inflectional property

  نفخر بمصرنا

- Arabic doesn't define case for letters (upper case letter)
- Arabic has some variants in spelling and typographic forms

  استراليا/أستراليا ,جرام/غرام

- Different sources of ambiguity (vowel, sense, lexical, and syntactic)
  - Vowels disambiguation

    مصر

  - Sense disambiguation

    محمود

Two kinds of NE approaches

Knowledge Engineering

- rule based
- developed by experienced language engineers
- make use of human intuition
- requires only small amount of training data
- development could be very time consuming
- some changes may be hard to accommodate

Learning Systems

- use statistics or other ML
- developers do not need LE expertise
- requires large amounts of annotated training data
- some changes may require re-annotation of the entire training corpus

List lookup approach - baseline

- System that recognises only entities stored in its lists (gazetteers).
- Advantages - Simple, fast, language independent, easy to retarget (just create lists)
- Disadvantages - collection and maintenance of lists, cannot deal with name variants, cannot resolve ambiguity

Rule Based NER

- **Create regular expressions to extract:**
  - Telephone number
  - E-mail
  - Capitalized names

Rule Based NER

- **Create regular expressions to extract:**
  - Telephone number
  - E-mail
  - Capitalized names

blocks of digits separated by hyphens

RegEx = (\d+\-)+\d+

- matches valid phone numbers like 900-865-1125 and 725-1234
- incorrectly extracts social security numbers 123-45-6789
- fails to identify numbers like 800.865.1125 and (800)865-CARE

Improved RegEx = (\d{3}[-.\ ()]){1,2}[\dA-Z]{4}

Rule Based NER

Use context patterns

[*PERSON*] earned [*MONEY*]

Ex. *Frank earned $20*

[*PERSON*] joined [*ORGANIZATION*]

Ex. *Sam joined IBM*

[*PERSON*],[*JOBTITLE*]

Ex. *Mary, the teacher*

still not so simple:

[*PERSON*|*ORGANIZATION*] fly to [*LOCATION*|*PERSON*|*EVENT*]

Ex. *Jerry flew to Japan*

Sarah flies to the party

Delta flies to Europe

Examples of context patterns

- [PERSON] earns [MONEY]
- [PERSON] joined [ORGANIZATION]
- [PERSON] left [ORGANIZATION]
- [PERSON] joined [ORGANIZATION] as [JOBTITLE]
- [ORGANIZATION]'s [JOBTITLE] [PERSON]
- [ORGANIZATION] [JOBTITLE] [PERSON]
- the [ORGANIZATION] [JOBTITLE]
- part of the [ORGANIZATION]
- [ORGANIZATION] headquarters in [LOCATION]
- price of [ORGANIZATION]
- sale of [ORGANIZATION]
- investors in [ORGANIZATION]
- [ORGANIZATION] is worth [MONEY]
- [JOBTITLE] [PERSON]
- [PERSON], [JOBTITLE]

Rule Based Arabic NER

Class	Example
Person	الرئيس <Nat> + <Person Name> + رئيس <Location>
Location	ميناء بلدة + <Location>
Organization	وكالة أنباء + <Organization>
Time	وفي الساعة + <Time> + بتوقيت جرينتش
Date	شهر + <Date> + الجاري
Money	سعر الذهب <Decimal> + <Money>

Learning System

- *Supervised* learning

  labeled training examples

  methods: Hidden Markov Models, k-Nearest Neighbors, Decision Trees, AdaBoost, SVM, …

  example: NE recognition, POS tagging, Parsing

- *Unsupervised learning*

  labels must be automatically discovered

  method: clustering

  example: NE disambiguation, text classification

Learning System

- *Semi-supervised* learning

  small percentage of training examples are labeled, the rest is unlabeled

  methods: bootstrapping, active learning, co-training, self-training

  example: NE recognition, POS tagging, Parsing, …

Machine Learning NER

Adam_B Smith_I works_O for_O IBM_B ,_O London_B ._O

- **NED**: Identify named entities using BIO tags
  - B beginning of an entity
  - I continues the entity
  - O word outside the entity

Machine Learning NER

Adam_B-PER Smith_I-PER works_O for_O **IBM_B-ORG** ,_O **London_B-LOC** ._O

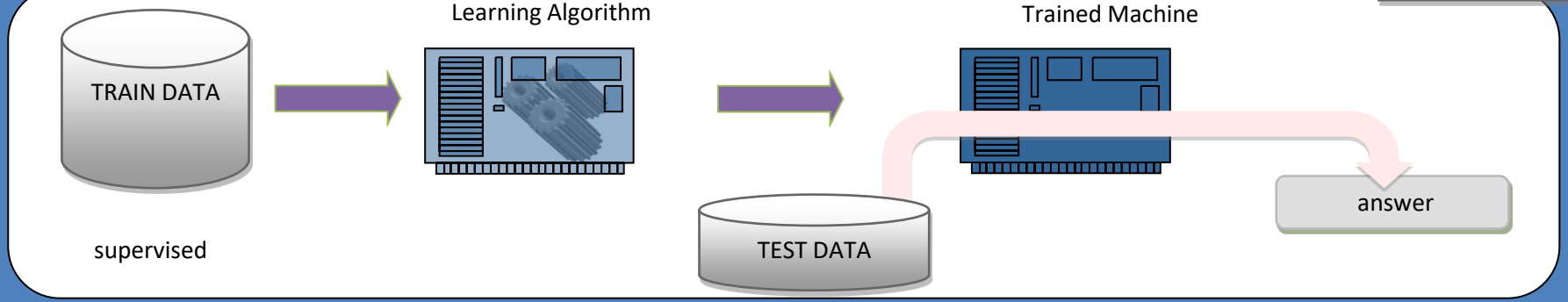
- **NED**: Identify named entities using BIO tags
  - B beginning of an entity
  - I continues the entity
  - O word outside the entity

- **NEC**: Classify into a predefined set of categories
  - Person names
  - Organizations (companies, governmental organizations, etc.)
  - Locations (cities, countries, etc.)
  - Miscellaneous (movie titles, sport events, etc.)

United States presidential election of 2008, scheduled for Tuesday November 4, 2008, will be the 56th consecutive quadernnial United States presidential election and will select the President and the Vice President of the United States. The Republican Party has chosen John McCain, the senior United States Senator from Arizona as its nominee; the Democratic Party has chosen Barak Obama, the junior United States Senator from Illinois, as its nominee.

Named Entity Detection

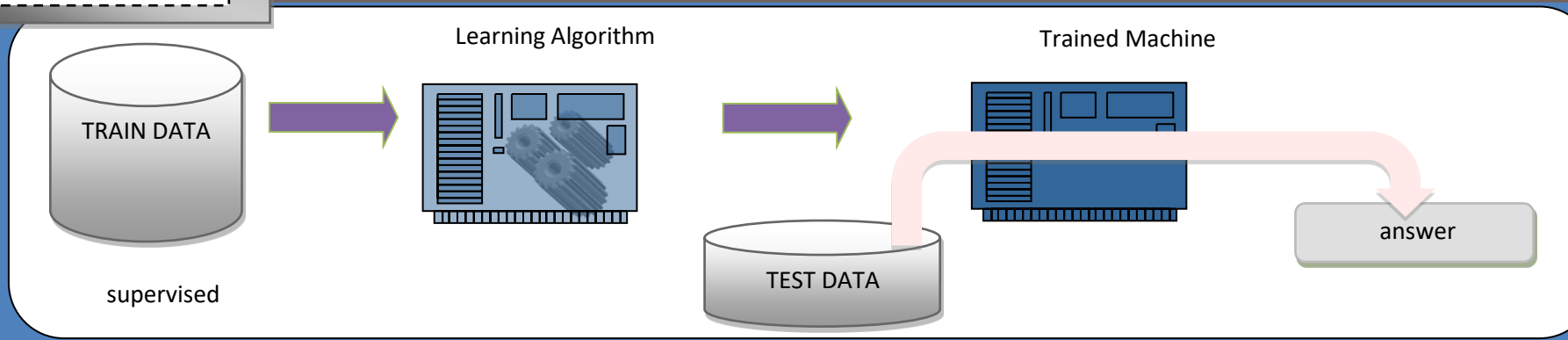
United,0,1,0,0,1,ed,ted,Un,Uni,1,0,1,1,1,null,null,null,States,presidential,election



United_B States_I presidential_O election_O of_O 2008_O ,_O scheduled_O for_O Tuesday_O November_O 4_O ,_O 2008_O ,_O will_O be_O the_O 56th_O consecutive_O quadernnial_O United_B States_I presidential_O election_O and_O will_O select_O the_O President_B and_O the_O Vice_B President_I of_I the_I United_I States_I. The_O Republican_B Party_I has_O chosen_O John_B McCain_I ,_O the_O senior_O United_B

Named Entity Classification

United,null,null,null,States,presidential,election,1,1,1,1,0,1,0,0,1,1,1,1,1,0,1,0,1,1,ed,ted,Uni,Un



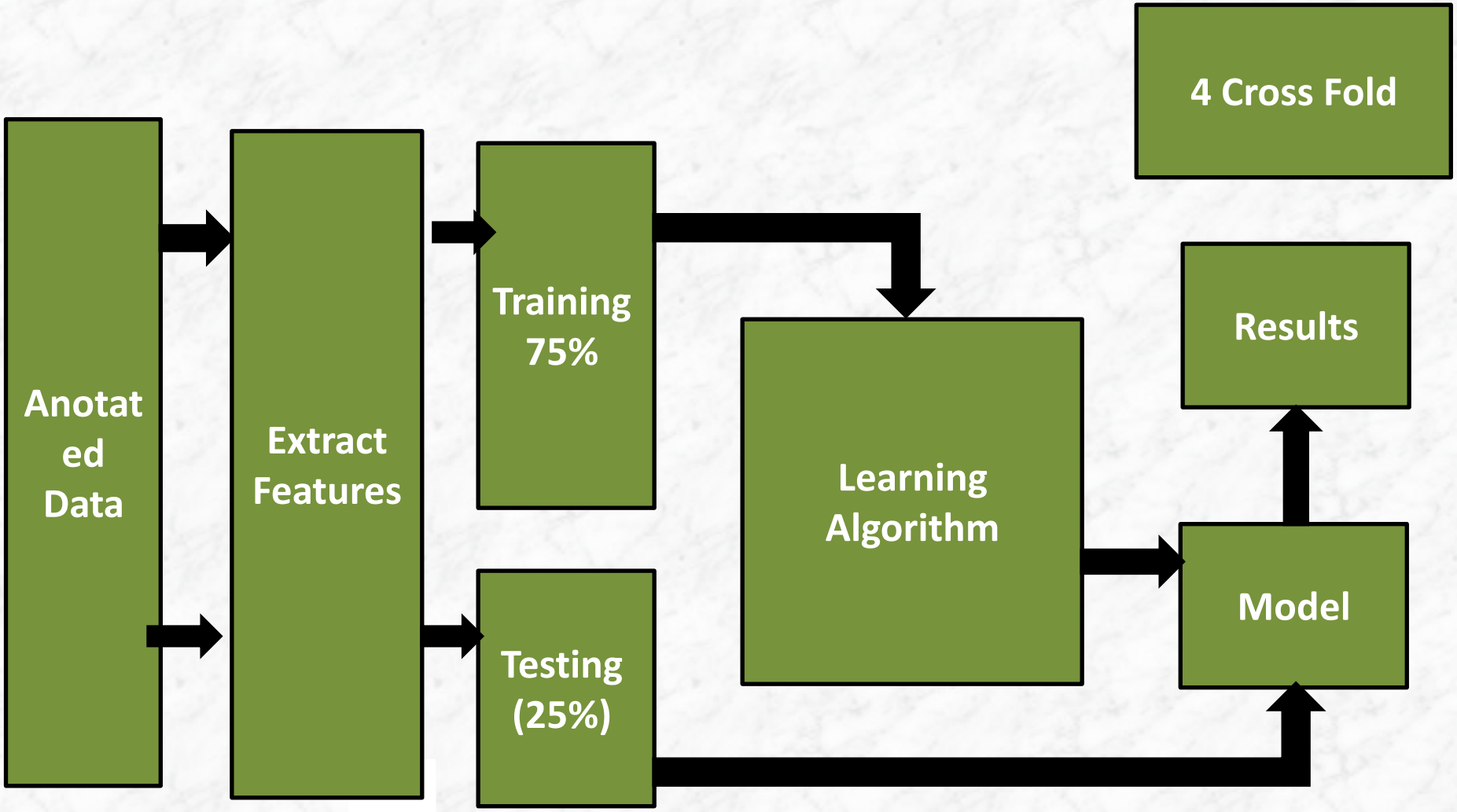
United_B-LOC States_I-LOC presidential_O election_O of_O 2008_O ,_O scheduled_O for_O Tuesday_O November_O 4_O ,_O 2008_O ,_O will_O be_O the_O 56th_O consecutive_O quadernnial_O United_B-LOC States_I-LOC presidential_O election_O and_O will_O select_O the_O President_B-PER and_O the_O Vice_B-PER President_I-PER of_I-PER the_I-PER United_I-PER States_I-PER. The_O Republican_B-ORG Party_I-ORG has_O chosen_O John_B-PER McCain_I-PER ,_O the_O senior_O United_B-PER States_I-PER Senator_I-PER from_O Arizona_B-LOC as_O its_O

Machine Learning Arabic NER

وصل خادم الحرمين الشريفين الملك فهد بن عبد العزيز الي مطار القاهرة بسلامة الله

وصل_O خادم_PER-B الحرمين_PER-I الشريفين_PER-I الملك_PER-I فهد_PER-I
بن_PER-I عبد_PER-I العزيز_PER-I الي_O مطار_ORG-B القاهرة_PER-I
بسلامة_O الله_O

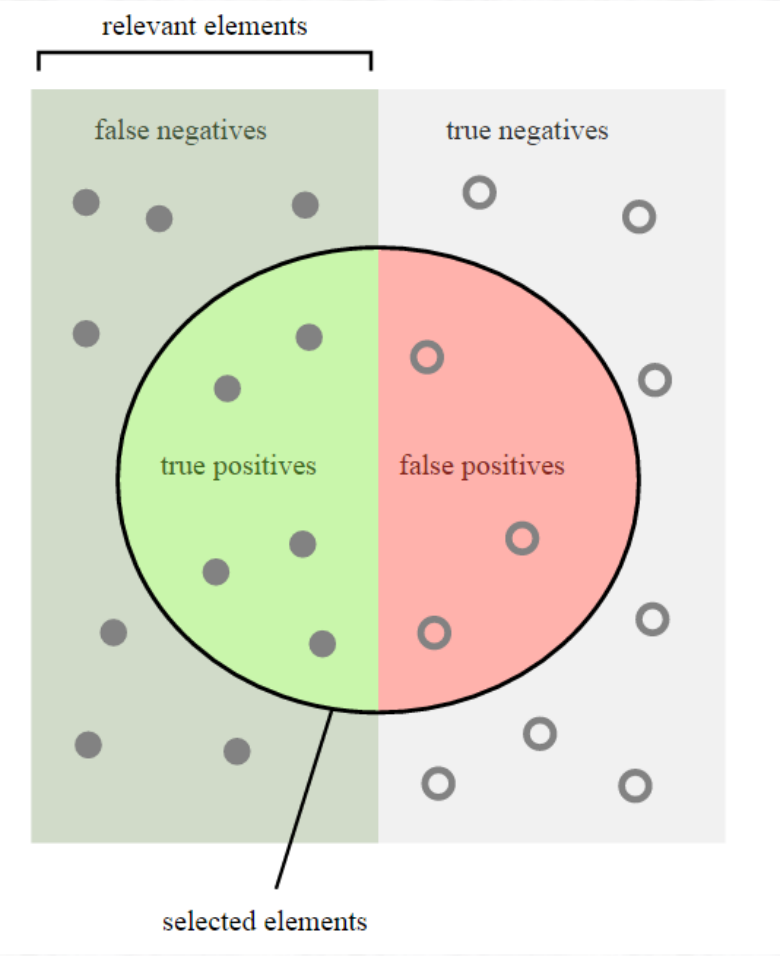
Supervised Learning Algorithms



Features for NE

- **Contextual**
  - current word W_0
  - words around W_0 in [-3,…,+3] window
- **Part-of-speech tag**
- **Trigger words**
  - for person (*Mr, Miss, Dr*)
  - for location (*city, street*)
  - for organization (*Ltd.*, *Co.*)
- **Gazetteers**
  - geographical
  - first name
  - surname
  - company names
- **Length in words of the entity being classified**

Precision and Recall Definition



How many selected items are relevant?

$$\text{Precision} = \frac{\text{true positives}}{\text{selected elements}}$$

How many relevant items are selected?

$$\text{Recall} = \frac{\text{true positives}}{\text{relevant elements}}$$

$$F\_measure = 2 * \frac{(Precision * Recall)}{(Precision + Recall)}$$

Problem

Assume the following:

- A database contains **80 records** on a particular topic
- A search was conducted on that topic and **60 records** were retrieved.
- Of the 60 records retrieved, **45 were relevant**. Calculate the precision and recall scores for the search

Solution:

- TP = The number of relevant records retrieved, 45
- SE = The number of retrieved records, 60
- FP = The number of irrelevant records retrieved. (60-45)

In this example

Recall = 45/80 * 100% = 56%

Precision = (45 / (45 + 15)) * 100% => 45/60 * 100% = 75%

F_Measure=(2(0.56*0.75)/(0.56+0.75))*100=64.12%