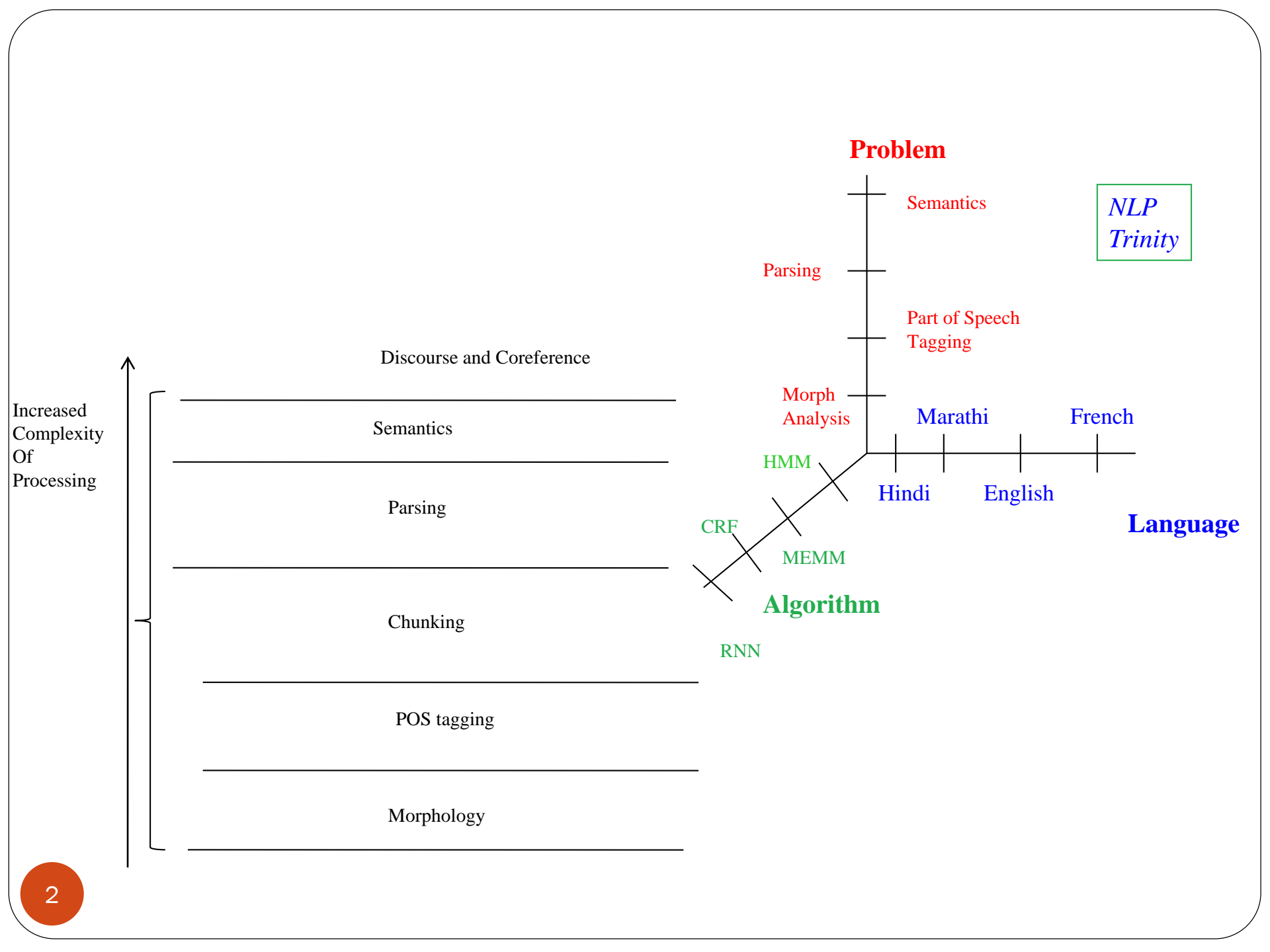


Introduction to Natural Language Processing

Asif Ekbal

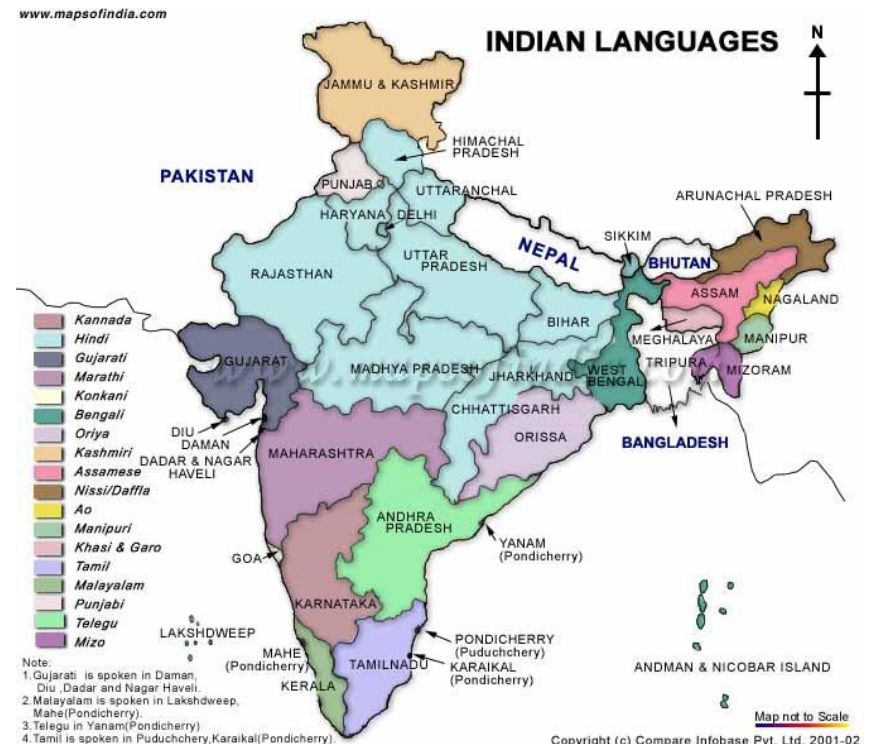
Dept. of Computer Science and Engineering
IIT Patna, Patna, India

Email: asif@iitp.ac.in, asif.ekbal@gmail.com



Multilinguality: Indian situation

- Major streams
 - Indo European
 - Dravidian
 - Sino Tibetan
 - Austro-Asiatic
- Some languages are ranked within 20 in in the world in terms of the populations speaking them
 - Hindi : 4th (~350 milion)
 - Bangla: 5^h (~230 million)
 - Marathi 10th (~84 million)



*Language Technology or Natural Language
Processing: Background & Relevance in
Indian Scenario*

Background: Indian Context

- India is a multi-lingual country with great linguistic and cultural diversities
- 22 official languages mentioned in the Indian constitution
- However, Census of India in 2001 reported-
 - **122 major languages**
 - **1,599 other regional languages**
 - **2,371 scripts**
 - **30 languages** are spoken by more than **one million native speakers**
 - **122** are spoken by more than **10,000 people**
- **20%** understand English
- **80%** cannot understand

Background

- Phenomenal growth in the number of internet users, social media (*Facebook, Twitter* etc.)
- Increasing tendency of using Indian language contents for exchanging information
- **Digital divide** cannot be tackled unless citizens are given flexibility in **communicating in their own languages**

Language Technology or Natural Language Processing (NLP) that deals with developing theories and techniques for effective communication in human languages play an important role towards creating this digital society

TDIL: MeitY, Govt. of India

- Technology Development for Indian Languages (TDIL) Programme
- **Objective:**
 - developing Information Processing Tools and Techniques to facilitate human-machine interaction without language barrier;
 - creating and accessing multilingual knowledge resources; and
 - integrating them to develop innovative user products and services

TDIL: Some major initiatives

- Development of English to Indian Language Machine Translation (**Anuvadakh**):

English to Hindi/Marathi/Bangla/Oriya/Tamil/Urdu/Gujrati/Bodo

- Development of English to Indian Language Machine Translation System with Angla-Bharti Technology: English to Bangla/Punjabi/Malaylam/Urdu/Hindi/Telugu

- Development of Indian Language to Indian Language Machine Translation System (**Sampark**)- 18 pairs of languages

-Hindi to Bengali, Bengali to Hindi, Marathi to Hindi, Hindi to Marathi, Hindi to Punjabi, Punjabi to Hindi, Hindi to Tamil, Tamil to Hindi, Hindi to Kannada, Kannada to Hindi, Hindi to Telugu, Telugu to Hindi, Hindi to Urdu, Urdu-Hindi, Malaylam to Tamil, Tamil to Malaylam, Tamil to Telugu, Telugu to Tamil

TDIL: Some major initiatives

- Development of Cross-Lingual Information Access (CLIA)
 - Assamese, Bengali, Hindi, Oriya, Punjabi, Tamil, Telugu, Marathi
- Development of Robust Document Analysis & Recognition System for Indian Languages (OCR)-14 languages
 - Assamese, Bengali, Devanagri, Gujrati, Gurumukhi, Kannada, Malaylam, Manipuri, Marathi, Oriya, Tamil, Telugu, Tibetan, Urdu
- Development of Text to Speech System in Indian Languages
- Development of Automatic Speech Recognition System in Indian Languages
- *Development of Hindi to English Machine Translation in Judicial Domain*

A Case-Study: **MyGov.in Portal**

Govt. Portal: MyGov.in



“ Let us join this mass movement towards **Surajya**, Realise the hopes and aspirations of the people and take India to greater heights! ”

TRENDING



Finalizing Bihar Sharif Smart City Proposal Round IV



Vidyanjali - (School Volunteer Programme)

IN FOCUS



BIRAC-Innovation Challenge Award'17: Solutions for Community Health (SoCH)



TRAI Invites Suggestions on Consultation Paper on 'In-Flight'

Swachh Bharat (Clean India) ▼



Do



Discuss



Poll/Survey



Blog



Talk

Sort By : Newest First 🔍

Cleanliness in school curriculum

How can schools innovatively include 'focusing on cleanliness' as a part of their curriculum where students learn about the need and importance of cleanliness from a ...

[See details](#) ▼

All Comments

Showing 15252 Submission(s)



NIKUNJ BHATT 2 years 4 months ago

Sir, curriculum mai safai abhiyan ki jankari dene se pehle ye dekhna jaruri hai k desh k her school, specially govt. schools mai sweeper ki facility ho ubi toh ye kam students khud kar rahe hai vo study kub karegae. Rural area mai halat bahut kharab hai. teachers pe already itna work load hai vo ye vyawastha kaise kare.keval fund dena samadhan nahi . ground level pe bahut se ese samasyaye hai jo keval teachers jantae hai . policymakers nahi

Govt. Portal: MyGov.in

- **Citizen-centric platform** empowers people to connect with the Government & contribute towards good governance
- Unique first of its kind participatory governance initiative involving the common citizen at large
- Idea is to bring the government closer to the common man by the **use of online platform** creating an interface for **healthy exchange of ideas** and **views** involving the common citizen and experts
- Ultimate goal is to contribute to the **social and economic transformation of India**
- Was launched on July 26, 2014 by the Hon'ble PM

Govt. Portal: MyGov.in

- This has been more than successful in keeping the citizens engaged on important policy issues and governance, be it **Clean Ganga, Girl Child Education, Skill Development** and **Healthy India** to name a few
- Has become a key part of the **policy and decision making** process of the country
- Platform has been able
 - to provide the citizens a voice in the governance process of the country and
 - create grounds for the citizens to become stakeholders not only in policy formulation and recommendation but also implementation through actionable tasks

Govt. Portal: MyGov.in

- **Major attributes:** Discussion, Tasks, Talks, Polls and Blogs on various groups based on the diverse governance and public policy issues
- Has more than **1.78 Million users** who contribute their ideas through discussions and also participate through the various earmarked tasks
- Platform gets more than **10,000 posts per weeks** on various issues

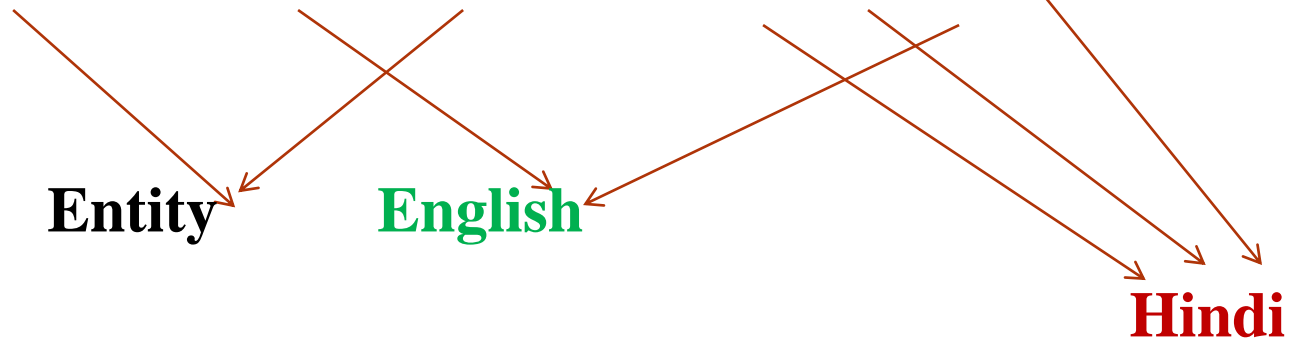
Feedbacks are analyzed and put together as suggestions for the concerned departments which are responsible to transform them into actionable agenda

- Infeasible to **mine** the most **relevant information** from this huge data
- Needs a method for automated analysis of this data
 - **Demands sophisticated NLP and ML techniques to build these**

Code-mixing

- Code-mixing refers to the mixing of two or more languages or language varieties in speech/text

Kolkata to Varanasi ka kya distance hai



Code-Mixing in MyGov.in: Few Examples

- *Sir ji aapka ye abhiyan acha ha isse naye bharat ka nirman hoga maine apne school ke student ke sath milkar hospital ki safai ki and jagrukta rali nikali jisse log gandagi kam failaye.*
- *Aaj her school main swachta abhiyan honi chye we do it*
- *india ko clean rakhne ke lie gandgi karne walo pe penalty lagani chahiye jo kaam das sal me hoga penalty lagane ke bad wo kuch hi dino me ho jaega*
- *Modi sir swachh bharat m aapke bjp poltician photo click krawane k liye safai krte h sathinye neta sirf pik click krte h bs.*
- *Our School also participated in Clean India Campaign . The students of class XII cleaned a Park and a Basket Ball area .*

Why to Analyse?

- Public opinions play important roles for the betterment of human lives
- Huge volumes and varieties of user-generated contents and user interaction networks constitute new opportunities for understanding social behavior
- Understanding deep feeling of public can help government to anticipate deep social changes and adapt to population expectations

Discipline known as Opinion Mining or Sentiment Analysis

NLP: Projected Growth

- *Growing in an exponential manner*
- *Expected to touch the market of **\$16 billion in 2021***
 - *With compound growth rate of 16% annually*
- **Reasons behind this growth**
 - Rising of the Chatbots
 - Urge of discovering the customer insights
 - Transfer of technology of messaging from manual to automated
 - Translation of contents, and
 - many other tasks which are required to be automated and involve language/Speech at some point
 - *Etc.*

Major Industries: *Amazon, Google, Microsoft, Facebook, IBM etc.*

NLP: Evolution

- Evolving from *human-computer interaction* to *human-computer conversation*
- The first critical part of NLP Advancements – Biometrics
- The second critical part of NLP advancements—Humanoid Robotics

NLP: In Governance

- NLP techniques for the delivery to the common people and to decrease the interaction gap between the citizen and the Government
- **Uses of NLP in Government Websites**
 - Making e-governance related information to be available in multiple languages
- **Natural Language Generation in e-Governance**
 - Chatbot
 - E.g. farmer can not read or write, but with the multilingual support and NLP generation, s/he can communicate the query in any language and get it resolved

NLP: In Business, Healthcare

- **Sentiment Analysis:** Analyzing public opinion
- **Email Filters:** Filtering out irrelevant emails
- **Voice Recognition:** Developing smart voice-driven services
- **Information Extraction**
- **NLP in Healthcare**
 - main concern and priority in nowadays the healthcare system is to provide better and 24/7 EHR experience
 - Voice-support systems, Predictive systems, Prescriptive analytics)
- **NLP in Healthcare**
 - can be used to reduce the communication and interaction gap between Healthcare technologies (such as patient portals which contain health records of a patient) and patients
 - Patients can interact in his/her own language
 - Easier for a patient to understand health status

NLP: In Healthcare

- **Increasing the dimension of high quality of care**
 - Healthcare reports generally contain parameters which require proper attention
 - Use of NLP can provide significant relief in case of calculating the measure of inpatient care and monitoring the clinical guidelines
- **Identification of the patients which require Improved Care Coordination**
 - Automated detection of cancer, detection of the root causes related to any substance disorder are some of the examples

NLP: In Finance

- **Credit Scoring Method**
 - Estimate risk factor of giving loan with the past histories
 - E.g. Lenddo EFL (with 115 employees), a Singapore-based company developed a software called Lenddo Score which uses machine learning and NLP to assess and calculate an individual's creditworthiness.
- **Document search**
 - Nuance Communications based in Massachusetts developed software known as Nuance Document Finance Solution, which is used to aid financial services companies in automatizing the documentation process
- **Fraud detection in banking**
- **Stock market prediction-** based on sentiment

NLP: In Other domains

- **National Security**

- Sentiment in Cross-border languages
- Hate Speech, Radicalization

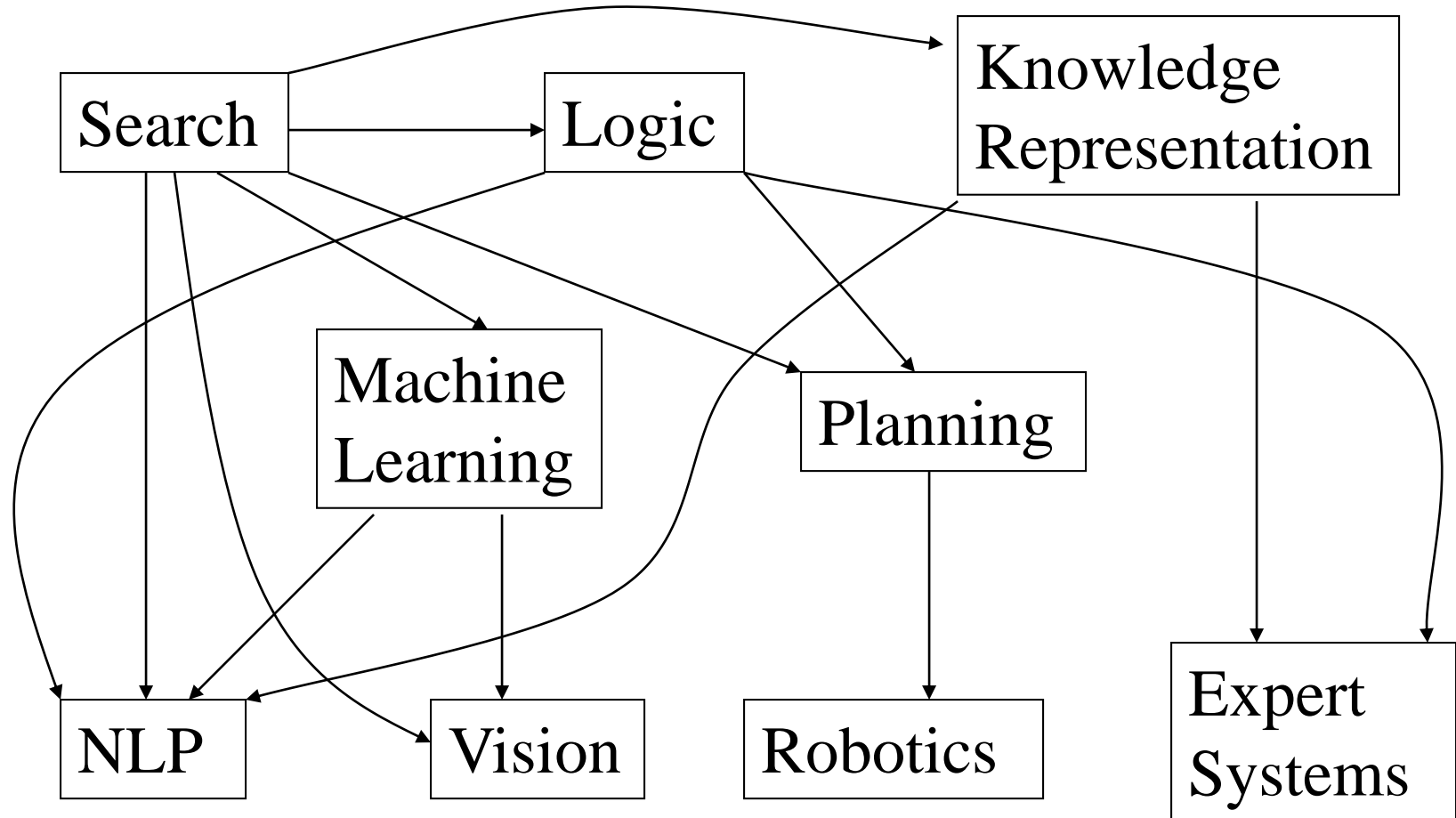
- **NLP in Recruitment**

- searching the appropriate applications from the data, and it also can be used for selecting the best applications from the data available

Natural Language Processing (NLP)

- NLP is the branch of computer science focused on developing systems that allow computers to communicate with people using everyday language
- Related to **Computational Linguistics**
 - Also concerns how computational methods can aid the understanding of human language

Perspectives of NLP: Areas of AI and their inter-dependencies



Evaluation Challenges

- Message Understanding Conference (MUC): Information Extraction (<http://www.cs.nyu.edu/cs/faculty/grishman/muc6.html>)
- Text Retrieval Conference (TREC): Information Retrieval (<http://trec.nist.gov/>)
- Document Understanding Conference (DUC): Summarization (<http://duc.nist.gov/duc2003/call.html>)
- Automatic Content Extraction (ACE): Information Extraction (<http://www.itl.nist.gov/iad/894.01/tests/ace/2004/>)
- Evaluation exercises on Semantic Evaluation (SemEval): WSD, Coreferences etc. (<http://en.wikipedia.org/wiki/SemEval>)
- Cross Language Evaluation Forum (CLEF): Cross-lingual Information retrieval (<http://www.clef-initiative.eu/>)
- Recognising Textual Entailment Challenge (RTE): Textual entailment (<http://www.pascal-network.org/Challenges/RTE/>)

Evaluation Challenges

- Morpho Challenge: unsupervised segmentation of words into morphemes (<http://www.cis.hut.fi/morphochallenge2005/>)
- Web People Search Evaluation Challenges (WePS): Information Extraction (<http://nlp.uned.es/weps/weps-2/>)
- CoNLL challenges: Chunking, Named Entity extraction etc. (<http://www.cnts.ua.ac.be/conll/>)
- Text Analysis Conference (TAC): Entailment etc. (<http://pascallin.ecs.soton.ac.uk/Challenges/RTE/>)
- BioCreative challenges: Biomedical text mining (<http://biocreative.sourceforge.net/>)
- Biomedical information extraction challenges
 - JNLPBA (<http://www-tsuji.is.s.u-tokyo.ac.jp/GENIA/ERtask/report.html>)
 - BioNLP 2009 (<http://www-tsuji.is.s.u-tokyo.ac.jp/GENIA/SharedTask/>)
 - BioNLP 2011 (<http://2011.bionlp-st.org/>)
 - BioNLP 2013 , 2014, 2015, 2016 etc.
- SemEval: Sentiment, Emotion, Question-Answering etc.

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

Definitions etc.

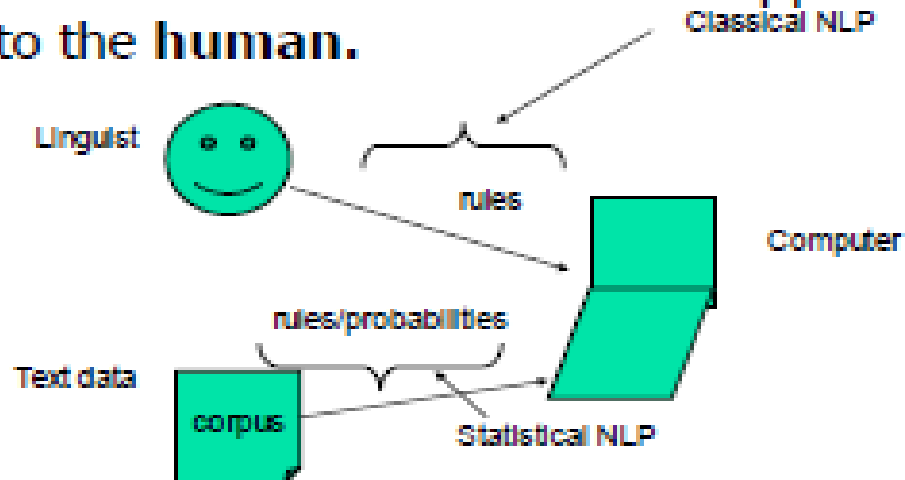
What is NLP?

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

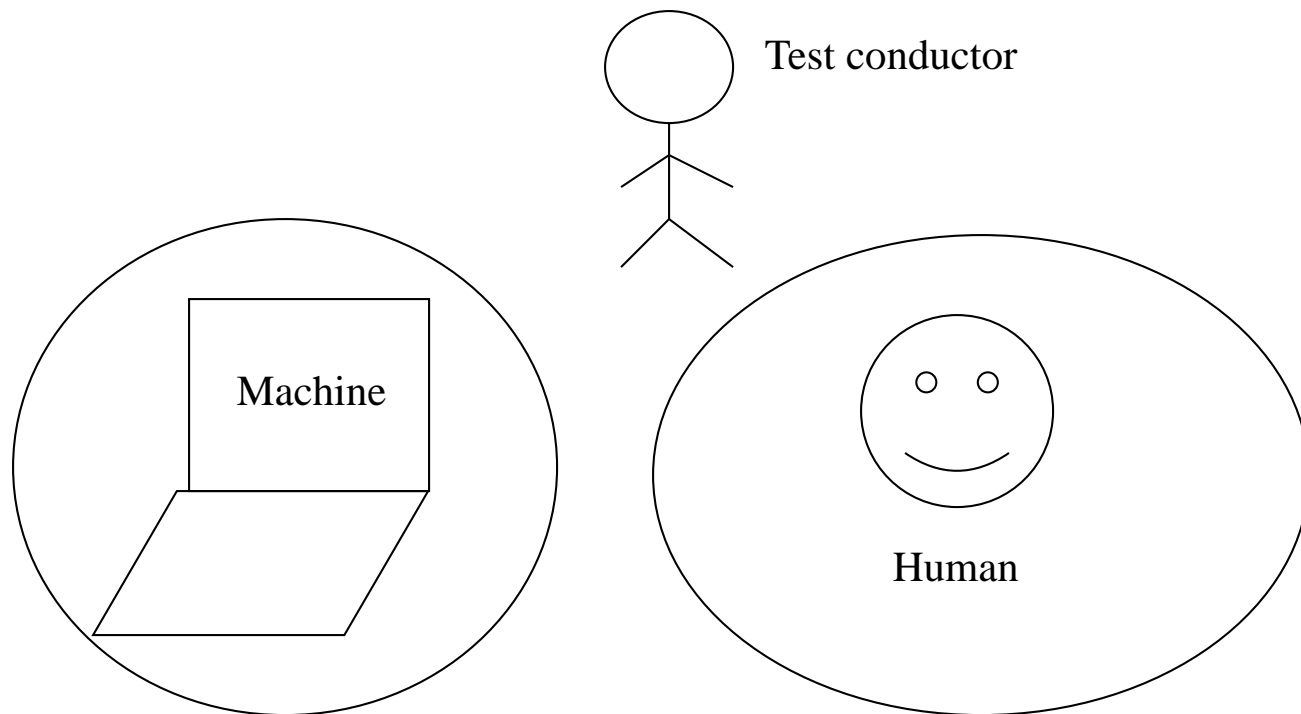
Two Views of NLP

1. Classical View
2. Statistical/Machine Learning View

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



The famous Turing Test: Language based Interaction (Computing Machinery and Intelligence:1950)



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

Natural Languages vs. Computer Languages

- *Ambiguity* is the primary difference between *natural* and *computer languages*
- Formal programming languages are designed to be unambiguous, i.e. they can be defined by a grammar that produces a unique parse (*in general*) for each sentence in the language
- Programming languages are also designed for efficient (deterministic) parsing, i.e. they are deterministic context-free languages (DCFLs)
 - A sentence in a DCFL can be parsed in $O(n)$ time where n is the length of the string

NLP architecture and stages of processing- **ambiguity** at every stage

- 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
 - *PhD students* are especially exhorted to attend as *such seminars* are integral to one's *post-graduate education*
 - Disfluency: *ah, um, ahem etc.*

The best part of my job is ... well ... the best part of my job is the responsibility.

Word Segmentation

- Breaking a string of characters (graphemes) into a sequence of words
- In some written languages (e.g. Chinese) words are not separated by spaces
- Even in English, characters other than white-space can be used to separate words [e.g. , ; . - : ()]
- Examples from English URLs:
 - jumptheshark.com \Rightarrow jump the shark .com
 - myspace.com/pluckerswingbar
 - \Rightarrow myspace .com pluckers wing bar
 - $\otimes \Rightarrow$ myspace .com plucker swing bar

Morphological Analysis

- **Morphology** is the field of linguistics that studies the internal structure of words (Wikipedia)
- A **morpheme** is the smallest linguistic unit that has semantic meaning (Wikipedia)
 - e.g. “carry”, “pre”, “ed”, “ly”, “s”
- Morphological analysis is the task of segmenting a word into its morphemes:
 - carried \Rightarrow carry + ed (past tense)
 - independently \Rightarrow in + (depend + ent) + ly
 - Googlers \Rightarrow (Google + er) + s (plural)
 - unlockable \Rightarrow un + (lock + able) ?
 \Rightarrow (un + lock) + able ?

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*)
- Crucial first step in NLP
- Languages rich in morphology: *e.g.*, Dravidian, Hungarian, Turkish, Indian languages
- 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 or to go after)*

Sense Disambiguation

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

Needs word relationships in a context

- *The chair emphasized 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**)

Watch: wrist watch/ watching something

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
- *Communifaking*: pretending to talk on mobile when you are actually not
- *Discomgooglation*: anxiety / discomfort at not being able to access internet
- *Helicopter Parenting*: over parenting

Ambiguity of Multiwords

- *The grandfather kicked the bucket after suffering from cancer.*
- *This job is a piece of cake*
- *Put the sweater on*
- *He is the dark horse of the match*

Google Translations of above sentences:

दादा कैंसर से पीड़ित होने के बाद बाल्टी लात मारी.
इस काम के केक का एक टुकड़ा है.
स्वेटर पर रखो.
वह मैच के अंधेरे घोड़ा है.

Ambiguity of Named Entities

- Bengali: চঞ্চল সরকার বাড়িতে আছে
English: *Government is restless at home. (*)*
Chanchal Sarkar is at home

Amsterdam airport: “Baby Changing Room”

- Hindi: दैनिक दबंग दुनिया
English: Daily domineering world

Actually name of a Hindi newspaper in Indore

- High degree of overlap between NEs and MWEs
- Treat differently - transliterate do not translate

Syntactic Tasks

Part of Speech (PoS) Tagging

- Annotate each word in a sentence with a PoS

I ate the spaghetti with meatballs.

Pro V Det N Prep N

- Useful for subsequent syntactic parsing and word sense disambiguation

John saw the saw and decided to take it to the table.

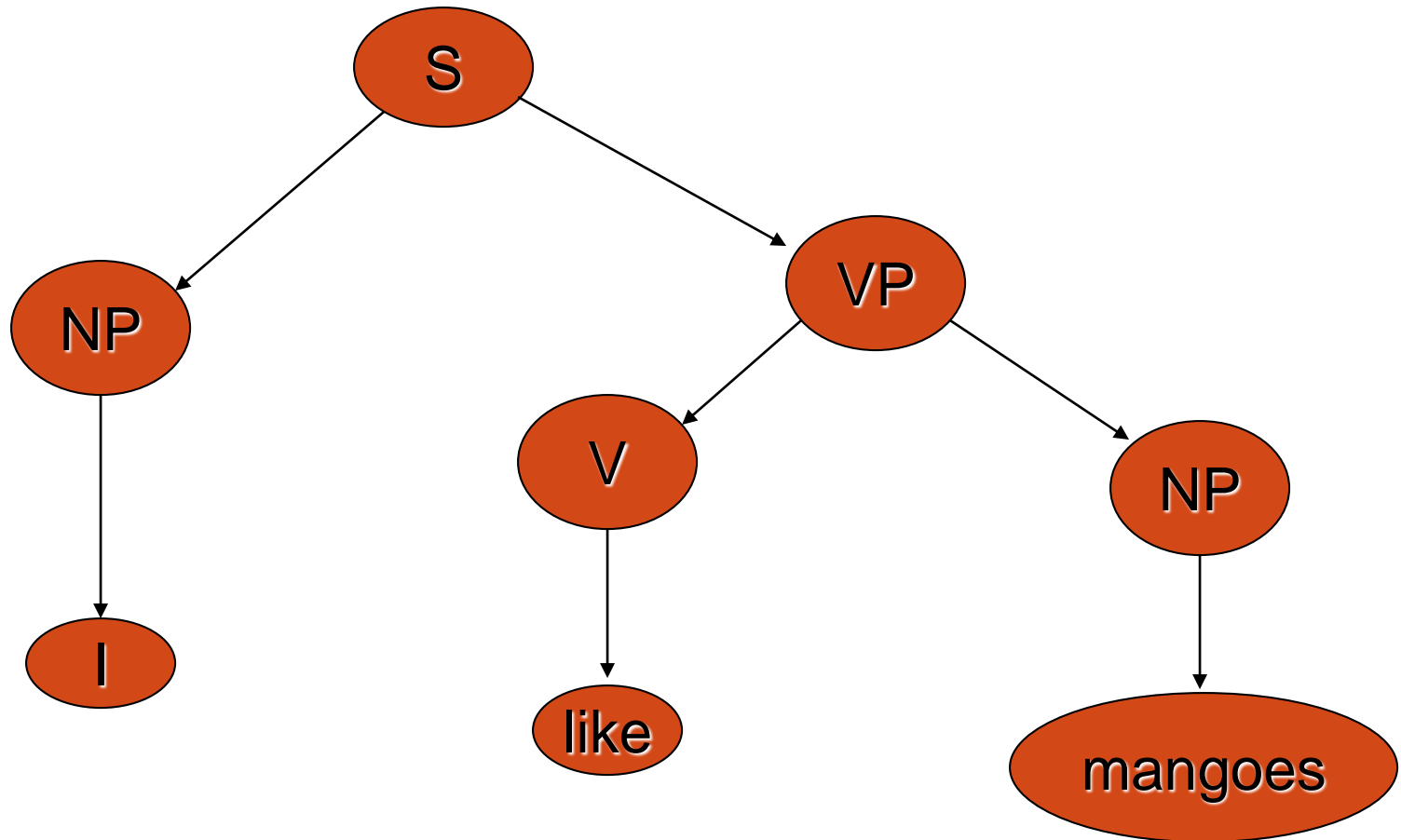
PN V Det N Con V Part V Pro Prep Det N

Phrase Chunking

- Find all non-recursive noun phrases (NPs) and verb phrases (VPs) in a sentence
 - [NP I] [VP ate] [NP the spaghetti] [PP with] [NP meatballs].
 - [NP He] [VP reckons] [NP the current account deficit] [VP will narrow] [PP to] [NP only # 1.8 billion] [PP in] [NP September]

Syntax Processing Stage

Structure Detection



Parsing Strategy

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

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*
- (Times of India, 26/2/08) *Aid for kins of cops killed in terrorist attacks*

Headache for Parsing: Garden Path sentences

- Garden Pathing: A **garden path sentence** is a grammatically correct sentence that starts in such a way that the readers' most likely interpretation will be incorrect
 - *The horse raced past the garden fell* → The horse – (that was) raced past the garden – fell
 - *The old man the boat* → The boat (is manned) by the old
 - *Twin Bomb Strike in Baghdad kill 25* (Times of India 05/09/07) → (*Twin Bomb Strike*) in Baghdad kill 25

Semantic Tasks

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*)
 - *Aapnaake aamake misti khoaate hobe*

Word Sense Disambiguation (WSD)

- Words in natural language usually have a fair number of different possible meanings
 - Ravi has a strong **interest** in computer science
 - Ravi pays a large amount of **interest** on his credit card
- For many tasks (*question answering, translation*), the proper sense of each ambiguous word in a sentence must be determined

Semantic Role Labeling (SRL)

- For each clause, determine the semantic role played by each noun phrase that is an argument to the verb
 - agent patient source destination instrument
 - John drove Mary from Austin to Dallas in his Toyota
 - The hammer broke the window
- Also referred to a “case role analysis,” “thematic analysis,” and “shallow semantic parsing”

Textual Entailment

- Determine whether one natural language sentence entails (implies) another under an ordinary interpretation

Textual Entailment Problems: from PASCAL Challenge

TEXT	HYPOTHESIS	ENTAILMENT
<i>Eyeing the huge market potential, currently led by Google, Yahoo took over search company Overture Services Inc last year.</i>	<i>Yahoo bought Overture.</i>	TRUE
<i>Microsoft's rival Sun Microsystems Inc. bought Star Office last month and plans to boost its development as a Web-based device running over the Net on personal computers and Internet appliances.</i>	<i>Microsoft bought Star Office.</i>	FALSE
<i>The National Institute for Psychobiology in Israel was established in May 1971 as the Israel Center for Psychobiology by Prof. Joel.</i>	<i>Israel was established in May 1971.</i>	FALSE
<i>Since its formation in 1948, Israel fought many wars with neighboring Arab countries.</i>	<i>Israel was established in 1948.</i>	TRUE

Pragmatics/Discourse Tasks

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


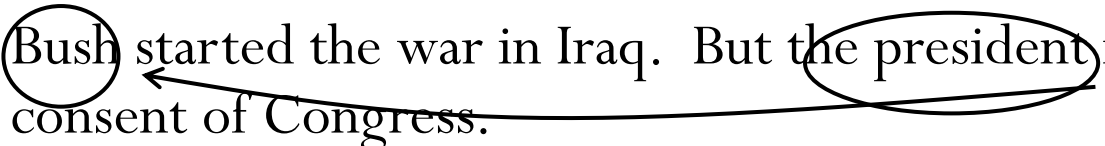
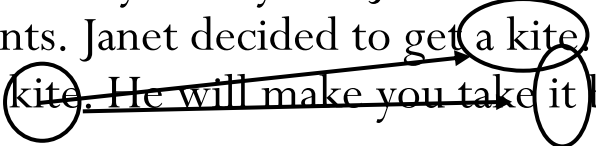
Ambiguity of *father*

father as parent

or

father as headmaster

Anaphora Resolution/ Co-Reference

- Determine which phrases in a document refer to the same underlying entity
 - John put the carrot on the plate and ate it.
 - Bush started the war in Iraq. But the president needed the consent of Congress.
- Some cases require difficult reasoning.
 - Today was Jack's birthday. Penny and Janet went to the store. They were going to get presents. Janet decided to get a kite. "Don't do that," said Penny. "Jack has a kite. He will make you take it back."

Other Tasks

Information Extraction (IE)

- Identify phrases in language that refer to specific types of entities and relations in text
- Named entity recognition is task of identifying names of people, places, organizations, etc. in text

people organizations places

- Michael Dell is the CEO of Dell Computer Corporation and lives in Austin Texas.
- Relation extraction identifies specific relations between entities.

- Michael Dell is the CEO of Dell Computer Corporation and lives in Austin Texas.

Question Answering

- Directly answer natural language questions based on information presented in a corpora of textual documents (e.g. the web)
 - When was Barack Obama born? (*factoid*)
 - August 4, 1961
 - Who was president when Barack Obama was born?
 - John F. Kennedy
 - How many presidents have there been since Barack Obama was born?
 - 9

Text Summarization

- Produce a short summary of a longer document or article
 - **Article:** With a split decision in the final two primaries and a flurry of superdelegate endorsements, [Sen. Barack Obama](#) sealed the Democratic presidential nomination last night after a grueling and history-making campaign against [Sen. Hillary Rodham Clinton](#) that will make him the first African American to head a major-party ticket. Before a chanting and cheering audience in St. Paul, Minn., the first-term senator from Illinois savored what once seemed an unlikely outcome to the Democratic race with a nod to the marathon that was ending and to what will be another hard-fought battle, against [Sen. John McCain](#), the presumptive Republican nominee....
 - **Summary:** Senator Barack Obama was declared the presumptive Democratic presidential nominee.

Sentiment Analysis

- Sentiment analysis
 - Extract subjective information usually from a set of documents, often using online reviews to determine "polarity" about specific objects
 - especially useful for identifying trends of *public opinion in the social media*, for the purpose of marketing

Machine Translation (MT)

- Translate a sentence from one natural language to another.
 - Hasta la vista, bebé \Rightarrow
Until we see each other again, baby.

Ambiguity Resolution is Required for Translation

- Syntactic and semantic ambiguities must be properly resolved for correct translation:
 - “John plays the guitar.” → “John toca la guitarra.”
 - “John plays soccer.” → “John juega el fútbol.”
- An apocryphal story is that an early MT system gave the following results when translating from English to Russian and then back to English:
 - “The spirit is willing but the flesh is weak.” ⇒ “The liquor is good but the meat is spoiled.”
 - “Out of sight, out of mind.” ⇒ “Invisible idiot.”

Resolving Ambiguity

- Choosing the correct interpretation of linguistic utterances requires knowledge of:
 - Syntax
 - An agent is typically the subject of the verb
 - Semantics
 - Michael and Ellen are names of people
 - Austin is the name of a city (and of a person)
 - Toyota is a car company and Prius is a brand of car
 - Pragmatics
 - World knowledge
 - Credit cards require users to pay financial interest
 - Agents must be animate and a hammer is not animate

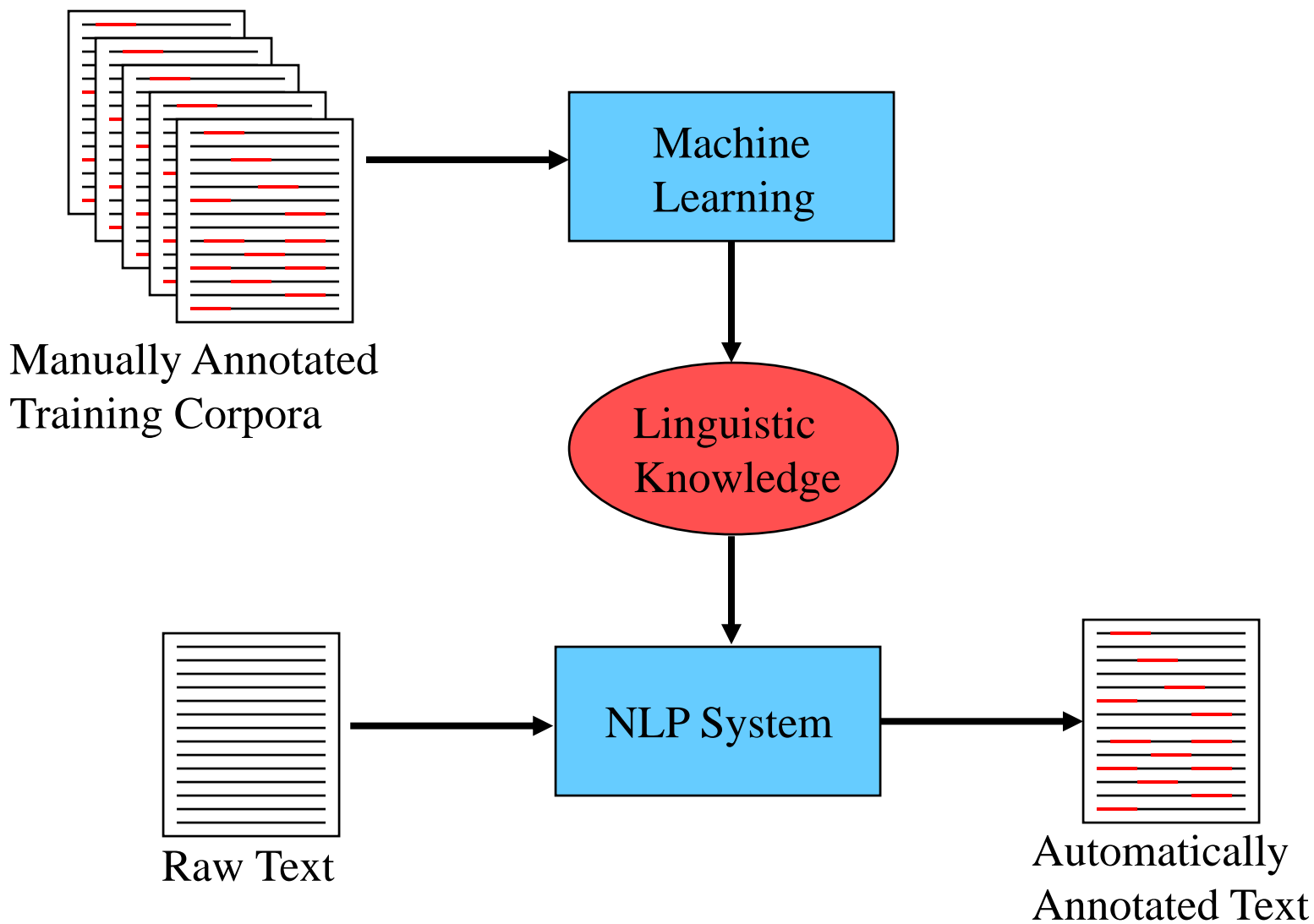
Manual Knowledge Acquisition

- Traditional, “rationalist” approaches to language processing require human specialists to specify and formalize the required knowledge
- Manual knowledge engineering is difficult, time-consuming, and error prone
- “Rules” in language have numerous exceptions and irregularities
 - “All grammars leak.” : Edward Sapir (1921)
- Manually developed systems were expensive to develop and their abilities were limited and “brittle” (not robust)

Automatic Learning Approach

- Use machine learning methods to automatically acquire the required knowledge from appropriately annotated text corpora
- Various referred to as the “corpus based,” “statistical,” or “empirical” approach
- Statistical learning methods were first applied to speech recognition in the late 1970’s and became the dominant approach in the 1980’s
- During the 1990’s, the statistical training approach expanded and came to dominate almost all areas of NLP

Learning Approach



Early History: 1950' s

- Shannon (the father of information theory) explored probabilistic models of natural language (1951)
- Chomsky (the extremely influential linguist) developed formal models of syntax, i.e. finite state and context-free grammars (1956)
- First computational parser developed at U Penn as a cascade of finite-state transducers (Joshi, 1961; Harris, 1962)
- Bayesian methods developed for *optical character recognition* (OCR) (Bledsoe & Browning, 1959).

History: 1960' s

- Work at MIT AI lab on question answering (BASEBALL) and dialog (ELIZA)
- Semantic network models of language for question answering (Simmons, 1965).
- First electronic corpus collected, Brown corpus, 1 million words (Kucera and Francis, 1967)
- Bayesian methods used to identify document authorship (*The Federalist* papers) (Mosteller & Wallace, 1964)

History: 1970' s

- “Natural language understanding” systems developed that tried to support deeper semantic interpretation
 - SHRDLU (Winograd, 1972) performs tasks in the “blocks world” based on NL instruction
 - Schank *et al.* (1972, 1977) developed systems for conceptual representation of language and for understanding short stories using hand-coded knowledge of scripts, plans, and goals.
- Prolog programming language developed to support logic-based parsing (Colmerauer, 1975).
- Initial development of hidden Markov models (HMMs) for statistical speech recognition (Baker, 1975; Jelinek, 1976).

History: 1980' s

- Development of more complex (mildly context sensitive) grammatical formalisms, e.g. unification grammar, tree-adjoining grammar etc
- Symbolic work on discourse processing and NL generation.
- Initial use of statistical (HMM) methods for syntactic analysis (POS tagging) (Church, 1988).

History: 1990' s

- Rise of statistical methods and empirical evaluation causes a “scientific revolution” in the field
- Initial annotated corpora developed for training and testing systems for POS tagging, parsing, WSD, information extraction, MT, etc.
- First statistical machine translation systems developed at IBM for Canadian Hansards corpus (Brown *et al.*, 1990)
- First robust statistical parsers developed (Magerman, 1995; Collins, 1996; Charniak, 1997)
- First systems for robust information extraction developed (e.g. MUC competitions)

History: 2000' s

- Increased use of a variety of ML methods, SVMs, logistic regression (i.e. max-ent), CRF' s, etc.
- Continued developed of corpora and competitions on shared data.
 - TREC Q/A
 - SENSEVAL/SEMEVAL
 - CONLL Shared Tasks (NER, SRL...)
- Increased emphasis on unsupervised, semi-supervised, and active learning as alternatives to purely supervised learning.
- Shifted focus to semantic tasks such as WSD and SRL.

History: 2000 onwards

- Information extraction from social networks
- Information retrieval
- Cross-lingual information access
- Machine Translation (statistical, hybrid etc.)
- Biomedical text mining
- Discourse processing

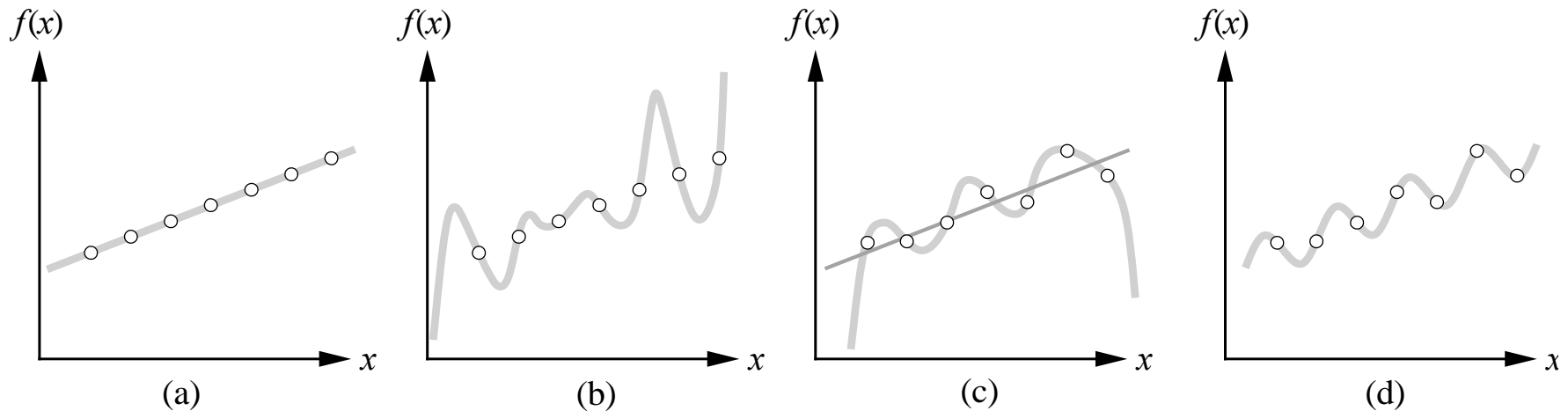
Machine Learning

- **Machine learning**: how to acquire a model on the basis of data / experience?
 - Learning parameters (e.g. probabilities)
 - Learning structure (e.g. BN graphs)
 - Learning hidden concepts (e.g. clustering)

Machine Learning

- **Unsupervised Learning**
 - No feedback from teacher; detect patterns
- **Reinforcement Learning**
 - Feedback consists of rewards/punishment
- **Supervised Learning**
 - Examples of correct answers are given
 - Discrete answers: *Classification*
 - Continuous answers: *Regression*

Supervised Machine Learning



Given a training set:

$$(x_1, y_1), (x_2, y_2), (x_3, y_3), \dots (x_n, y_n)$$

Where each y_i was generated by an unknown $y = f(x)$,

Discover a function h that approximates the true function f

Example: Spam Filter

- Input: $x = \text{email}$
- Output: $y = \text{“spam” or “ham”}$
- **Setup:**
 - Get a large collection of example emails, each labeled “spam” or “ham”
 - Note: someone has to hand label all this data!
 - Want to learn to predict labels of new, future emails
- **Features:** The attributes used to make the ham / spam decision
 - Words: FREE!
 - Text Patterns: \$dd, CAPS
 - Non-text: SenderInContacts
 - ...



Dear Sir.

First, I must solicit your confidence in this transaction, this is by virture of its nature as being utterly confidential and top secret. ...



TO BE REMOVED FROM FUTURE MAILINGS, SIMPLY REPLY TO THIS MESSAGE AND PUT "REMOVE" IN THE SUBJECT.

99 MILLION EMAIL ADDRESSES FOR ONLY \$99



Ok, I know this is blatantly OT but I'm beginning to go insane. Had an old Dell Dimension XPS sitting in the corner and decided to put it to use, I know it was working pre being stuck in the corner, but when I plugged it in, hit the power nothing happened.

Example: Digit Recognition

- Input: x = images (pixel grids)
- Output: y = a digit 0-9
- Setup:
 - Get a large collection of example images, each labeled with a digit
 - Note: someone has to hand label all this data!
 - Want to learn to predict labels of new, future digit images
- Features: The attributes used to make the digit decision
 - Pixels: (6,8)=ON
 - Shape Patterns: NumComponents, AspectRatio, NumLoops
 - ...

 0

 1

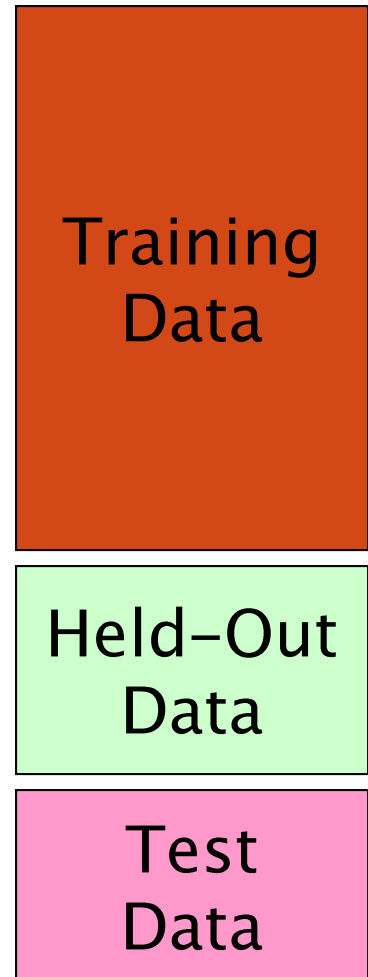
 2

 1

 ??

How to Learn

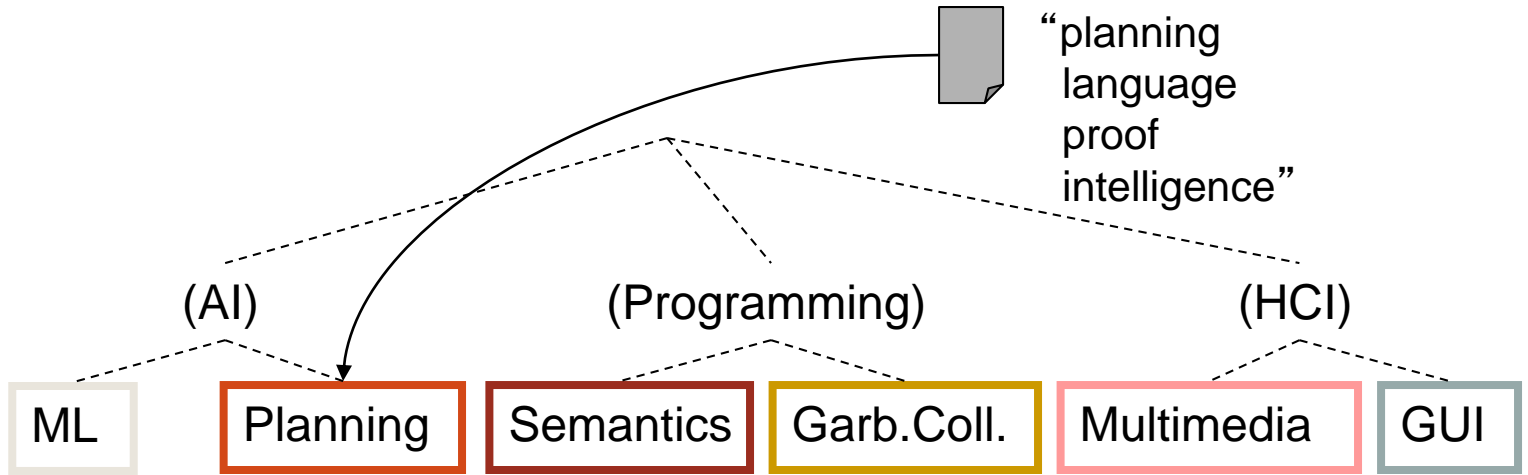
- **Data:** labeled instances, e.g. emails marked spam/ham
 - Training set
 - Held out (validation) set
 - Test set
- **Features:** attribute-value pairs which characterize each x
- **Experimentation cycle**
 - Learn parameters (e.g. model probabilities) on training set
 - Tune hyperparameters on held-out set
 - Compute accuracy on test set
 - Very important: never “peek” at the test set!
- **Evaluation**
 - Accuracy: fraction of instances predicted correctly
- **Overfitting and generalization**
 - Want a classifier which does well on *test* data
 - Overfitting: fitting the training data very closely, but not generalizing well to test data



Document Classification

Test Data:

Classes:



Training Data:

learning intelligence algorithm Reinforcement network...	<u>planning</u> temporal reasoning plan <u>language...</u>	programming semantics <u>language</u> <u>proof...</u>	garbage collection memory optimization region...
--	--	---	--	-----	-----

More Text Classification Examples

Many search engine functionalities use classification

Assigning labels to documents or web-pages:

- Labels are most often topics such as Yahoo-categories
 - *"finance," "sports," "news>world>asia>business"*
- Labels may be genres (or, categories)
 - *"editorials" "movie-reviews" "news"*
- Labels may be opinion on a person/product
 - *"like", "hate", "neutral"*
- Labels may be domain-specific
 - *"interesting-to-me" : "not-interesting-to-me"*
 - *language identification: English, French, Chinese, ...*
 - *search vertical: about Linux versus not*
 - *"link spam" : "not link spam"*

Classification Methods: History

- **Manual classification**
 - Used by the original Yahoo! Directory
 - Looksmart, about.com, ODP, PubMed
 - Very accurate when job is done by experts
 - Consistent when the problem size and team is small
 - Difficult and expensive to scale
 - Means we need automatic classification methods for big problems

Classification Methods: History

- **Automatic classification**
 - Hand-coded rule-based systems
 - One technique used by Reuters, CIA, etc.
 - It's what Google Alerts is doing
 - Widely deployed in government and enterprise
 - Companies provide “IDE” (integrated development environment) for writing such rules
 - E.g., assign category if document contains a given boolean combination of words
 - Standing queries: Commercial systems have complex query languages (everything in IR query languages +score accumulators)
 - Accuracy is often very high if a rule has been carefully refined over time by a subject expert
 - Building and maintaining these rules is expensive
 - Rules could vary with the change of domain

Classification Methods: History

- *Supervised learning of a document-label assignment function*
 - Many systems *partly rely on machine learning* (Autonomy, Microsoft, Enkata, Yahoo!, Google News, ...)
 - k-Nearest Neighbors (simple, powerful)
 - Naive Bayes (simple, common method)
 - Support-vector machines (new, more powerful)
 - ... plus many other methods
 - Requirement: requires hand-classified training data
 - But data can be built up (and refined) by amateurs
- Many commercial systems use a mixture of methods

NLP and ML: From Past to Present

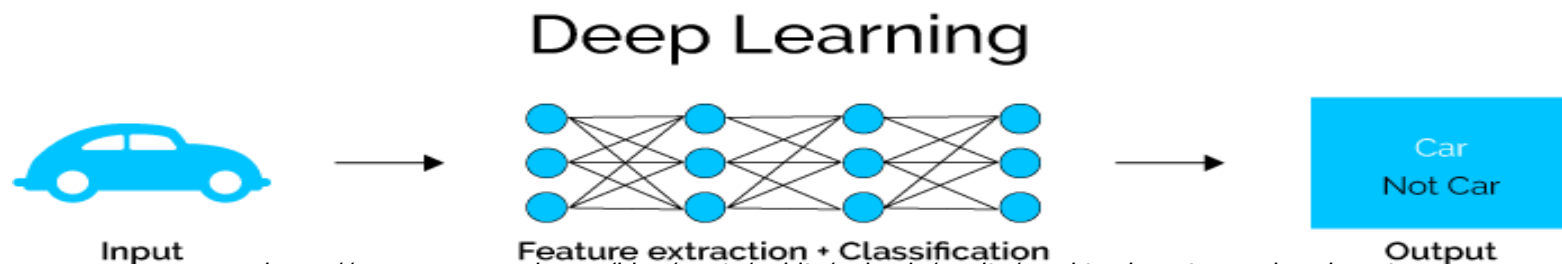
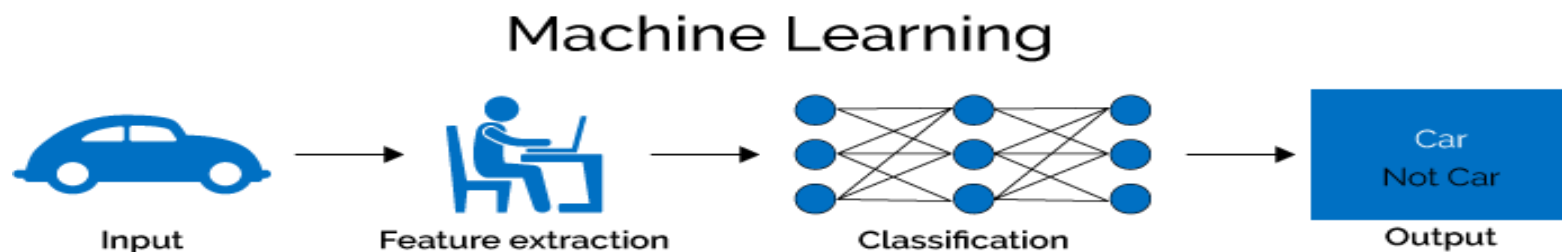
- NLP based systems have enabled wide-range of applications
 - Google's powerful search engines, Google's MT
 - Alexa etc.
 - Amazon Comprehend Medical services
 - Cognitive Analytics and NLP, Spam detection, NLP in Recruitment
 - Sentiment Analysis, Hate Speech detection, Fake News detection
- Shallow ML algorithms (corresponds to Statistical NLP)
 - Used extensively (HMM, MaxEnt, CRF, SVM, Logistic Regression etc.)
 - Requires handcrafting of features
 - Time-consuming
 - Curse of dimensionality (because of joint modeling of language models)

NLP and ML: From Past to Present

- Deep Learning algorithms
 - No feature engineering
 - Success of distributed representations (Neural language models)
- Some recent developments
 - The rise of distributed representations (e.g., Word2vec, GLOVE, ELMO, BERT etc)
 - Convolutional, recurrent, recursive neural networks, Transformer, Reinforcement learning
 - Unsupervised sentence representation learning
 - Combining deep learning models with memory-augmenting strategies
- Explainable AI

Deep Learning (DL)

- Subfield of learning **representations** of data
- Exceptionally effective at **learning patterns**
- Deep learning algorithms attempt to learn (*multiple levels of*) representations by using a **hierarchy of multiple layers**
- If you provide the system **tons of information**, it begins to understand it and respond in useful ways

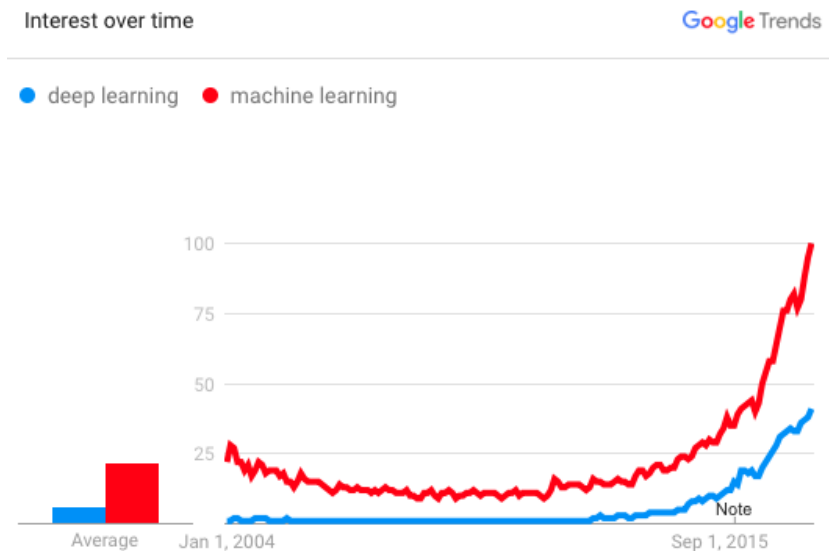


<https://www.xenonstack.com/blog/static/public/uploads/media/machine-learning-vs-deep-learning.png>

Why is DL useful?

- Manually designed features are often **over-specified**, **incomplete** and take a **long time to design** and validate
- Learned Features are **easy to adapt**, **fast** to learn
- Deep learning provides a very **flexible**, (almost?) **universal**, learnable framework for representing world, visual and linguistic information
- Can learn both unsupervised and supervised
- Effective **end-to-end** learning
- Utilize large amounts of training data

In ~2010 DL started
outperforming other ML
techniques
first in speech and vision, then NLP



News: March 27, 2019

Yoshua Bengio, Geoffrey Hinton, and Yann LeCun received the

Turing Award-2018 (equivalent to Nobel Prize of Computing)

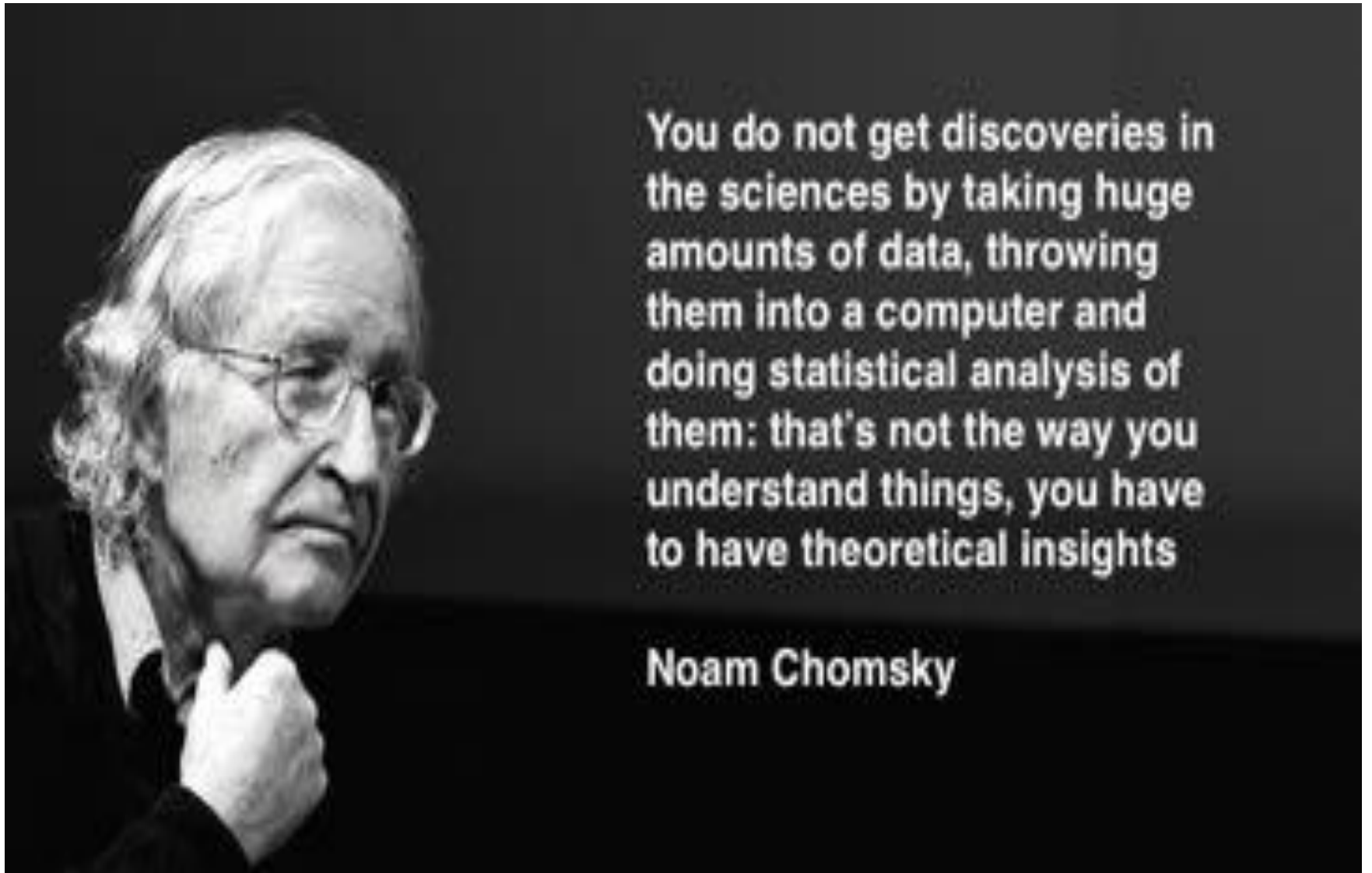
- *for Modern AI (specifically for deep learning research)*

Bengio- University of Toronto and Google

Hinton- University of Montreal

LeCun- Facebook's chief AI scientist and a professor at NYU

Statistics are no panacea!



You do not get discoveries in the sciences by taking huge amounts of data, throwing them into a computer and doing statistical analysis of them: that's not the way you understand things, you have to have theoretical insights

Noam Chomsky

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