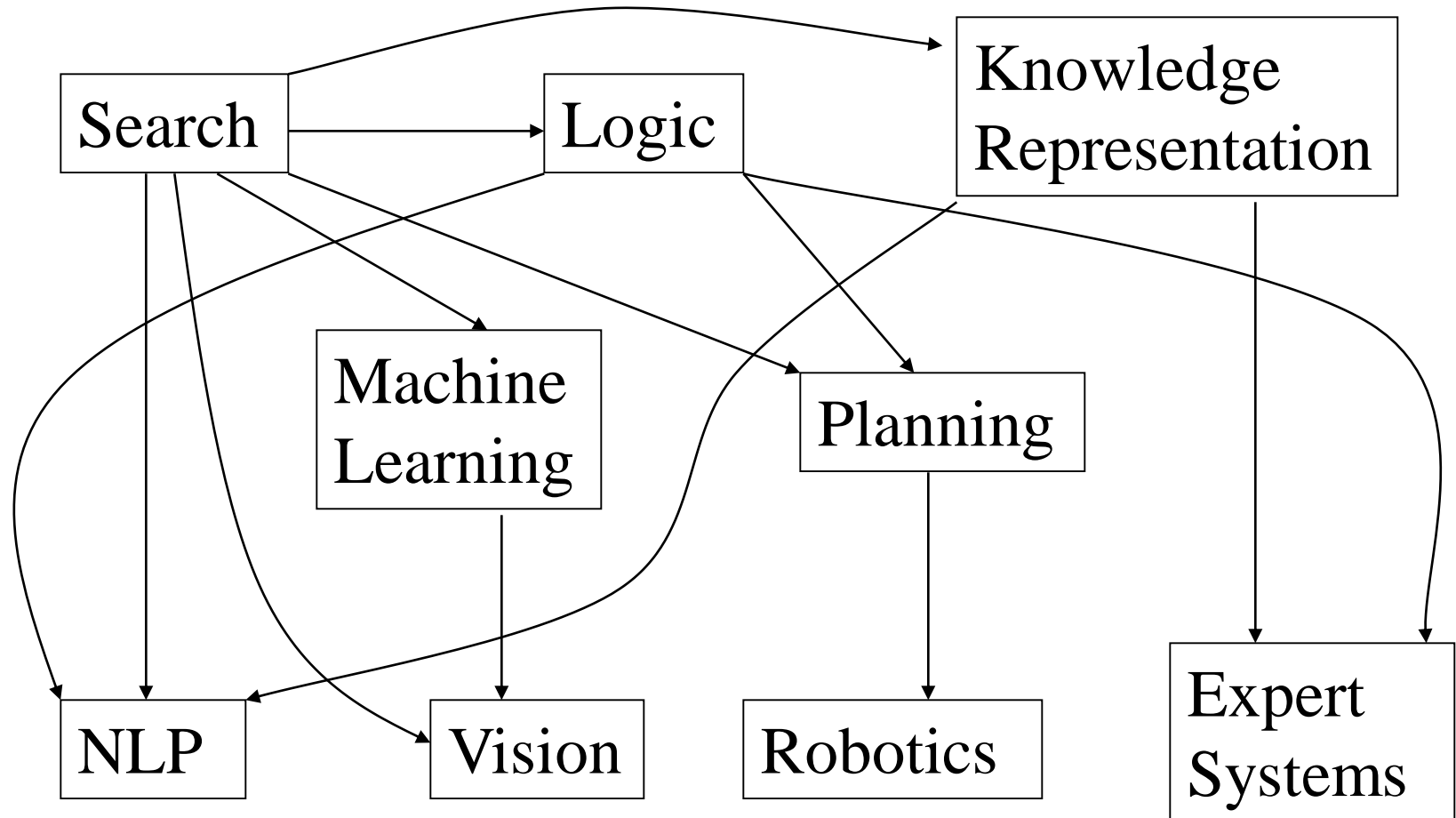


# Natural Language Processing

## (Lecture 1 – Introduction)

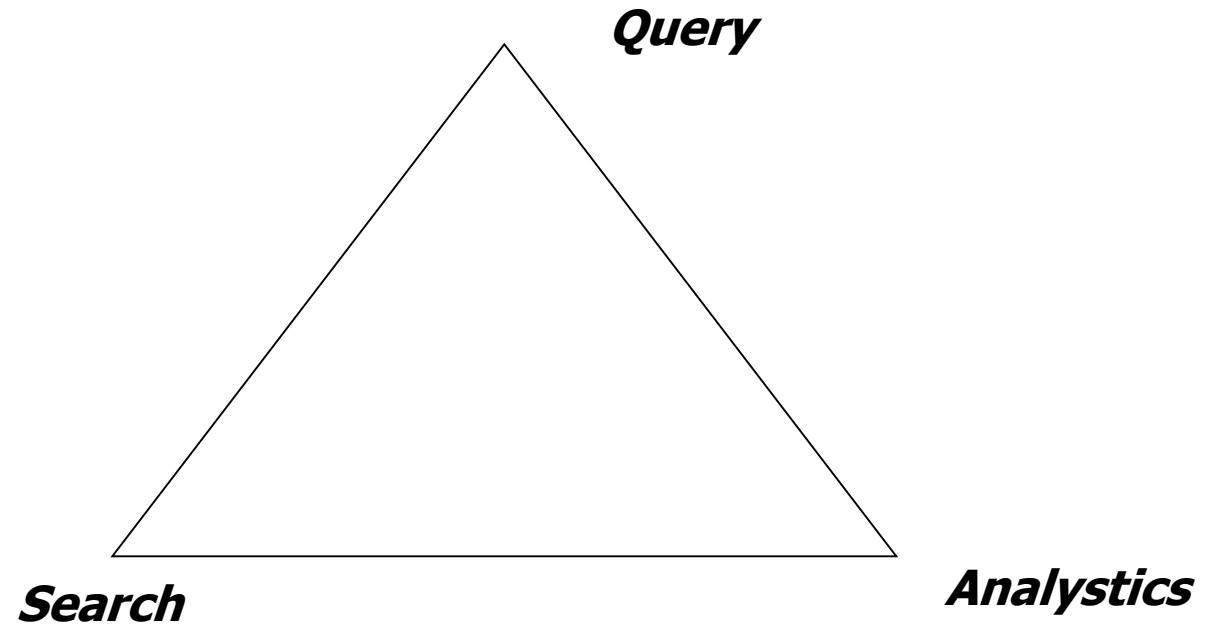
# Perpectivising NLP: Areas of AI and their inter-dependencies



# Web brings in new perspectives

- Web 2.0
- (wikipedia) In studying and/or promoting web-technology, the phrase **Web 2.0** can refer to a perceived second generation of web-based communities and hosted services — such as social-networking sites, wikis, and folksonomies — which aim to facilitate creativity, collaboration, and sharing between users.
- According to Tim O'Reilly, "Web 2.0 is the business revolution in the computer industry caused by the move to the Internet as platform, and an attempt to understand the rules for success on that new platform."

# QSA Triangle



# Areas being investigated

- Business Intelligence on the Internet Platform
- Opinion Mining
- Reputation Management
- Sentiment Analysis (*some observations at the end*)

NLP is thought to play a key role

# Books etc.

- Main Text(s):

- Natural Language Understanding: James Allan
- Speech and NLP: Jurafsky and Martin
- Foundations of Statistical NLP: Manning and Schutze

- Other References:

- NLP a Paninian Perspective: Bharati, Cahitanya and Sangal
- Statistical NLP: Charniak

- Journals

- Computational Linguistics, Natural Language Engineering, AI, AI Magazine, IEEE SMC

- Conferences

- ACL, EACL, COLING, MT Summit, EMNLP, IJCNLP, HLT, ICON, SIGIR, WWW, ICML, ECML

# Allied Disciplines

Philosophy	Semantics, Meaning of “meaning”, Logic (syllogism)
Linguistics	Study of Syntax, Lexicon, Lexical Semantics etc.
Probability and Statistics	Corpus Linguistics, Testing of Hypotheses, System Evaluation
Cognitive Science	Computational Models of Language Processing, Language Acquisition
Psychology	Behavioristic insights into Language Processing, Psychological Models
Brain Science	Language Processing Areas in Brain
Physics	Information Theory, Entropy, Random Fields
Computer Sc. & Engg.	Systems for NLP

# Topics to be covered

- Shallow Processing
  - Part of Speech Tagging and Chunking using HMM, MEMM, CRF, and Rule Based Systems
  - EM Algorithm
- Language Modeling
  - N-grams
  - Probabilistic CFGs
- Basic Linguistics
  - Morphemes and Morphological Processing
  - Parse Trees and Syntactic Processing: Constituent Parsing and Dependency Parsing
- Deep Parsing
  - Classical Approaches: Top-Down, Bottom-UP and Hybrid Methods
  - Chart Parsing, Earley Parsing
  - Statistical Approach: Probabilistic Parsing, Tree Bank Corpora



# Topics to be covered (contd.)

- Knowledge Representation and NLP
  - Predicate Calculus, Semantic Net, Frames, Conceptual Dependency, Universal Networking Language (UNL)
- Lexical Semantics
  - Lexicons, Lexical Networks and Ontology
  - Word Sense Disambiguation
- Applications
  - Machine Translation
  - IR
  - Summarization
  - Question Answering

# Grading

- Based on
  - Midsem
  - Endsem
  - Assignments
  - Seminar
  - Project (possibly)

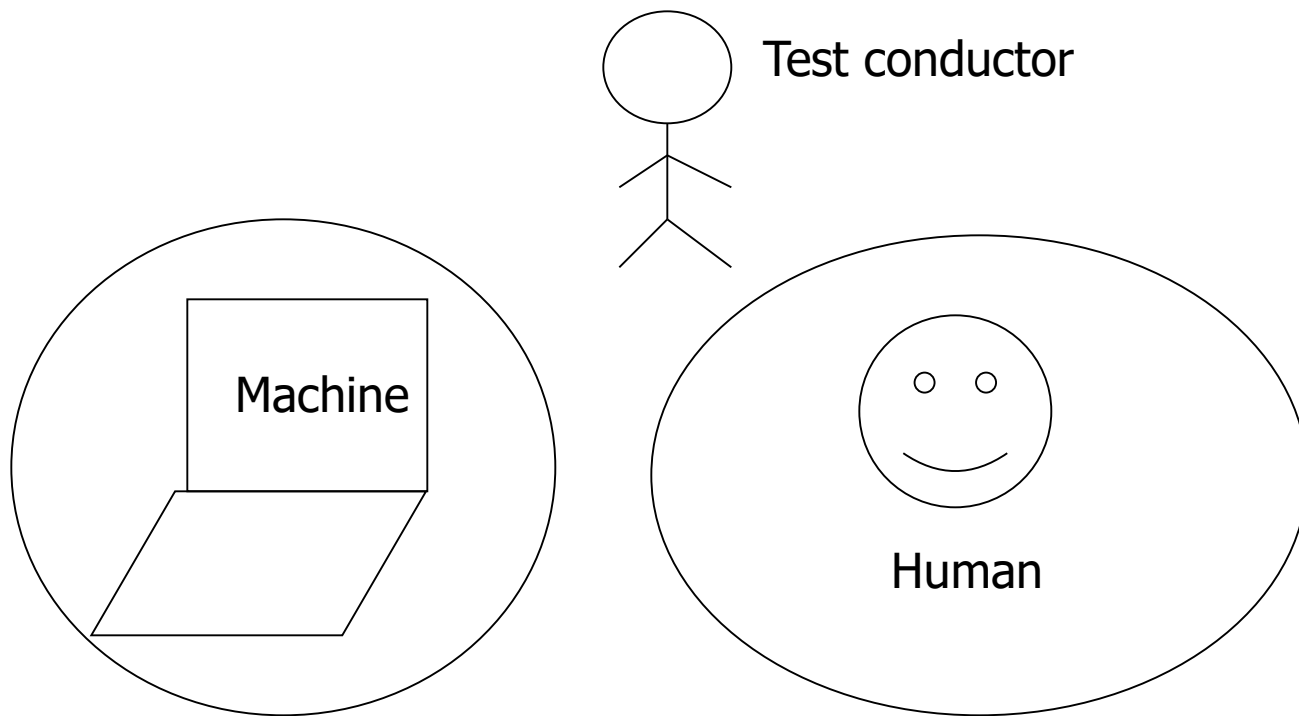
*Except the first two everything else in groups of 4. Weightages will be revealed soon.*

Definitions etc.

# What is NLP

- Branch of AI
- 2 Goals
  - Science Goal: Understand the language processing behaviour
  - Engineering Goal: Build systems that analyse and generate language; reduce the man machine gap

# The famous Turing Test: Language Based Interaction

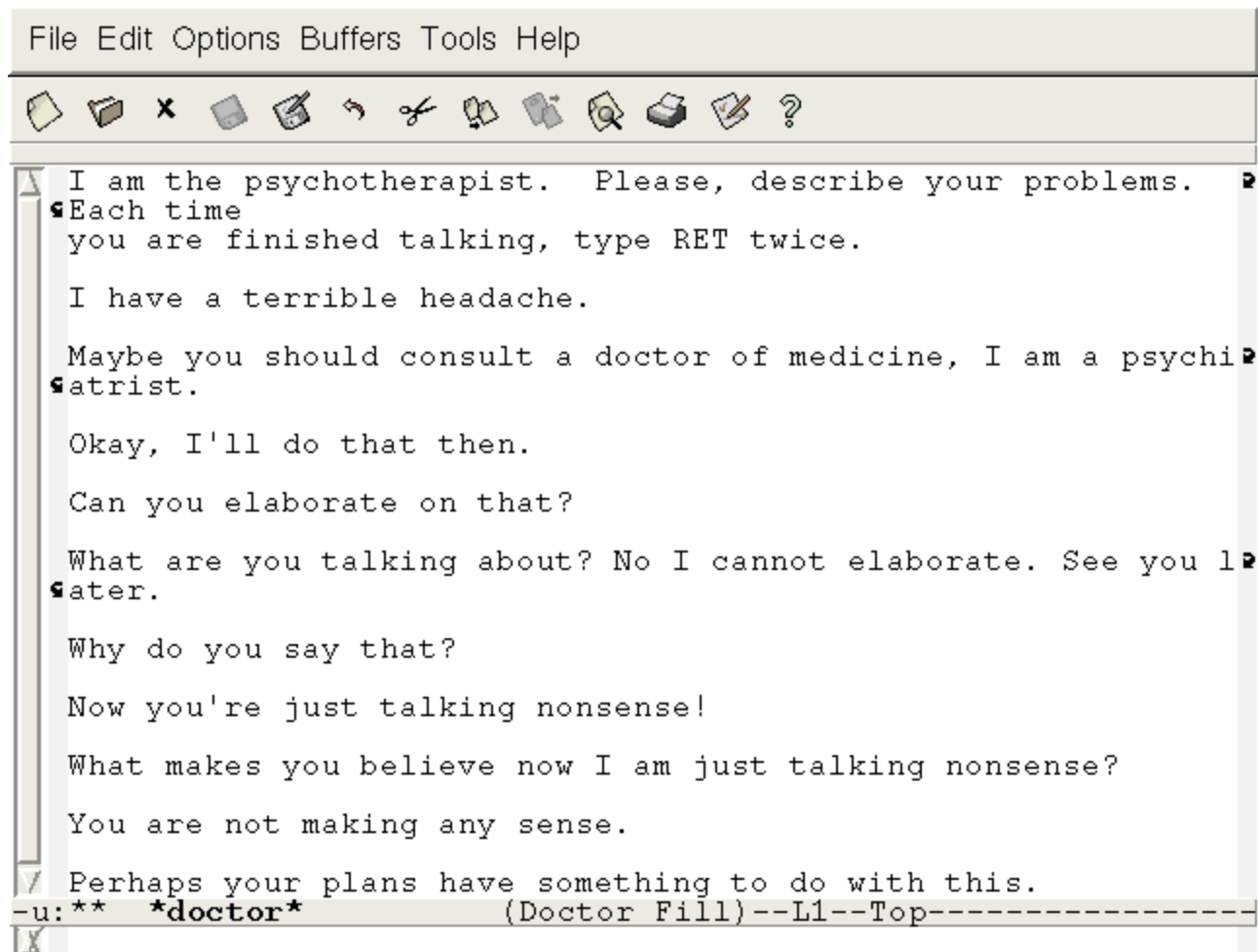


***Can the test conductor find out which is the machine and which the human***

# Inspired *Eliza*

- <http://www.manifestation.com/neurotoys/eliza.php3>

# Inspired *Eliza* (another sample interaction)



The screenshot shows a window titled "Eliza" with a menu bar (File, Edit, Options, Buffers, Tools, Help) and a toolbar with icons for file operations. The main text area contains a dialogue between a user and a program. The program's responses are indented. At the bottom, there is a status bar with the text: -u: \*\* \*doctor\* (Doctor Fill)--L1--Top-----

```
File Edit Options Buffers Tools Help

I am the psychotherapist. Please, describe your problems.
Each time
you are finished talking, type RET twice.

I have a terrible headache.

Maybe you should consult a doctor of medicine, I am a psychiatrist.

Okay, I'll do that then.

Can you elaborate on that?

What are you talking about? No I cannot elaborate. See you later.

Why do you say that?

Now you're just talking nonsense!

What makes you believe now I am just talking nonsense?

You are not making any sense.

Perhaps your plans have something to do with this.
-u: ** *doctor* (Doctor Fill)--L1--Top-----
```

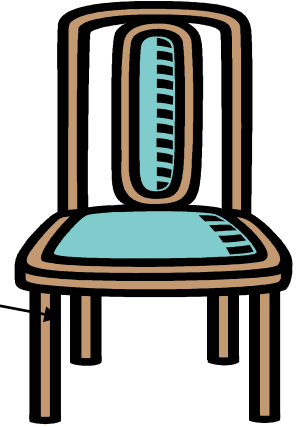
“What is it” question: NLP is concerned with *Grounding*

**Ground the language into perceptual, motor and cognitive capacities.**

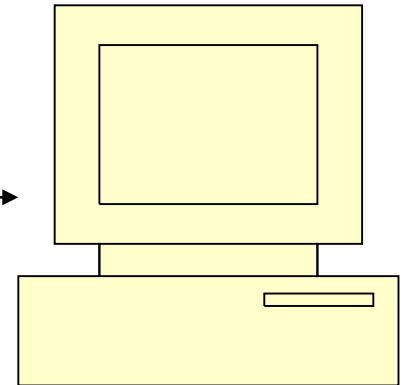


# Grounding

**Chair**



**Computer**

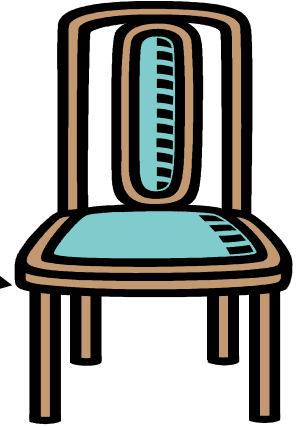


# Grounding faces 3 challenges

- Ambiguity.
- Co-reference resolution  
(*anaphora* is a kind of it).
- Elipsis.

# Ambiguity

**Chair**



# Co-reference Resolution

**Sequence of commands to the robot:**

***Place the wrench on the table.***

***Then paint it.***

**What does *it* refer to?**

# Elipsis

Sequence of command to the Robot:

*Move the table to the corner.*

*Also the chair.*

Second command needs completing by using the first part of the previous command.

# Two Views of NLP and the Associated Challenges

1. Classical View
2. Statistical/Machine Learning View

# Stages of processing *(traditional view)*

- Phonetics and phonology
- Morphology
- Lexical Analysis
- Syntactic Analysis
- Semantic Analysis
- Pragmatics
- Discourse

# Phonetics

- Processing of speech
- Challenges
  - Homophones: *bank (finance)* vs. *bank (river bank)*
  - Near Homophones: *maatras* vs. *maatra (hin)*
  - Word Boundary
    - *aajaayenge (aa jaayenge (will come) or aaj aayenge (will come today)*
    - *I got [ua]plate*
  - Phrase boundary
    - *mtech1 students are especially exhorted to attend as such seminars are integral to one's post-graduate education*
  - Disfluency: *ah, um, ahem etc.*



# Morphology

- Word formation rules from *root* words
- Nouns: Plural (*boy-boys*); Gender marking (czar-czarina)
- Verbs: Tense (*stretch-stretched*); Aspect (*e.g. perfective sit-had sat*); Modality (*e.g. request khaanaa → khaaiie*)
- First crucial first step in NLP
- Languages rich in morphology: e.g., Dravidian, Hungarian, Turkish
- Languages poor in morphology: Chinese, English
- Languages with rich morphology have the advantage of easier processing at higher stages of processing
- A task of interest to computer science: *Finite State Machines for Word Morphology*

# Lexical Analysis

- Essentially refers to dictionary access and obtaining the properties of the word

*e.g. dog*

*noun (lexical property)*

*take-'s'-in-plural (morph property)*

*animate (semantic property)*

*4-legged (-do-)*

*carnivore (-do)*

Challenge: *Lexical or word sense  
disambiguation*

# Lexical Disambiguation

First step: *part of Speech Disambiguation*

- *Dog* as a *noun* (animal)
- *Dog* as a verb (*to pursue*)

Sense Disambiguation

- *Dog* (as *animal*)
- *Dog* (as *a very detestable person*)

Needs word relationships in a context

- *The chair emphasised the need for adult education*

Very common in day to day communications

Satellite Channel Ad: *Watch what you want, when you want* (two senses of watch)

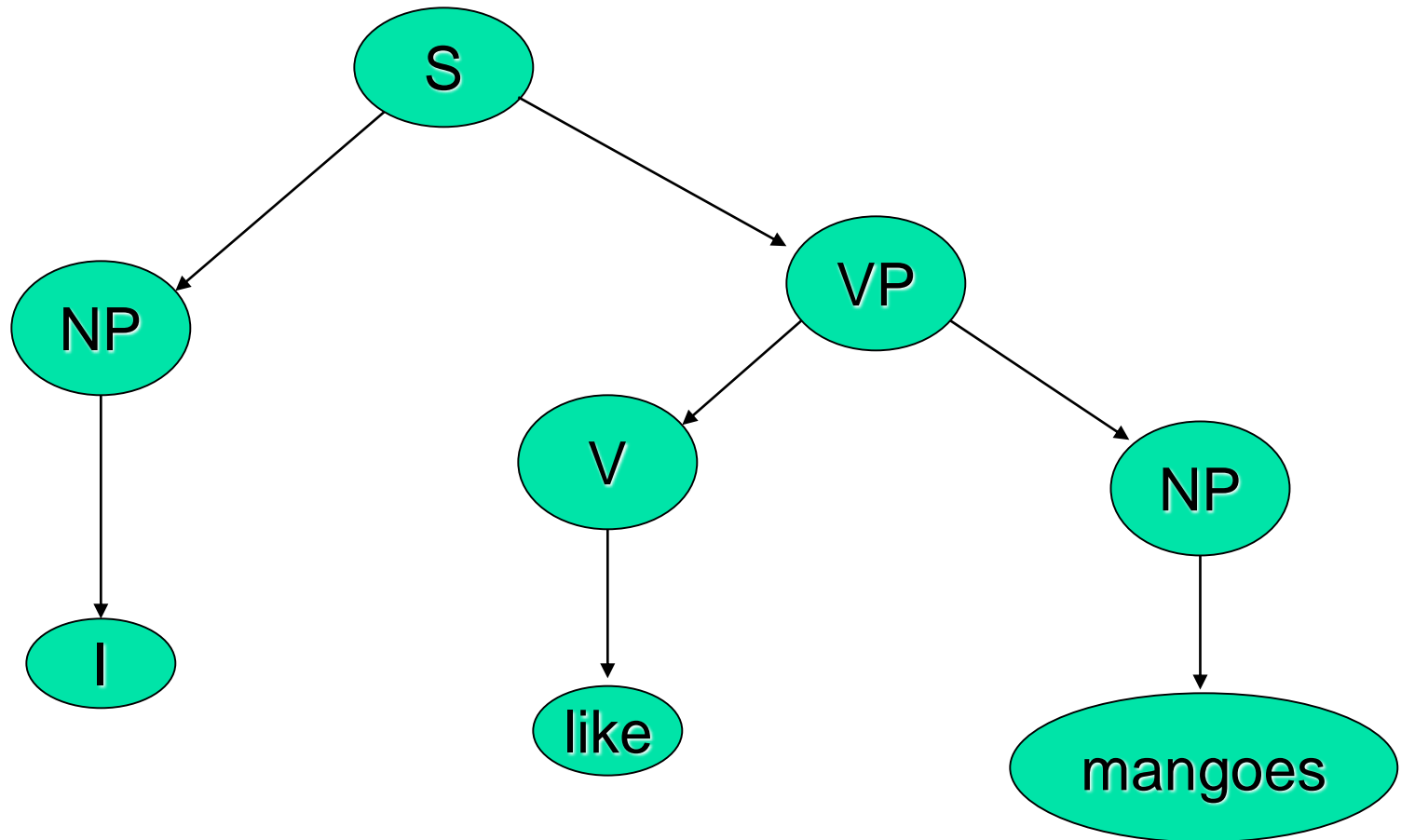
e.g., Ground breaking ceremony/research

Technological developments bring in new terms, additional meanings/nuances for existing terms

- Justify as in *justify the right margin* (word processing context)
- *Xeroxed*: a new verb
- *Digital Trace*: a new expression

# Syntax Processing Stage

## Structure Detection



# Parsing Strategy

- Driven by grammar
  - $S \rightarrow NP VP$
  - $NP \rightarrow N \mid PRON$
  - $VP \rightarrow V NP \mid V PP$
  - $N \rightarrow \text{Mangoes}$
  - $PRON \rightarrow I$
  - $V \rightarrow \text{like}$

# Challenges in Syntactic Processing: Structural Ambiguity

- Scope

1. *The old men and women were taken to safe locations*  
(*old men and women*) vs. (*(old men) and women*)
2. *No smoking areas will allow Hookas inside*

- Preposition Phrase Attachment

- *I saw the boy with a telescope*  
(who has the *telescope*?)
- I saw the mountain with a telescope  
(world knowledge: *mountain* cannot be an *instrument of seeing*)
- I saw the boy with the pony-tail  
(world knowledge: *pony-tail* cannot be an *instrument of seeing*)

Very ubiquitous: newspaper headline "*20 years later, BMC pays father 20 lakhs for causing son's death*"

# Structural Ambiguity...

- Overheard
  - *I did not know my PDA had a phone for 3 months*
- An actual sentence in the newspaper
  - *The camera man shot the man with the gun when he was near Tendulkar*



# Headache for parsing: Garden Path sentences

- Consider

- *The horse raced past the garden* (sentence complete)
- *The old man* (phrase complete)
- *Twin Bomb Strike in Baghdad* (news paper heading: complete)

# Headache for Parsing

- Garden Pathing

- *The horse raced past the garden fell*
- *The old man the boat*
- *Twin Bomb Strike in Baghdad kill 25 (Times of India 5/9/07)*

# Semantic Analysis

- Representation in terms of
  - Predicate calculus/Semantic Nets/Frames/Conceptual Dependencies and Scripts
- *John gave a book to Mary*
  - Give action: Agent: John, Object: Book, Recipient: Mary
- Challenge: ambiguity in semantic role labeling
  - *(Eng) Visiting aunts can be a nuisance*
  - *(Hin) aapko mujhe mithaai khilaanii padegii*  
*(ambiguous in Marathi and Bengali too; not in Dravidian languages)*

# Pragmatics

- Very hard problem
- Model user intention
  - *Tourist (in a hurry, checking out of the hotel, motioning to the service boy): Boy, go upstairs and see if my sandals are under the divan. Do not be late. I just have 15 minutes to catch the train.*
  - *Boy (running upstairs and coming back panting): yes sir, they are there.*
- World knowledge
  - *WHY INDIA NEEDS A SECOND OCTOBER (ToI, 2/10/07)*

# Discourse

Processing of *sequence* of sentences

*Mother to John:*

*John go to school. It is open today. Should you bunk? Father will be very angry.*

Ambiguity of *open*

*bunk* what?

*Why will the father be angry?*

Complex chain of reasoning and application of world knowledge

Ambiguity of *father*

*father* as *parent*

or

*father* as *headmaster*

# Complexity of Connected Text

*John was returning from school  
dejected – today was the math test*

*He couldn't control the class*

*Teacher shouldn't have made him  
responsible*

*After all he is just a janitor*

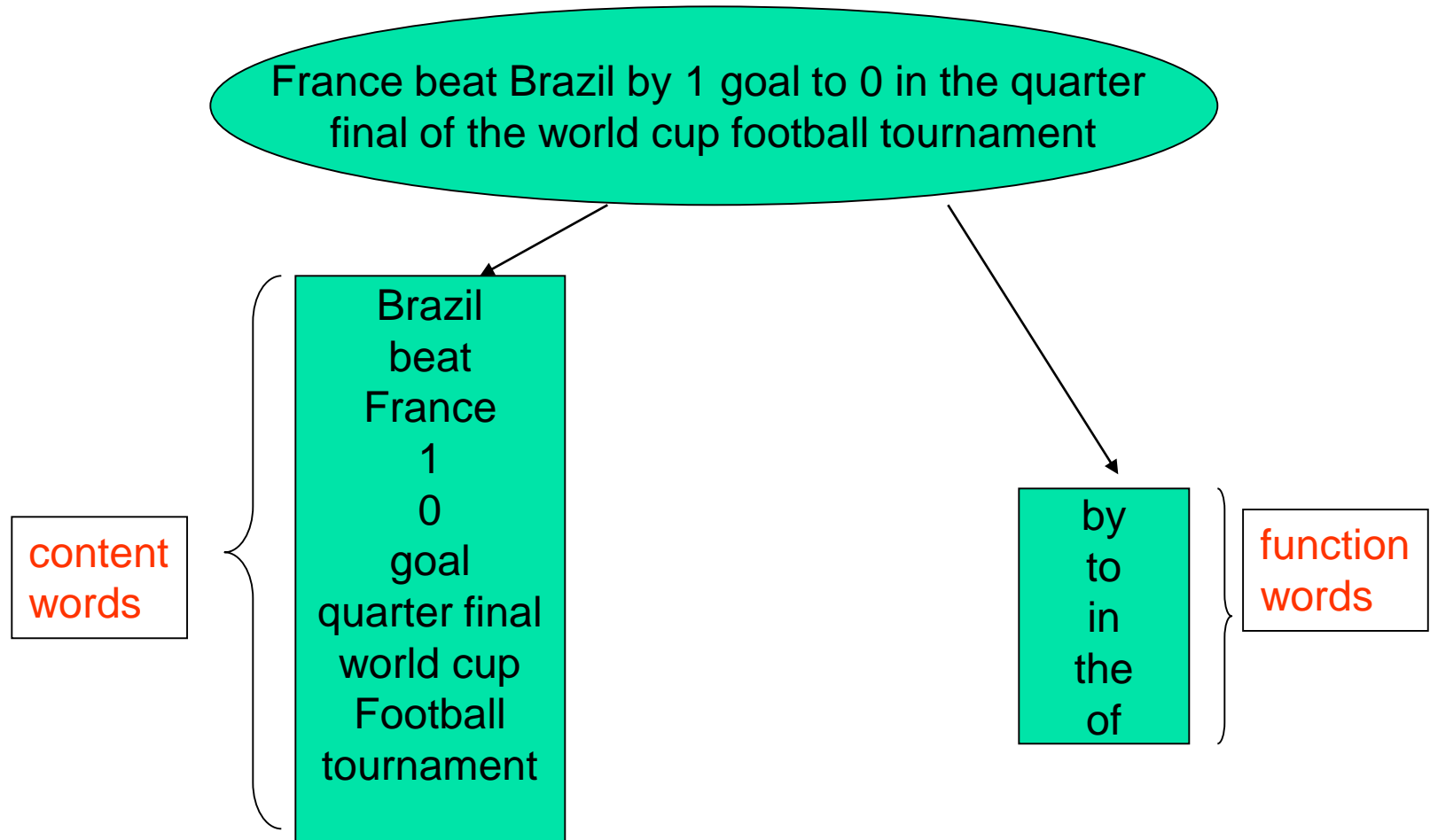
# Machine Learning and NLP

# NLP as an ML task

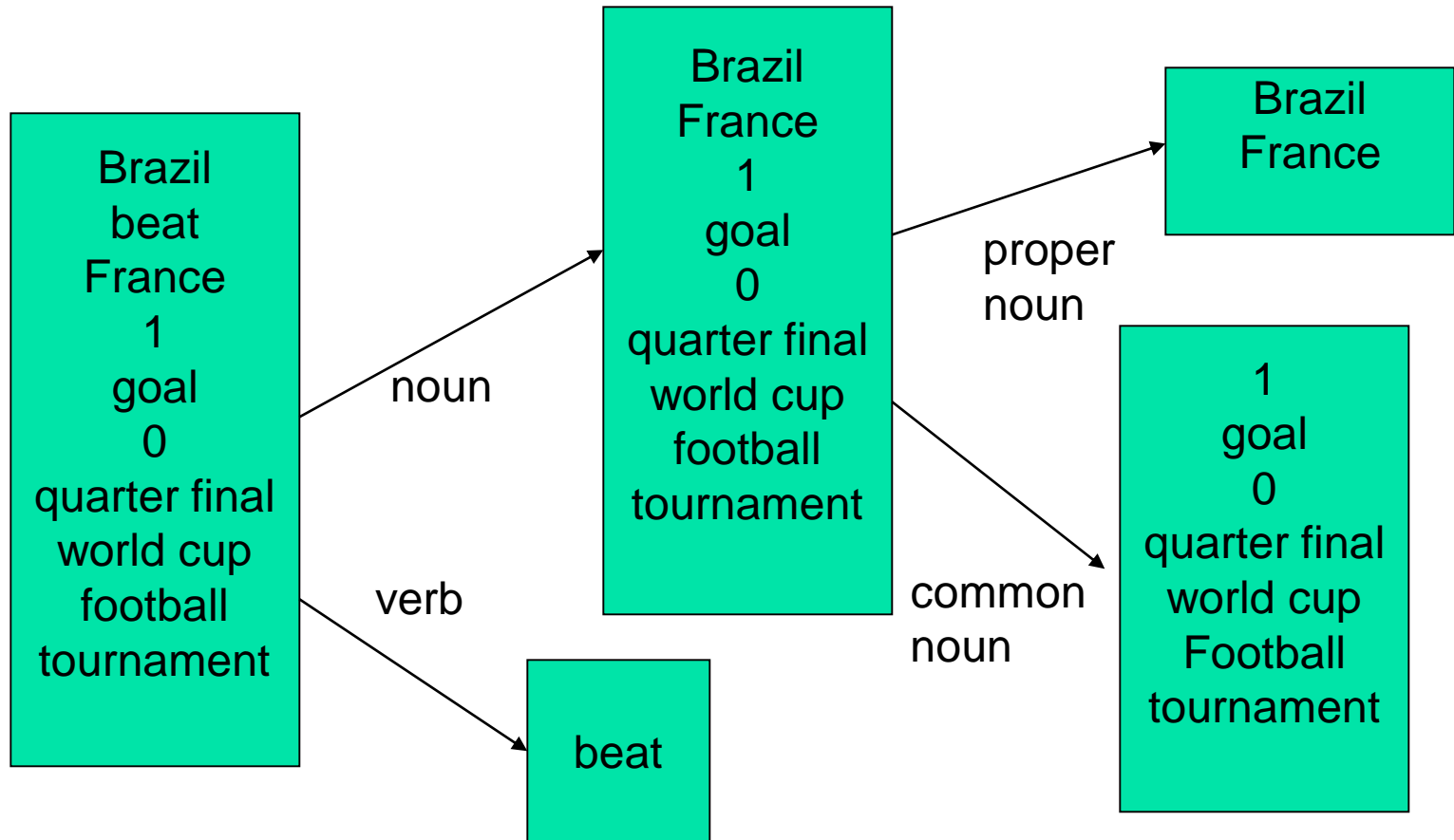
- France *beat Brazil by 1 goal to 0 in the quarter-final of the world cup football tournament. (English)*
- *braazil ne phraans ko vishwa kap phutbal spardhaa ke kwaartaar phaainal me 1-0 gol ke baraabarii se haraayaa. (Hindi)*



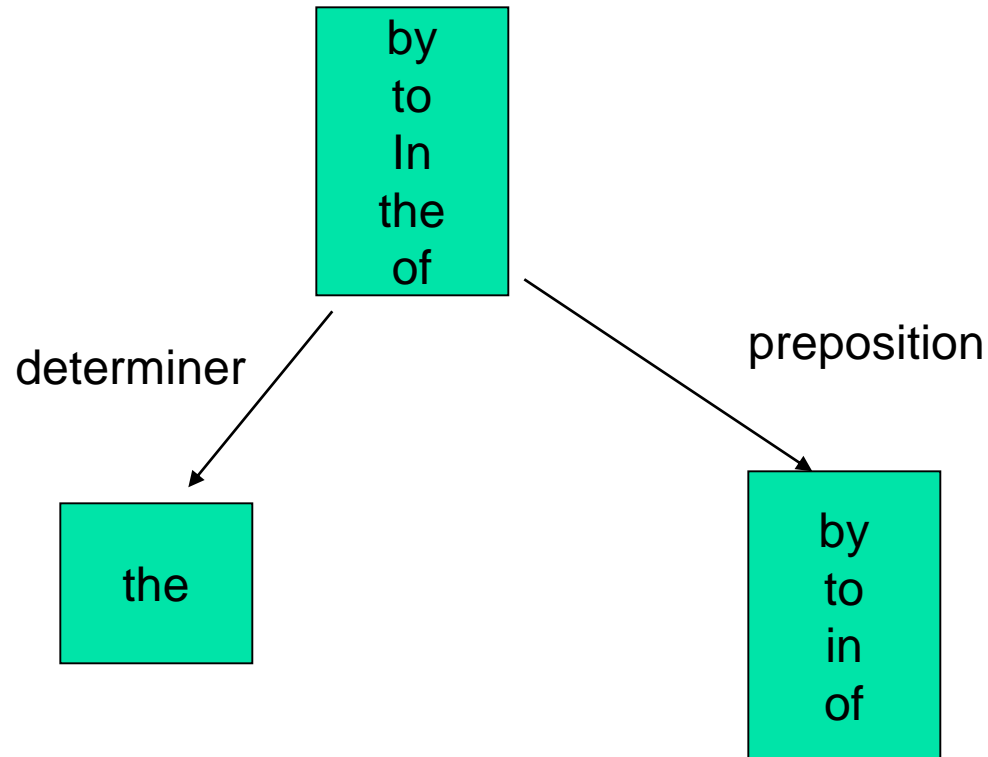
# Categories of the Words in the Sentence



# Further Classification *1/2*



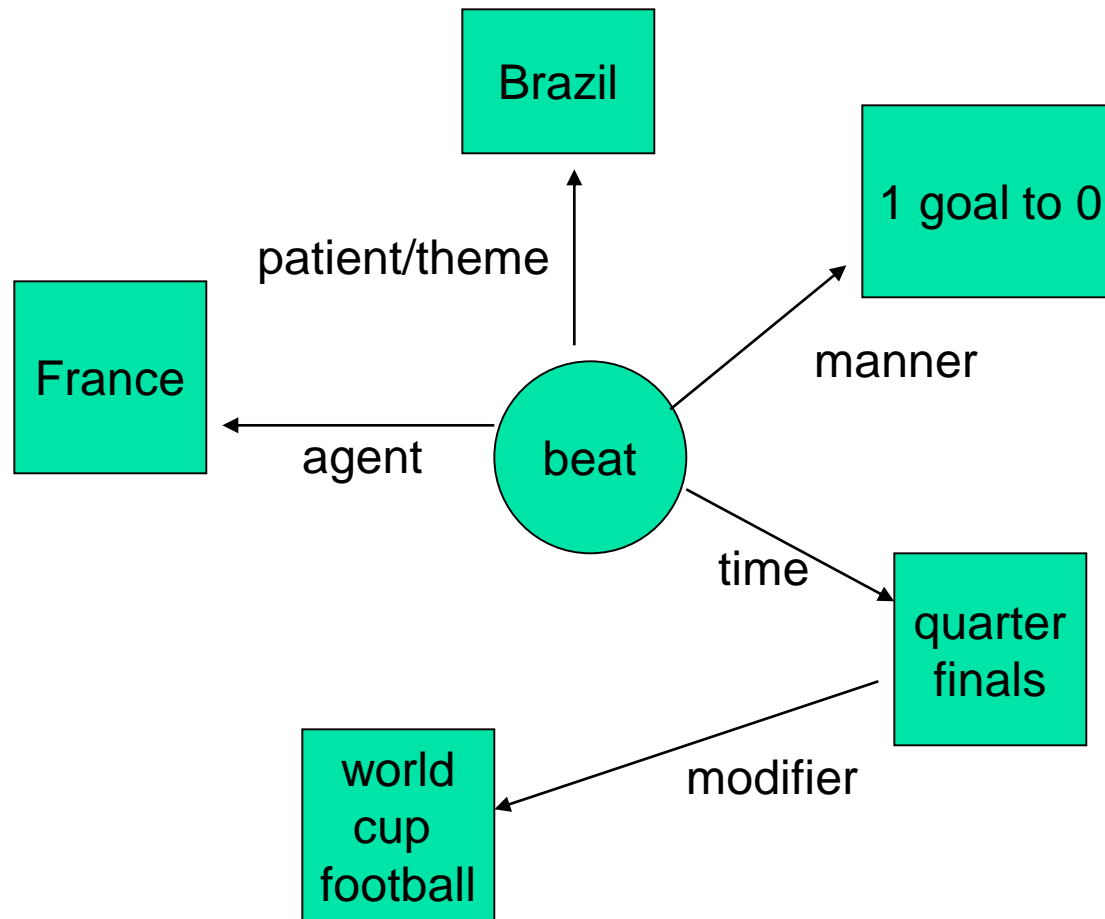
# Further Classification 2/2



# Why all this?

- information need
  - *who did what*
  - *to whom*
  - *by what*
  - *when*
  - *where*
  - *in what manner*

# Semantic roles



# Semantic Role Labeling: *a classification task*

- *France beat Brazil by 1 goal to 0 in the quarter-final of the world cup football tournament*
  - Brazil: *agent or object?*
  - Agent: *Brazil or France or Quarter Final or World Cup?*
- *Given an entity, what role does it play?*
- *Given a role, it is played by which entity?*

# A lower level of classification: Part of Speech (POS) Tag Labeling

- *France beat Brazil by 1 goal to 0 in the quarter-final of the world cup football tournament*
  - beat: *verb of noun (heart beat, e.g.)?*
  - Final: *noun or adjective?*

# Uncertainty in classification:

## Ambiguity

- *Visiting aunts can be a nuisance*
  - Visiting:
    - *adjective or gerund* (POS tag ambiguity)
  - Role of *aunt*:
    - *agent of visit* (aunts are visitors)
    - *object of visit* (aunts are being visited)
- Minimize uncertainty of classification with **cues** from the sentence



# What *cues*?

- Position with respect to the verb:
  - *France* to the left of *beat* and *Brazil* to the right: agent-object role marking (English)
- Case marking:
  - *France* ne (Hindi); ne (Marathi): agent role
  - *Brazil* ko (Hindi); laa (Marathi): object role
- Morphology: *haraaayaa* (hindi); *haravlaa* (Marathi):
  - verb POS tag as indicated by the distinctive suffixes

# Cues are like *attribute-value pairs* prompting machine learning from NL data

- Constituent ML tasks
  - Goal: classification or clustering
  - Features/attributes (word position, morphology, word label *etc.*)
  - Values of features
  - Training data (corpus: annotated or un-annotated)
  - Test data (test corpus)
  - Accuracy of decision (precision, recall, F-value, MAP *etc.*)
  - Test of significance (sample space to generality)

# What is the output of an ML-NLP System (1/2)

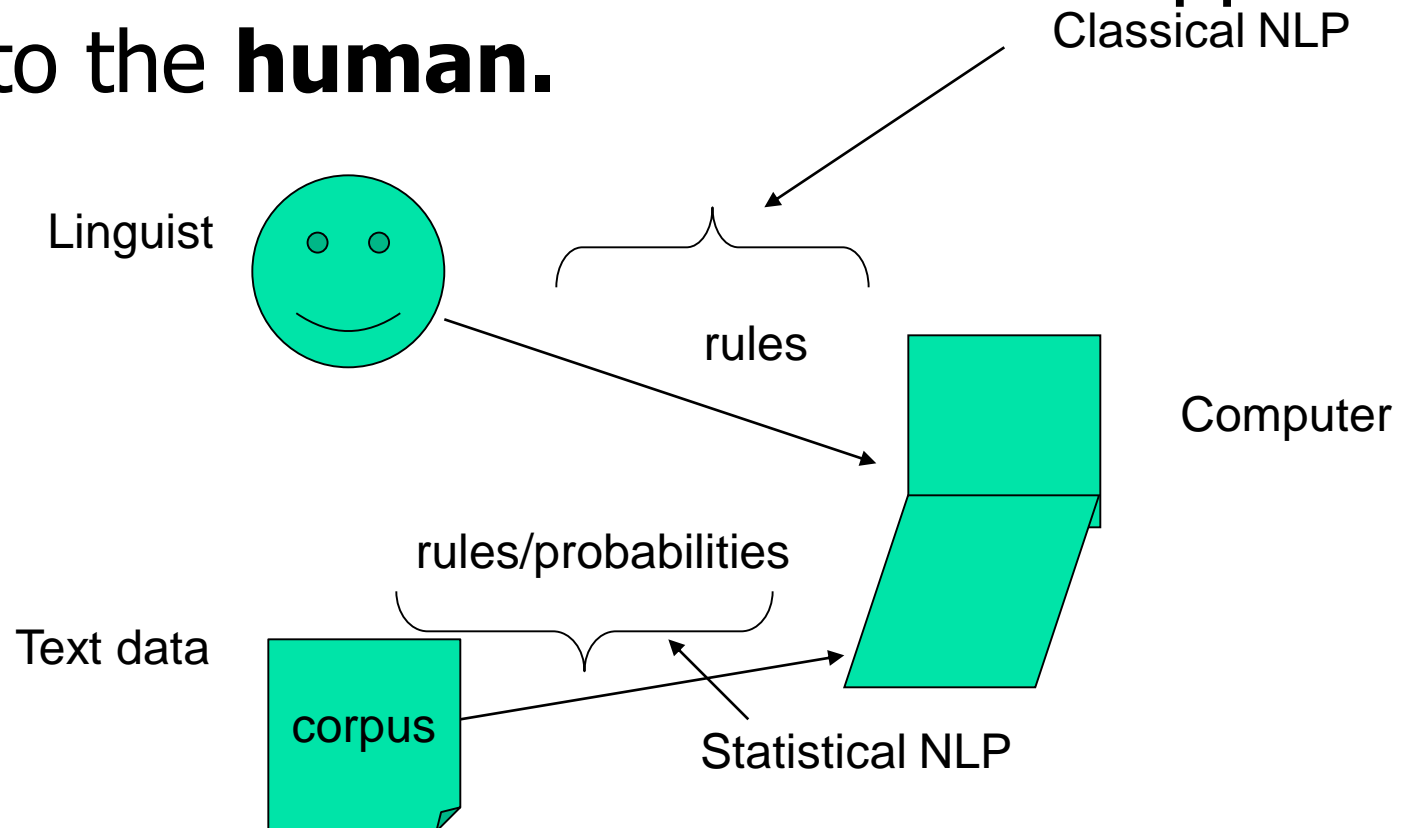
- Option 1: A set of rules, *e.g.*,
  - *If the word to the left of the verb is a noun and has animacy feature, then it is the likely **agent** of the action denoted by the verb.*
    - *The child broke the toy (child is the agent)*
    - *The window broke (window is not the agent; inanimate)*

# What is the output of an ML-NLP System (2/2)

- Option 2: a set of probability values
  - *$P(\text{agent} \mid \text{word is to the left of verb and has animacy}) > P(\text{object} \mid \text{word is to the left of verb and has animacy}) > P(\text{instrument} \mid \text{word is to the left of verb and has animacy})$  etc.*

# How is this different from classical NLP

- The burden is on the **data** as opposed to the **human**.



Classification appears as  
sequence labeling

# A set of Sequence Labeling Tasks: *smaller to larger units*

- *Words:*
  - Part of Speech tagging
  - Named Entity tagging
  - Sense marking
- *Phrases:* Chunking
- *Sentences:* Parsing
- *Paragraphs:* Co-reference annotating

# Example of word labeling: POS Tagging

<s>

Come January, and the IIT campus is abuzz with  
new and returning students.

</s>



<s>

Come\_VB January\_NNP ,\_, and\_CC the\_DT  
IIT\_NNP campus\_NN is\_VBZ abuzz\_JJ with\_IN  
new\_JJ and\_CC returning\_VBG students\_NNS .\_.  
</s>



# Example of word labeling: Named Entity Tagging

<month\_name>

January

</month\_name>

<org\_name>

IIT

</org\_name>

# Example of word labeling: Sense Marking

<u>Word</u>	<u>Synset</u>	<u>WN-synset-</u>
<u>no</u>		
<i>come</i>	<i>{arrive, get, come}</i>	<i>01947900</i>
	.	
	.	
	.	
<i>abuzz</i>	<i>{abuzz, buzzing, droning}</i>	<i>01859419</i>

# Example of phrase labeling: Chunking

Come July, and the IIT campus is

abuzz with new and returning students .

# Example of Sentence labeling: Parsing

[S<sub>1</sub>[S[S[VP[VB Come][NP[NNP July]]]]]

[,]

[CC and]

[S [NP [DT the] [JJ IIT] [NN campus]]

[VP [AUX is]

[ADJP [JJ abuzz]

[PP[IN with]

[NP[ADJP [JJ new] [CC and] [ VBG returning]]

[NNS students]]]]]]

[.]]]

# Modeling Through the Noisy Channel

5 problems in NLP

## 5 Classical Problems in NLP: being tackled now by statistical approaches

- Part of Speech Tagging
- Statistical Spell Checking
- Automatic Speech Recognition
- Probabilistic Parsing
- Statistical Machine Translation

# Problem-1: PoS tagging

## **Input:**

1. sentences (string of words to be tagged)
2. tagset

**Output:** single best tag for each word

# PoS tagging: Example

Sentence:

*The national committee remarked on a number of other issues.*

*Tagged output:*

*The/DET national/ADJ committee/NOU remarked/VRB  
on/PRP a/DET number/NOU of/PRP other/ADJ  
issues/NOU.*



# Stochastic Models (Contd..)

Best tag  $t^*$ ,

$$t^* = \arg \max_t P(t \mid w)$$

Bayes Rule gives,

$$P(t \mid w) = \frac{P(t)P(w \mid t)}{P(w)} = \frac{P(w, t)}{P(w)}$$

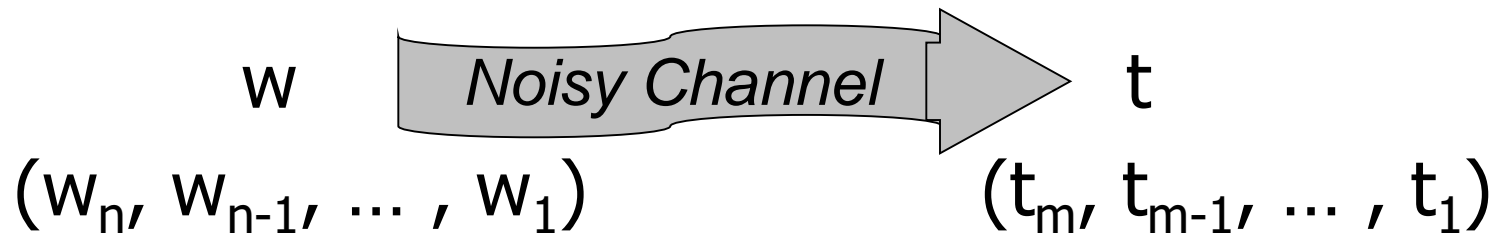
$$P(w, t)$$

Joint Distribution

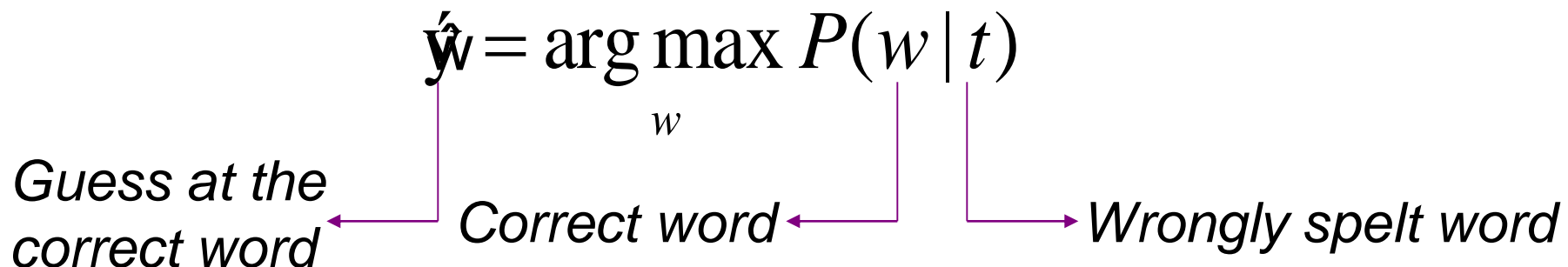
$$P(t \mid w)$$

Conditional Distribution

# Problem 2: Probabilistic Spell Checker



Given  $t$ , find the most probable  $w$  : Find that  $\hat{w}$  for which  $P(w|t)$  is maximum, where  $t$ ,  $w$  and  $\hat{w}$  are strings:



# Spell checker: apply Bayes Rule

$$\hat{w} = \arg \max_w (p(w) \cdot p(t | w))$$

- Why apply Bayes rule?
  - Finding  $p(w/t)$  vs.  $p(t/w)$  ?
- $p(w/t)$  or  $p(t/w)$  have to be computed by counting  $c(w,t)$  or  $c(t,w)$  and then normalizing them
- Assumptions :
  - $t$  is obtained from  $w$  by a single error.
  - The words consist of only alphabets

# Spell checker: Confusion Matrix (1/3)

Confusion Matrix: 26x26

- Data structure to store  $c(a,b)$
- Different matrices for insertion, deletion, substitution and transposition
- *Substitution*
  - The number of instances in which a is wrongly substituted by b in the training corpus (denoted  $\text{sub}(x,y)$  )

# Confusion Matrix (2/3)

- *Insertion*

- The number of times a letter  $y$  is inserted after  $x$  wrongly ( denoted  $\text{ins}(x,y)$  )

- *Transposition*

- The number of times  $xy$  is wrongly transposed to  $yx$  ( denoted  $\text{trans}(x,y)$  )

- *Deletion*

- The number of times  $y$  is deleted wrongly after  $x$  ( denoted  $\text{del}(x,y)$  )

# Confusion Matrix (3/3)

- If  $x$  and  $y$  are alphabets,
  - $\text{sub}(x,y) = \#$  times  $y$  is written for  $x$  (substitution)
  - $\text{ins}(x,y) = \#$  times  $x$  is written as  $xy$
  - $\text{del}(x,y) = \#$  times  $xy$  is written as  $x$
  - $\text{trans}(x,y) = \#$  times  $xy$  is written as  $yx$

# Probabilities from confusion matrix

- $P(t/w) = P(t/w)_S + P(t/w)_I + P(t/w)_D + P(t/w)_X$

where

$$P(t/w)_S = \text{sub}(x,y) / \text{count of } x$$

$$P(t/w)_I = \text{ins}(x,y) / \text{count of } x$$

$$P(t/w)_D = \text{del}(x,y) / \text{count of } x$$

$$P(t/w)_X = \text{trans}(x,y) / \text{count of } x$$

- These are considered to be mutually exclusive events

# Spell checking: Example

- Correct document has  $w_s$
- Wrong document has  $t_s$
- $P(\text{maple}|\text{aple}) =$   
     $\# (\text{maple was wanted instead of aple}) / \# (\text{aple})$
- $P(\text{apple}|\text{aple})$  and  $P(\text{applet}|\text{aple})$  calculated similarly
- Leads to problems due to data sparsity.
- Hence, use Bayes rule.



# Problem 3: Probabilistic Speech Recognition

- **Problem Definition : Given a sequence of speech signals, identify the words.**
- **2 steps :**
  - **Segmentation (Word Boundary Detection)**
  - **Identify the word**
- **Isolated Word Recognition :**
  - **Identify  $W$  given  $SS$  (speech signal)**

$$\hat{W} = \arg \max_W P(W | SS)$$

# Speech recognition: Identifying the word

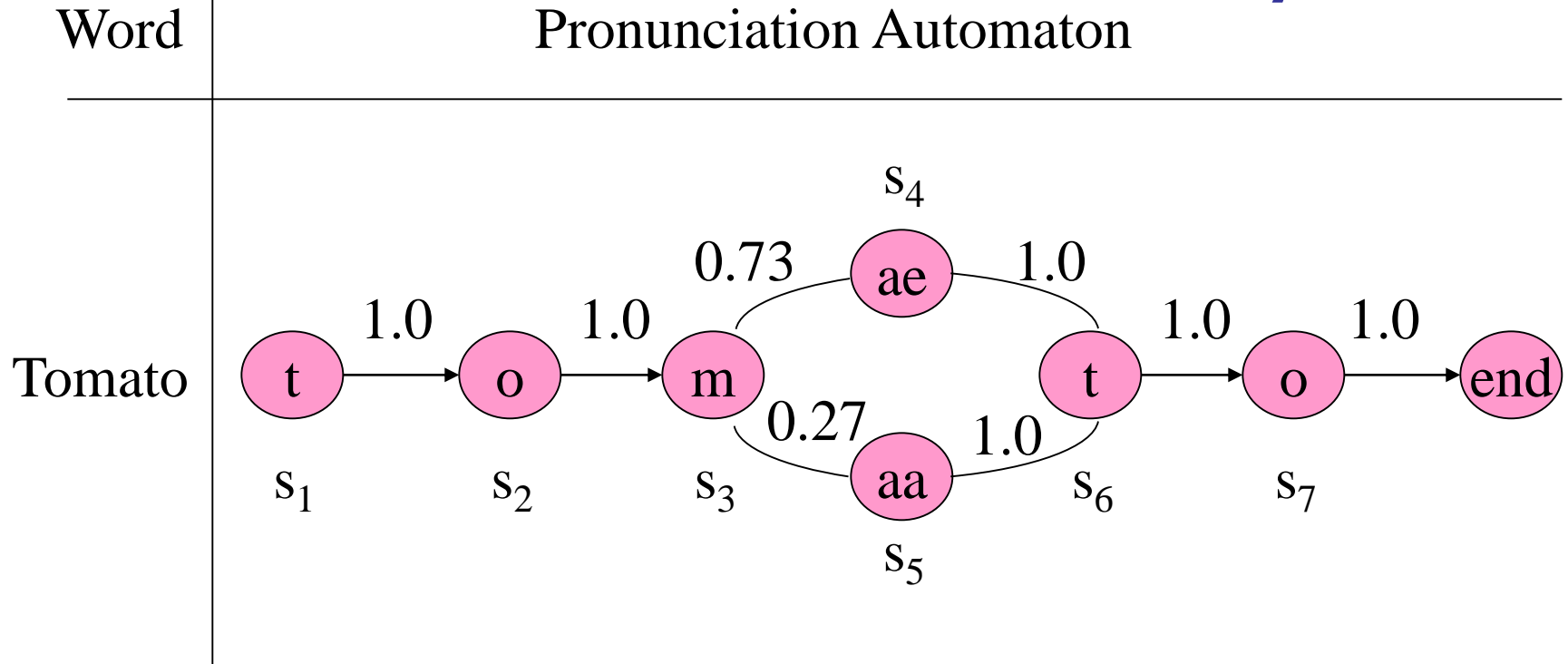
$$\begin{aligned}\hat{W} &= \arg \max_W P(W | SS) \\ &= \arg \max_W P(W)P(SS | W)\end{aligned}$$

- $P(SS/W)$  = likelihood called “phonological model” → intuitively more tractable!
- $P(W)$  = prior probability called “language

$$P(W) = \frac{\# \text{ } W \text{ appears in the corpus}}{\# \text{ words in the corpus}}$$

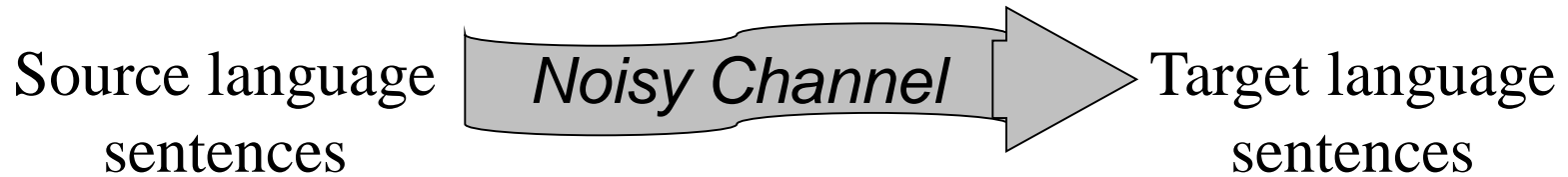
# Pronunciation Dictionary

## Pronunciation Automaton



- $P(SS/W)$  is maintained in this way.
- $P(t o m a e t o / \text{Word is "tomato"})$  = Product of arc probabilities

# Problem 4: Statistical Machine Translation



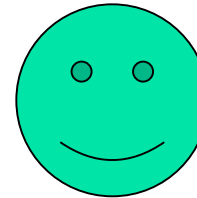
- What sentence in the target language will maximise the probability

$$P(\textit{target sentence}/\textit{source sentence})$$

# Statistical MT: Parallel Texts

- Parallel texts
  - Instruction manuals
  - Hong Kong legislation
  - Macao legislation
  - Canadian parliament Hansards
  - United nation reports
  - Official journal of the European Communities
  - Trilingual documents in Indian states

- Observation:



Every time I see *banco*, the translation is *bank* or *bench* ... if it is *banco de*, then it always becomes *bank* and never *bench*

Courtesy: a presentation by K. Knight

# SMT: formalism

- Source language:  $F$
- Target language:  $E$
- Source language sentence:  $f$
- Target language sentence:  $e$
- Source language word:  $w_f$
- Target language word:  $w_e$

# SMT Model

- To translate  $f$ :
  - Assume that all sentences in  $E$  are translations of  $f$  with some probability!
  - Choose the translation with the highest probability

$$\hat{e} = \arg \max_e (p(e | f))$$

# SMT: Apply Bayes Rule

$$\hat{e} = \arg \max_e (p(e).p(f | e))$$

$P(e)$  is called the **language model** and stands for **fluency**

and

$P(f|e)$  is called the **translation model** and stands for **faithfulness**



# Reason for Applying Bayes Rule

- The way  $P(f/e)$  and  $P(e/f)$  are usually calculated
  - Word translation based
  - Word order
  - Collocations (For example, *strong tea*)
- Example:
  - $f$ : *It is raining*
  - Candidates for  $e$  (in Hindi):
    - *bAriSa Ho raHI HE* (rain happening is)
    - *Ho bAriSa raHI HE* (is rain happening)
    - *bAriSa Ho raHA HE* (rain happening\_masculine is)

Is NLP Really Needed

# Post-1

- POST----5 TITLE: "Wants to invest in IPO? Think again" | <br /><br />Here's a sobering thought for those who believe in investing in IPOs. Listing gains "the return on the IPO scrip at the close of listing day over the allotment price" have been falling substantially in the past two years. Average listing gains have fallen from 38% in 2005 to as low as 2% in the first half of 2007. Of the 159 book-built initial public offerings (IPOs) in India between 2000 and 2007, two-thirds saw listing gains. However, these gains have eroded sharply in recent years. Experts say this trend can be attributed to the aggressive pricing strategy that investment bankers adopt before an IPO. "While the drop in average listing gains is not a good sign, it could be due to the fact that IPO issue managers are getting aggressive with pricing of the issues," says Anand Rathi, chief economist, Sujan Hajra. While the listing gain was 38% in 2005 over 34 issues, it fell to 30% in 2006 over 61 issues and to 2% in 2007 till mid-April over 34 issues. The overall listing gain for 159 issues listed since 2000 has been 23%, according to an analysis by Anand Rathi Securities. Aggressive pricing means the scrip has often been priced at the high end of the pricing range, which would restrict the upward movement of the stock, leading to reduced listing gains for the investor. It also tends to suggest investors should not indiscriminately pump in money into IPOs. But some market experts point out that India fares better than other countries. "Internationally, there have been periods of negative returns and low positive returns in India should not be considered a bad thing."

# Post-2

- POST-----7TITLE: "[IIM-Jobs] \*\*\*\*\* Bank: International Projects Group - Manager"| <br />Please send your CV & cover letter to anup.abraham@\*\*\*\*\*bank.com \*\*\*\*\* Bank, through its International Banking Group (IBG), is expanding beyond the Indian market with an intent to become a significant player in the global marketplace. The exciting growth in the overseas markets is driven not only by India linked opportunities, but also by opportunities of impact that we see as a local player in these overseas markets and / or as a bank with global footprint. IBG comprises of Retail banking, Corporate banking & Treasury in 17 overseas markets we are present in. Technology is seen as key part of the business strategy, and critical to business innovation & capability scale up. The International Projects Group in IBG takes ownership of defining & delivering business critical IT projects, and directly impact business growth. Role: Manager & International Projects Group Purpose of the role: Define IT initiatives and manage IT projects to achieve business goals. The project domain will be retail, corporate & treasury. The incumbent will work with teams across functions (including internal technology teams & IT vendors for development/implementation) and locations to deliver significant & measurable impact to the business. Location: Mumbai (Short travel to overseas locations may be needed) Key Deliverables: Conceptualize IT initiatives, define business requirements

# Sentiment Classification

- Positive, negative, neutral – 3 class
- Sports, economics, literature - multi class
- Create a representation for the document
- Classify the representation

The most popular way of representing a document is feature vector (indicator sequence).

# Established Techniques

- Naïve Bayes Classifier (NBC)
- Support Vector Machines (SVM)
- Neural Networks
- K nearest neighbor classifier
- Latent Semantic Indexing
- Decision Tree ID3
- Concept based indexing

# Successful Approaches

The following are successful approaches as reported in literature.

- NBC – simple to understand and implement
- SVM – complex, requires foundations of perceptions

# Mathematical Setting

We have training set

A: Positive Sentiment Docs

B: Negative Sentiment Docs

Indicator/feature  
vectors to be formed

Let the class of positive and negative documents be  $C_+$  and  $C_-$ , respectively.

Given a new document **D** label it positive if

$$P(C_+|D) > P(C_-|D)$$



# Priori Probability

Docu ment	Vector	Classif ication
D1	V1	+
D2	V2	-
D3	V3	+
..	..	..
$D_{4000}$	$V_{4000}$	-

Let  $T$  = Total no of documents

And let  $|+| = M$

So,  $|-| = T - M$

$$P(\text{D being positive}) = M/T$$

Priori probability is calculated without considering any features of the new document.

# Apply Bayes Theorem

Steps followed for the NBC algorithm:

- Calculate Prior Probability of the classes.  $P(C_+)$  and  $P(C_-)$
- Calculate feature probabilities of new document.  $P(D|C_+)$  and  $P(D|C_-)$
- Probability of a document **D** belonging to a class **C** can be calculated by Baye's Theorem as follows:

$$P(C|D) = \frac{P(C) * P(D|C)}{P(D)}$$

- Document belongs to  $C_+$  , if

$$P(C_+) * P(D|C_+) > P(C_-) * P(D|C_-)$$

# Calculating $P(D|C_+)$

$P(D|C_+)$  is the probability of class  $C_+$  given  $D$ . This is calculated as follows:

- Identify a set of features/indicators to evaluate a document and generate a feature vector ( $V_D$ ).  $V_D = \langle x_1, x_2, x_3 \dots x_n \rangle$

- Hence,  $P(D|C_+) = P(V_D|C_+)$

$$\begin{aligned} &= P(\langle x_1, x_2, x_3 \dots x_n \rangle | C_+) \\ &= \frac{|\langle x_1, x_2, x_3 \dots x_n \rangle, C_+|}{|C_+|} \end{aligned}$$

- Based on the assumption that all features are Independently Identically Distributed (IID)

$$\begin{aligned} &= P(\langle x_1, x_2, x_3 \dots x_n \rangle | C_+) \\ &= P(x_1 | C_+) * P(x_2 | C_+) * P(x_3 | C_+) * \dots * P(x_n | C_+) \\ &= \prod_{i=1}^n P(x_i | C_+) \end{aligned}$$

- $P(x_i | C_+)$  can now be calculated as  $\frac{|x_i|}{|C_+|}$

# Baseline Accuracy

- Just on Tokens as features, **80%** accuracy
- 20% probability of a document being misclassified
- On large sets this is significant



# To improve accuracy...

---

- Clean corpora
- POS tag
- Concentrate on critical POS tags (e.g. *adjective*)
- Remove 'objective' sentences ('of' ones)
- Do aggregation

Use minimal to sophisticated NLP