CS626 : Natural Language Processing, Speech and the Web

(Lecture 1,2,3 – Introduction, POS tagging)

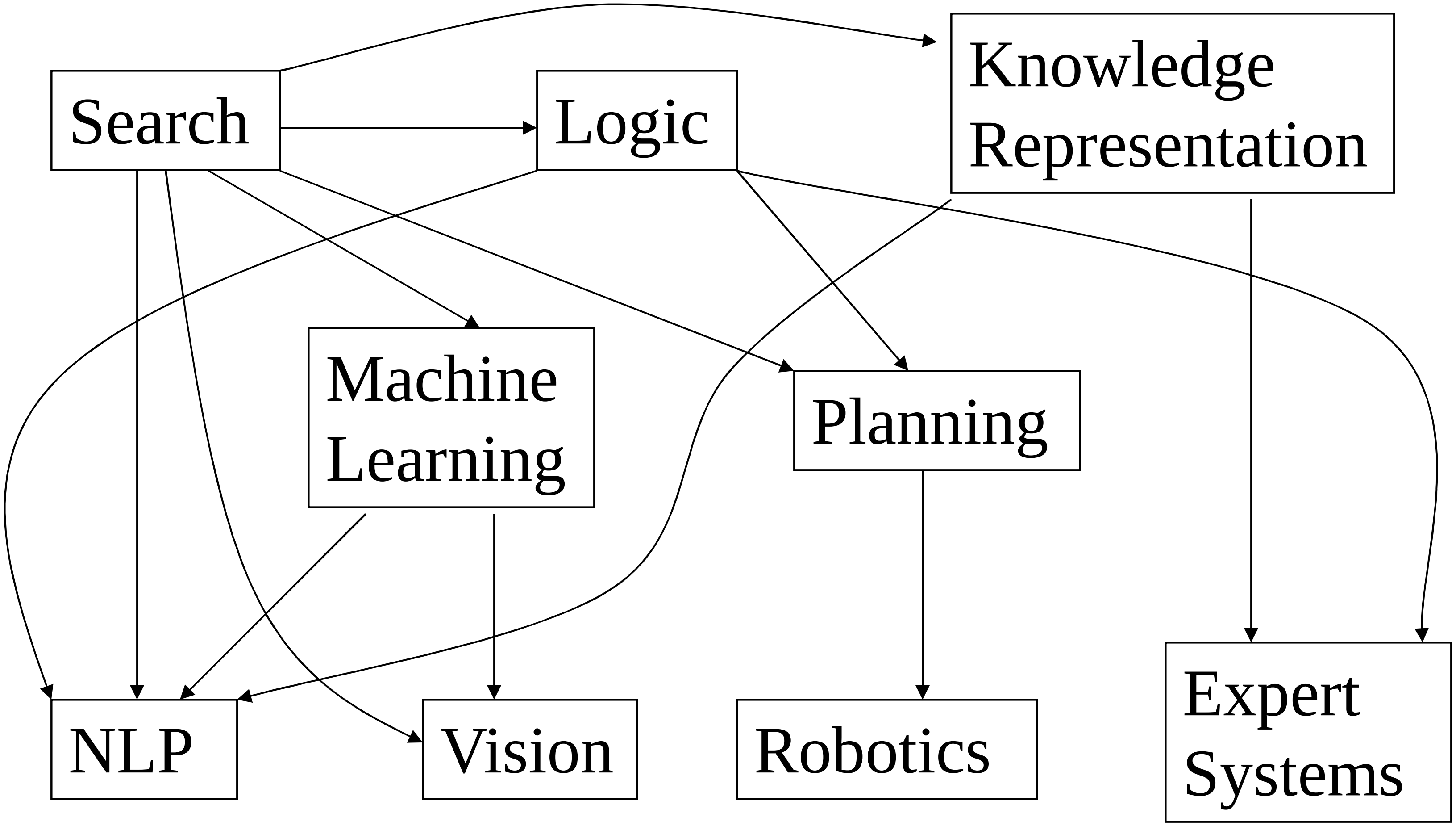
Pushpak Bhattacharyya
CSE Dept.,
IIT Bombay

19th , 22nd and 26th July, 2012

Logistics

- Faculty instructors: Dr. Pushpak Bhattacharyya (www.cse.iitb.ac.in/~pb)
- TAs: Avishek Dan, Bibek Behera {avishekdan,bibek}@cse.iitb.ac.in
- Course home page (to be created)
  - http://www.cse.iitb.ac.in/~cs626-sem1-2012
- Moodle account
- SIC 201
- Slot 8: Mon-2 to 3.25 PM and Thu-2 to 3.25 PM

Perpectivising NLP: Areas of AI and their inter-dependencies



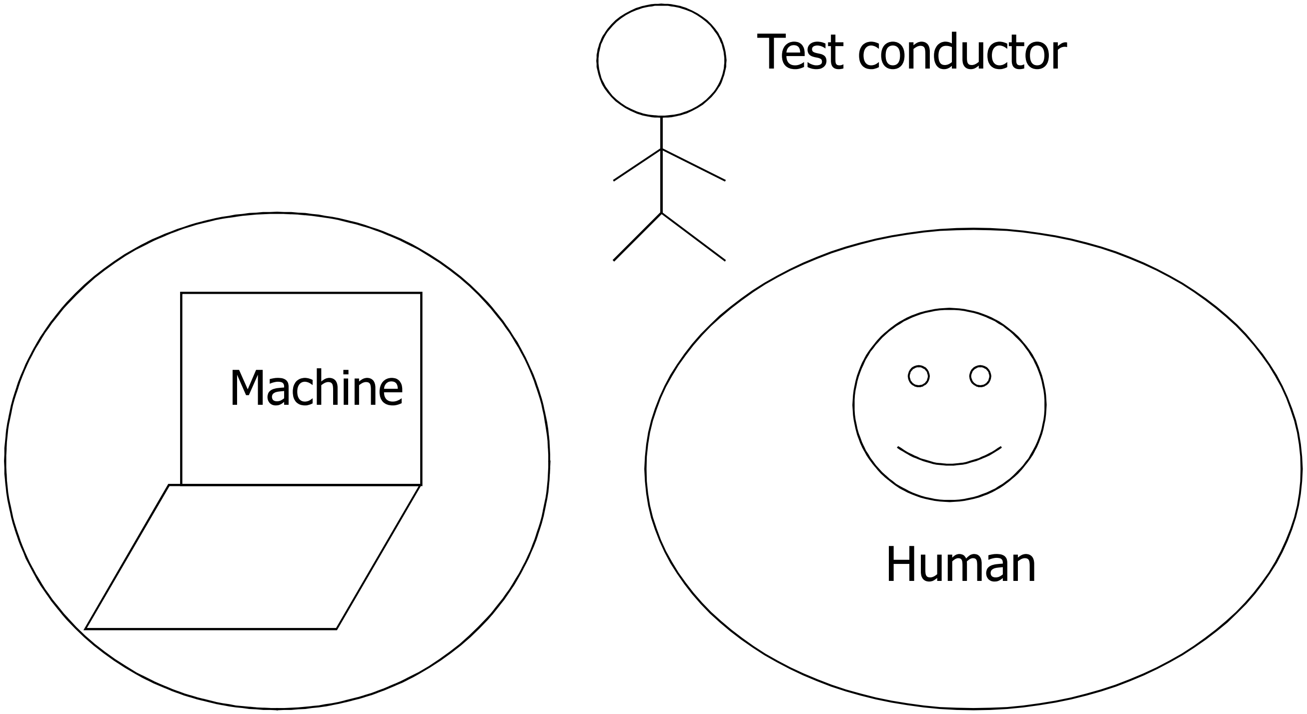
What is NLP

- 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 and the Associated Challenges

1. Classical View
2. Statistical/Machine Learning View

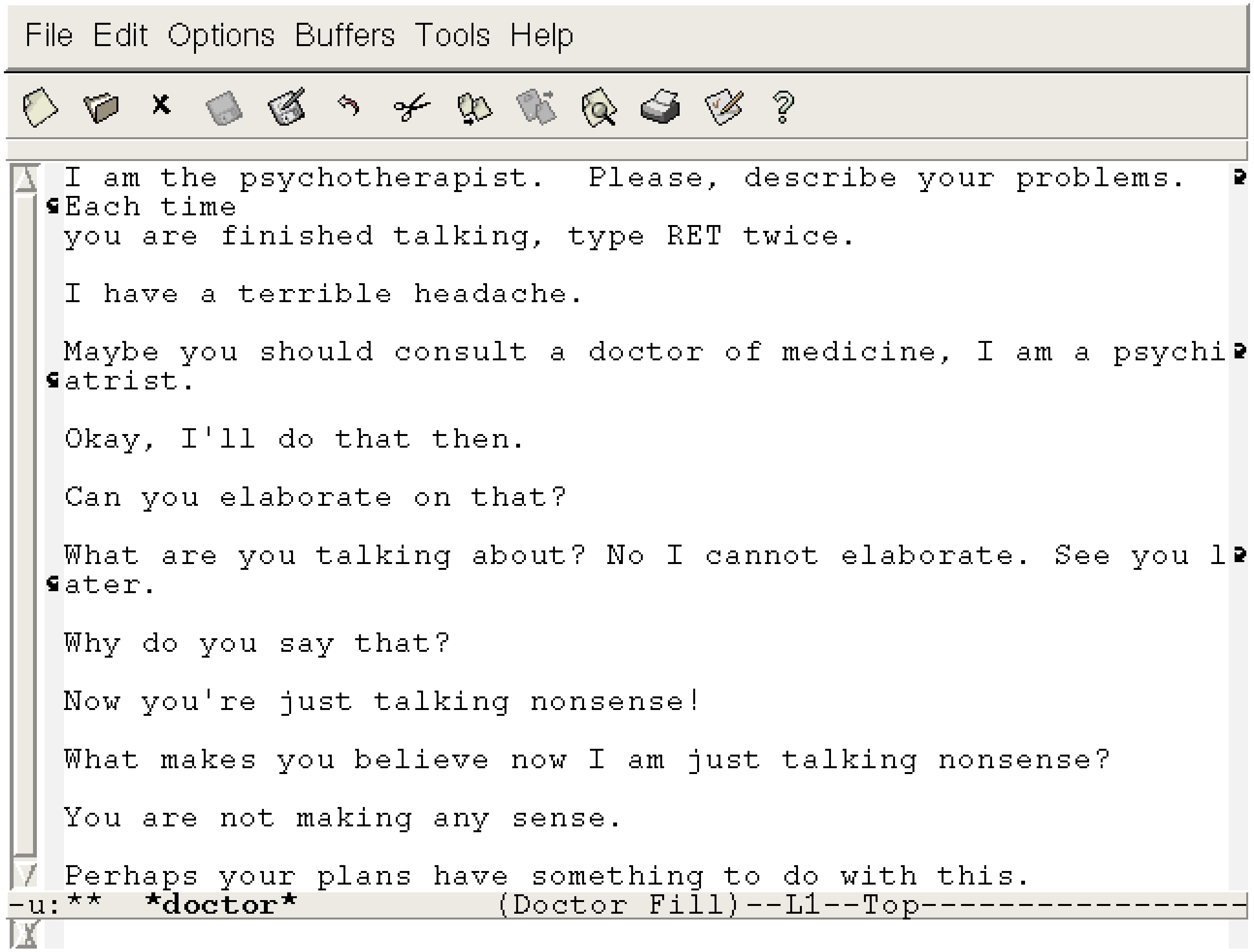
The famous Turing Test: Language Based Interaction



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

Inspired *Eliza*

- A Sample of Interaction:

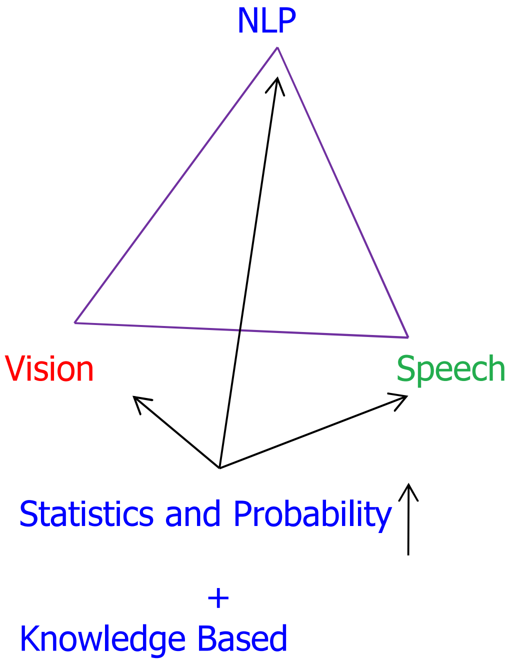


```
I am the psychotherapist.  Please, describe your problems.
Each time
you are finished talking, type RET twice.

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-----------------
```

NLP→Cryptography→Turing

- Machine Translation is an important branch of NLP, has once been looked upon as a problem in cryptography!
- Warren Weaver, one of the pioneering minds in machine translation, wrote in 1947:
  - *When I look at an article in Russian, I say: 'This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode'. [Weaver, 1947, 1949]*

NLP: Two pictures



NLP

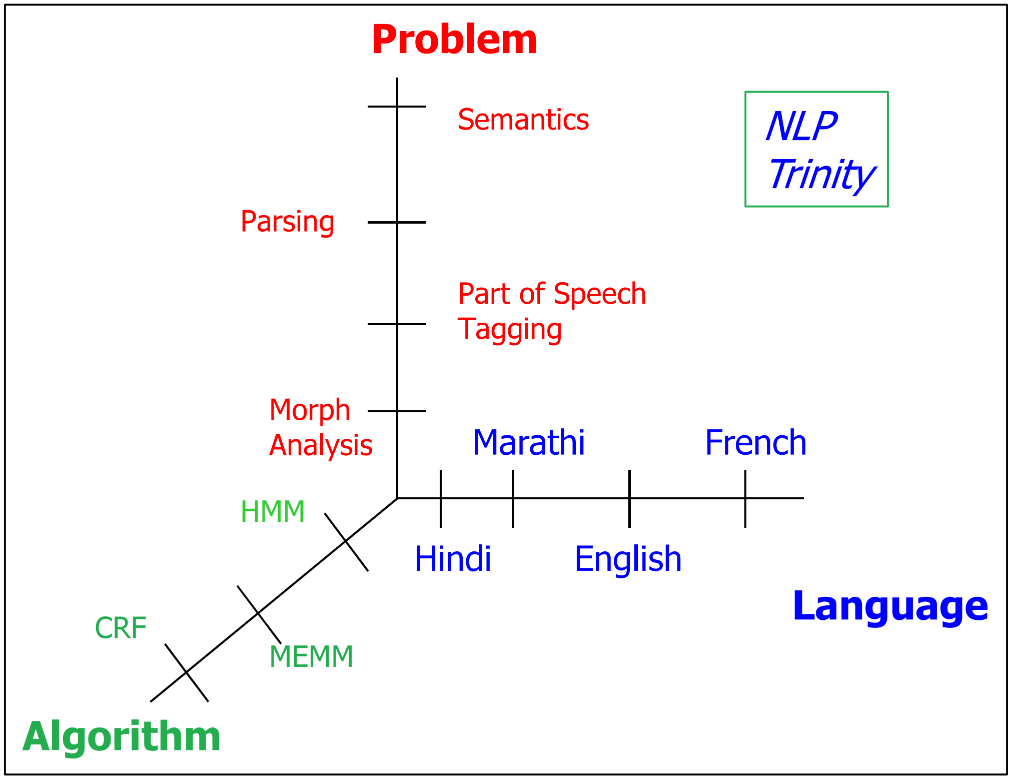
Vision

Speech

Statistics and Probability

+

Knowledge Based



Problem

Semantics

Parsing

Part of Speech Tagging

Morph Analysis

NLP Trinity

Marathi

French

Hindi

English

Language

HMM

CRF

MEMM

Algorithm

Stages of processing

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

sound and associated challenges

- Homophones: *bank (finance)* vs. *bank (river bank)*
- Near Homophones: *maatraa* vs. *maatra (hin)*
- Word Boundary
    - आजायेंगे (*aajaayenge) (aa jaayenge (will come)* or *aaj aayenge (will come today)*
    - *I got [ua]plate*
  - Disfluency: *ah, um, ahem etc.*

Recently heard:

The king of Abu Dhabi had expired last year. The nation was mourning. A few children were playing in the evening in a garden.

An elderly person: "Why are you playing? It is mourning time.

Children: "No, it is evening time. Why shouldn't we play."

Today's (24/6/12) Times of India Headline: Google CEO Larry Page looses his voice.

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
- *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

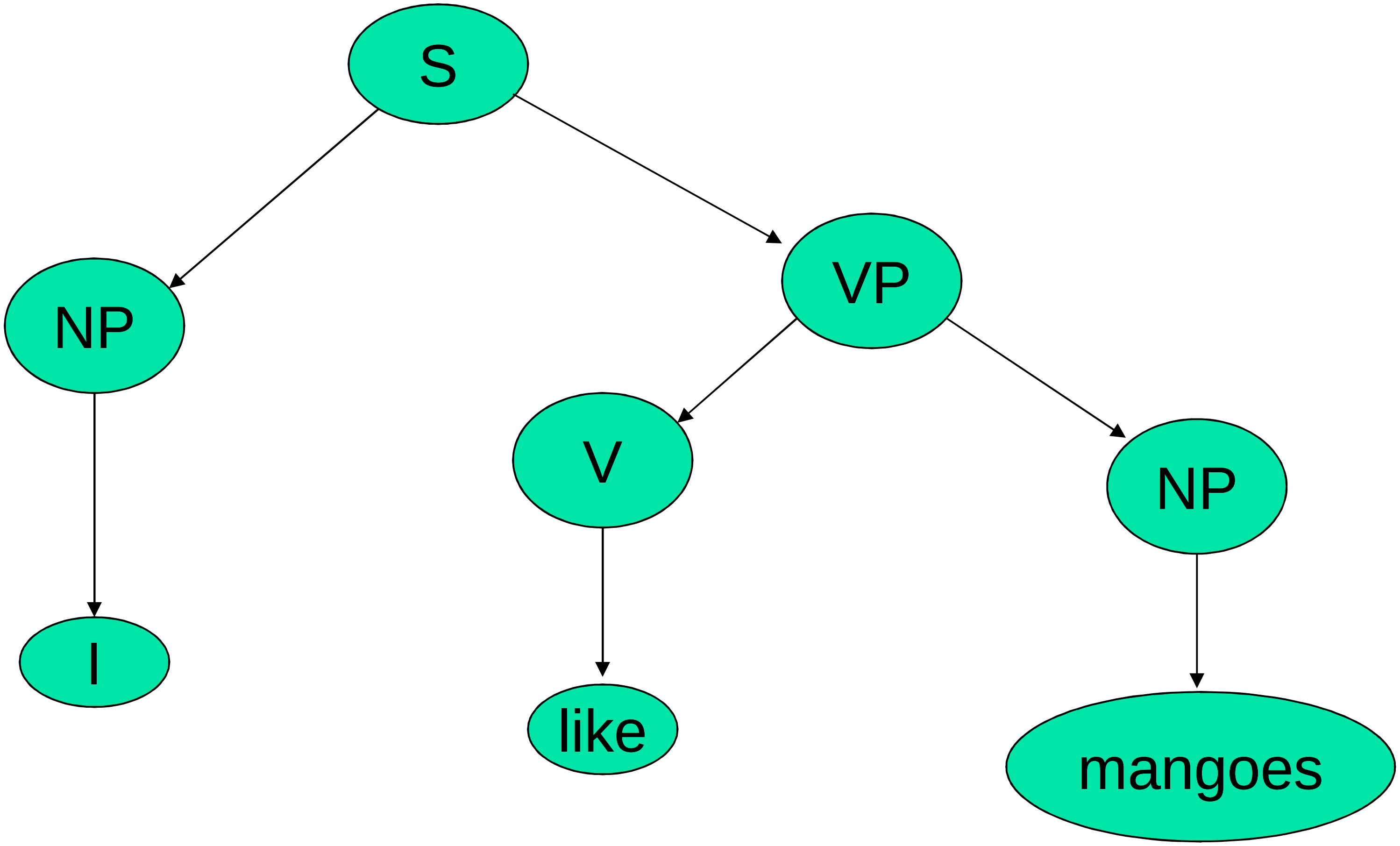
- Bengali: চঞ্চল সরকার বাড়িতে আছে
  English: *Government is restless at home*. (*)
  Chanchal Sarkar is at home

- Hindi: दैनिक दबंग दुनिया
  English: everyday bold world
  Actually name of a Hindi newspaper in Indore

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

Syntax Processing Stage

Structure Detection



Parsing Strategy

- Driven by grammar
  - S-> NP VP
  - NP-> N | PRON
  - VP-> V NP | V PP
  - N-> Mangoes
  - PRON-> I
  - V-> 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"*

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

Textual Humour (1/2)

1. Teacher (angrily): did you miss the class yesterday?
   Student: not much
2. A man coming back to his parked car sees the sticker "Parking fine". He goes and thanks the policeman for appreciating his parking skill.
3. *Son*: mother, I broke the neighbour's lamp shade.
   Mother: then we have to give them a new one.
   Son: no need, aunty said the lamp shade is irreplaceable.
4. *Ram*: I got a Jaguar car for my unemployed youngest son.
   Shyam: That's a great exchange!
5. Shane Warne should bowl maiden overs, instead of bowling maidens over

Textual Humour (2/2)

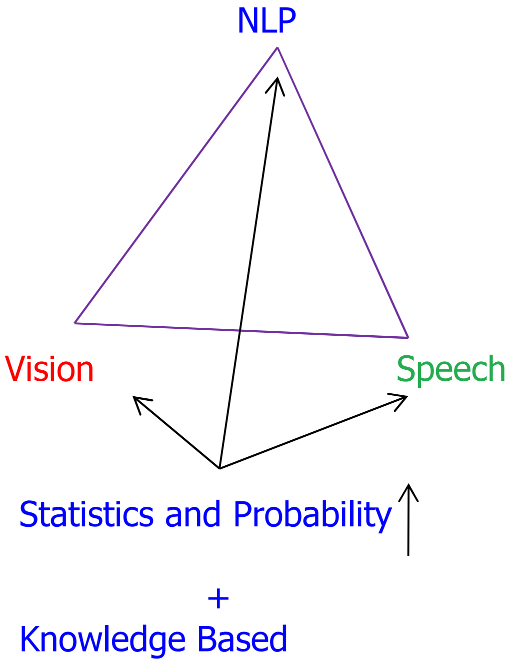
- It is not hard to meet the expenses now a day, you find them everywhere

- Teacher: What do you think is the capital of Ethiopia?
  Student: What do you think?
  Teacher: I do not think, I know
  Student: I do not think I know

Stages of processing

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

Two pictures



NLP

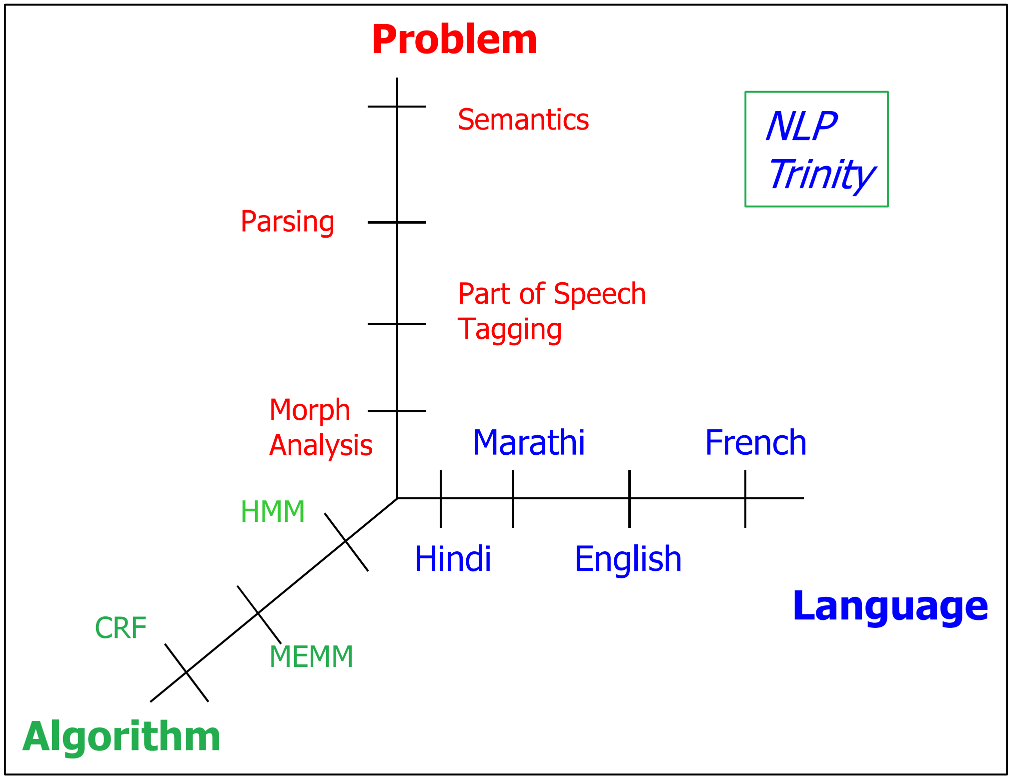
Vision

Speech

Statistics and Probability

+

Knowledge Based



Problem

Semantics

Parsing

Part of Speech Tagging

Morph Analysis

NLP Trinity

Marathi

French

HMM

Hindi

English

Language

CRF

MEMM

Algorithm

Part of Speech Tagging and Hidden Markov Model

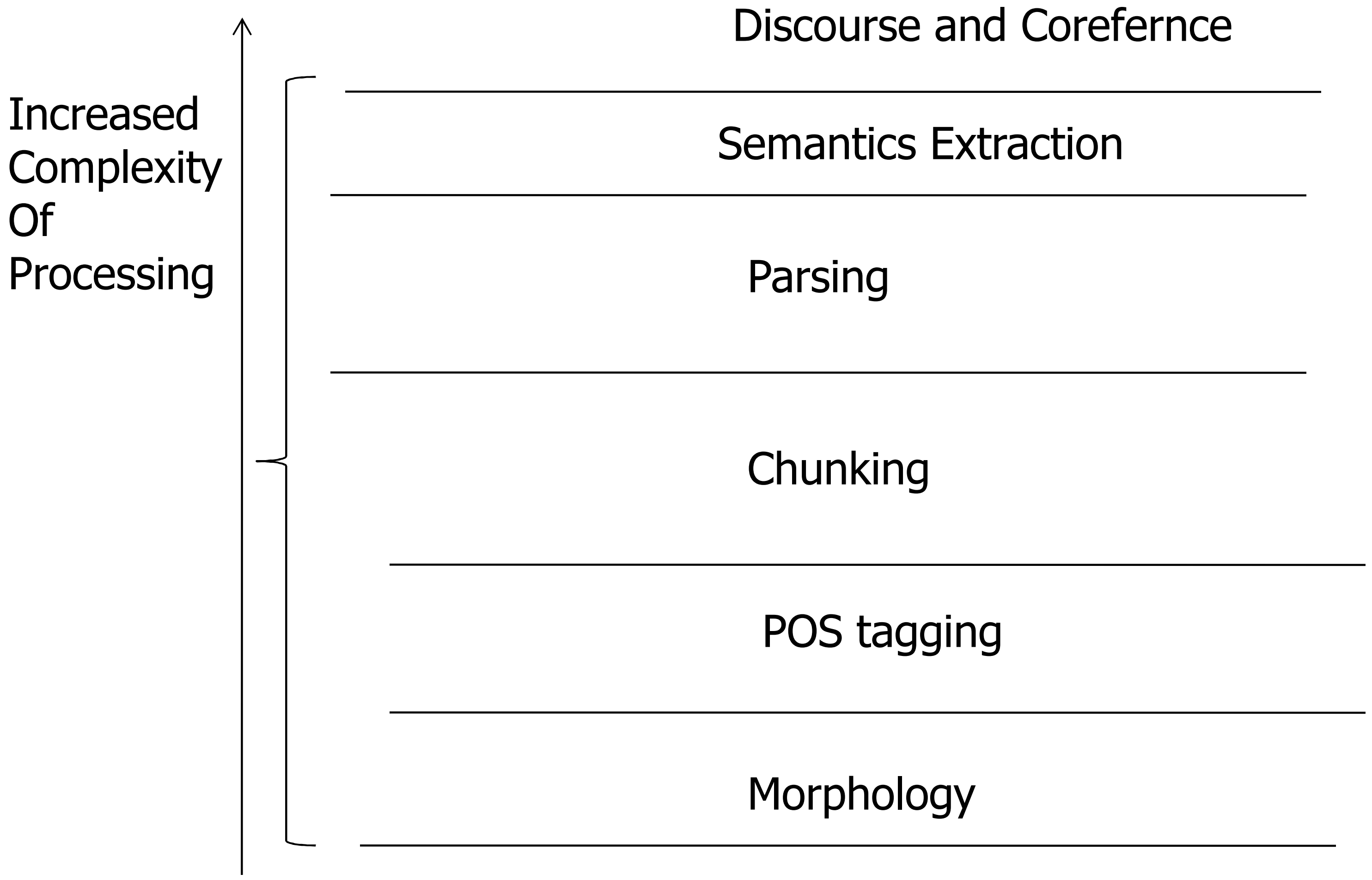
Part of Speech Tagging

- POS Tagging is a process that attaches each word in a sentence with a suitable tag from a given set of tags.
- The set of tags is called the Tag-set.
- Standard Tag-set : Penn Treebank (for English).

POS tagging: Definition

- Tagging is the assignment of a singlepart-of-speech tag to each word (and punctuation marker) in a corpus.
  - “_“ The_DT guys_NNS that_WDT make_VBP traditional_JJ hardware_NN are_VBP really_RB being_VBG obsoleted_VBN by_IN microprocessor-based_JJ machines_NNS ,_, ”_” said_VBD Mr._NNP Benton_NNP ._.

Where does POS tagging fit in



Increased Complexity Of Processing

Discourse and Corefernce

Semantics Extraction

Parsing

Chunking

POS tagging

Morphology

Mathematics of POS tagging

Entity Labeling

- Label a sequence of entities with labels from a set

$\mathbf{e_1}$	**e2**	**e3**	**Entities**
L_1	L_2	L_3	Labels
			Sample Face Images
Sad (S)	Peaceful (P)	Angry (A)	Visual Sentiment Analysis
The movie was great	The movie was horrible	Shahrukh is the hero	Sample text
Positive (P)	Negative (N)	Objective (O)	Sentiment Analysis

Argmax computation (1/2)

Best tag sequence
= T*
= argmax P(T|W)
= argmax P(T)P(W|T)   (by Baye's Theorem)

$P(T) = P(t_0{=}\wedge\ t_1 t_2 \ldots t_{n+1}{=}.)$

$= P(t_0)P(t_1|t_0)P(t_2|t_1t_0)P(t_3|t_2t_1t_0) \ldots$

$P(t_n|t_{n-1}t_{n-2}\ldots t_0)P(t_{n+1}|t_nt_{n-1}\ldots t_0)$

$= P(t_0)P(t_1|t_0)P(t_2|t_1) \ldots P(t_n|t_{n-1})P(t_{n+1}|t_n)$

$= \prod_{i=0}^{N+1} P(t_i|t_{i-1})$ Bigram Assumption

Argmax computation (2/2)

$$P(W|T) = P(w_0|t_0\text{-}t_{n+1})P(w_1|w_0t_0\text{-}t_{n+1})P(w_2|w_1w_0t_0\text{-}t_{n+1}) \ldots$$
$$P(w_n|w_0\text{-}w_{n-1}t_0\text{-}t_{n+1})P(w_{n+1}|w_0\text{-}w_nt_0\text{-}t_{n+1})$$

Assumption: A word is determined completely by its tag. This is inspired by speech recognition

$$= P(w_o|t_o)P(w_1|t_1) \ldots P(w_{n+1}|t_{n+1})$$

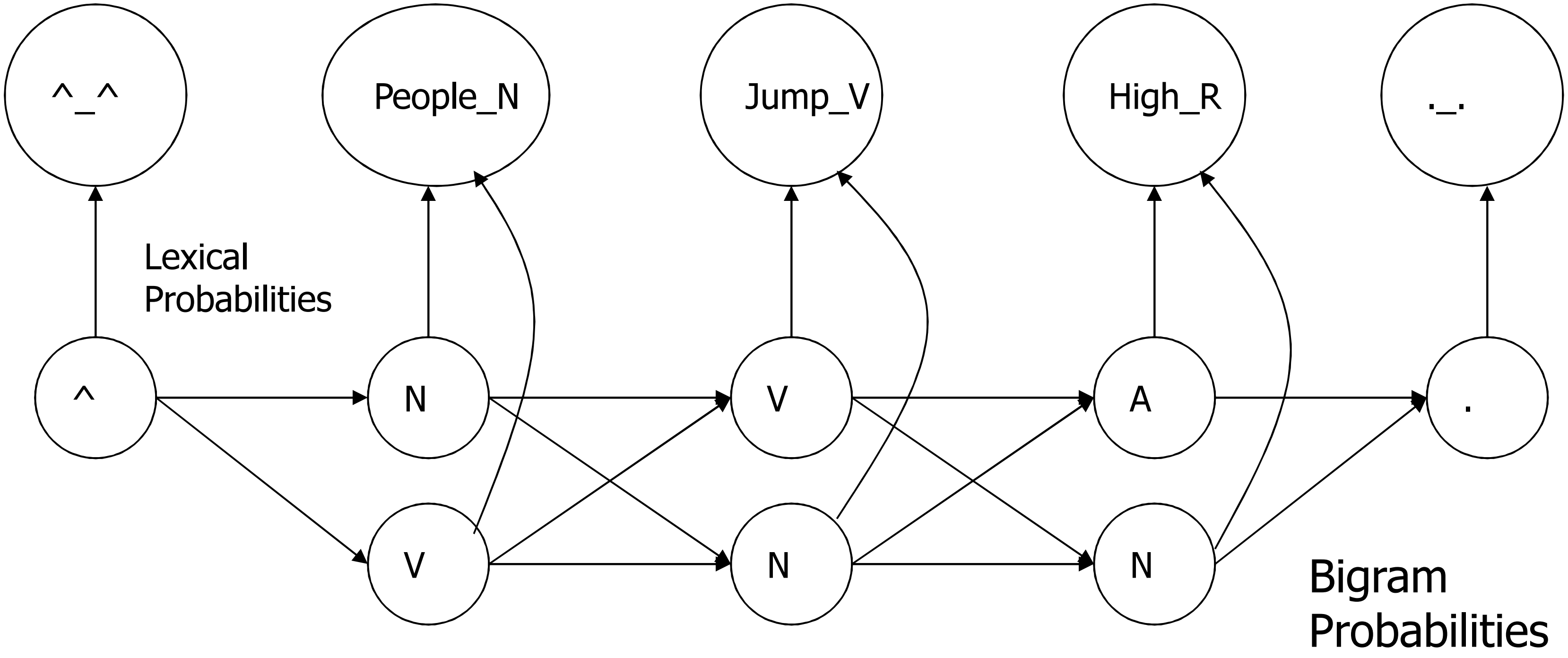
$$= \prod_{i=0}^{n+1} P(w_i|t_i)$$

$$= \prod_{i=1}^{n+1} P(w_i|t_i) \qquad \text{(Lexical Probability Assumption)}$$

Reading List

- TnT (http://www.aclweb.org/anthology-new/A/A00/A00-1031.pdf)
- Brill Tagger (http://delivery.acm.org/10.1145/1080000/1075553/p112-brill.pdf?ip=182.19.16.71&acc=OPEN&CFID=129797466&CFTOKEN=72601926&__acm__=1342975719_082233e0ca9b5d1d67a9997c03a649d1)
- Hindi POS Tagger built by IIT Bombay (http://www.cse.iitb.ac.in/pb/papers/ACL-2006-Hindi-POS-Tagging.pdf)
- Projection (http://www.dipanjandas.com/files/posInduction.pdf)

Generative Model

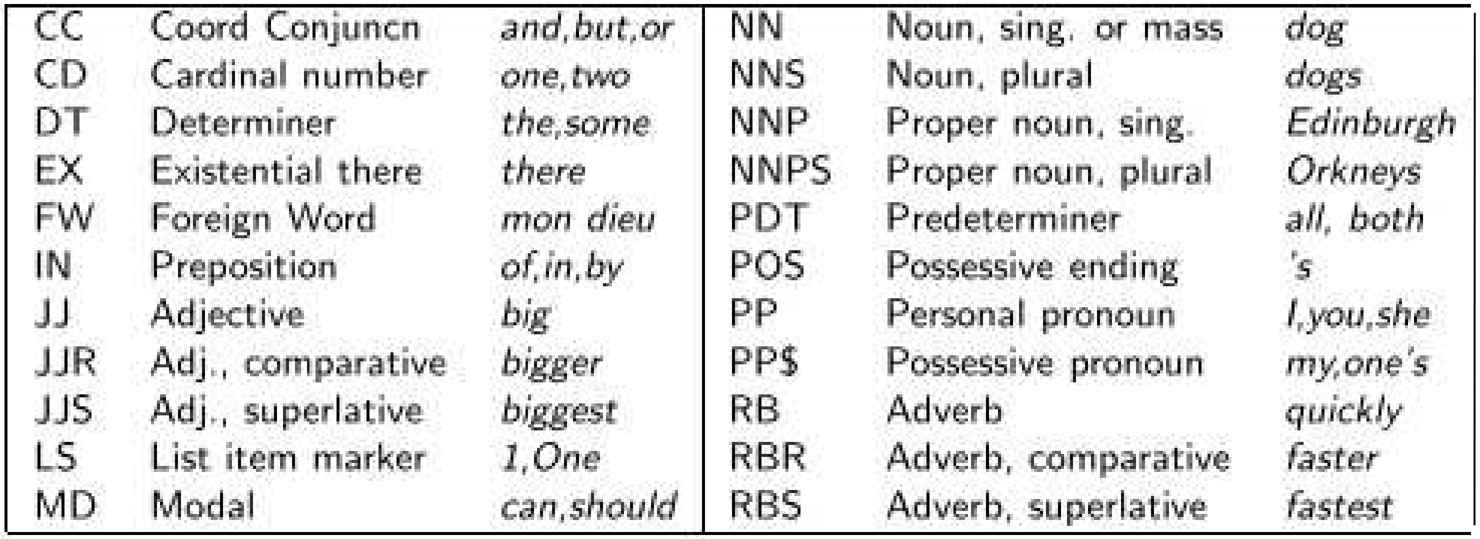


This model is called Generative model.
Here words are observed from tags as states.
This is similar to HMM.

Observations leading to why probability is needed

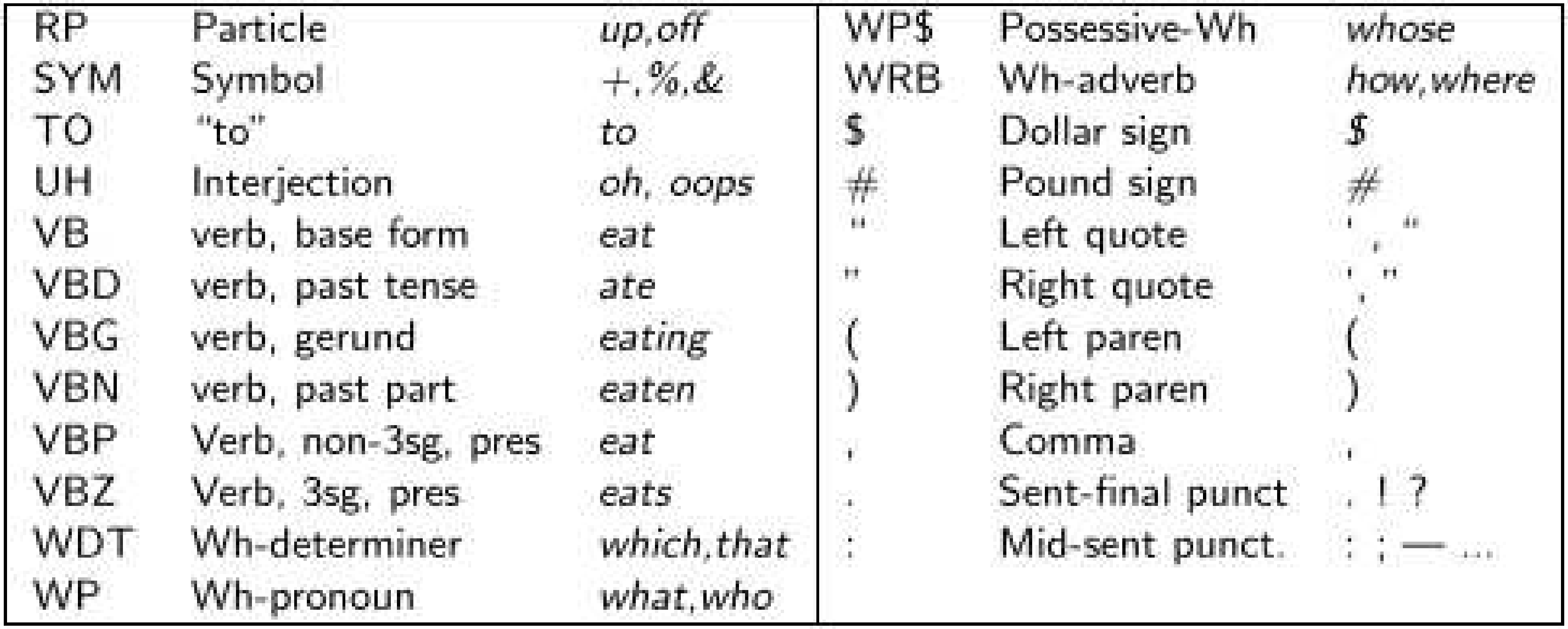
- Many intelligence tasks are sequence labeling tasks
- Tasks carried out in layers
- Within a layer, there are limited windows of information
- This naturally calls for strategies for dealing with uncertainty
- Probability and Markov process give a way

Penn tagset (1/2)



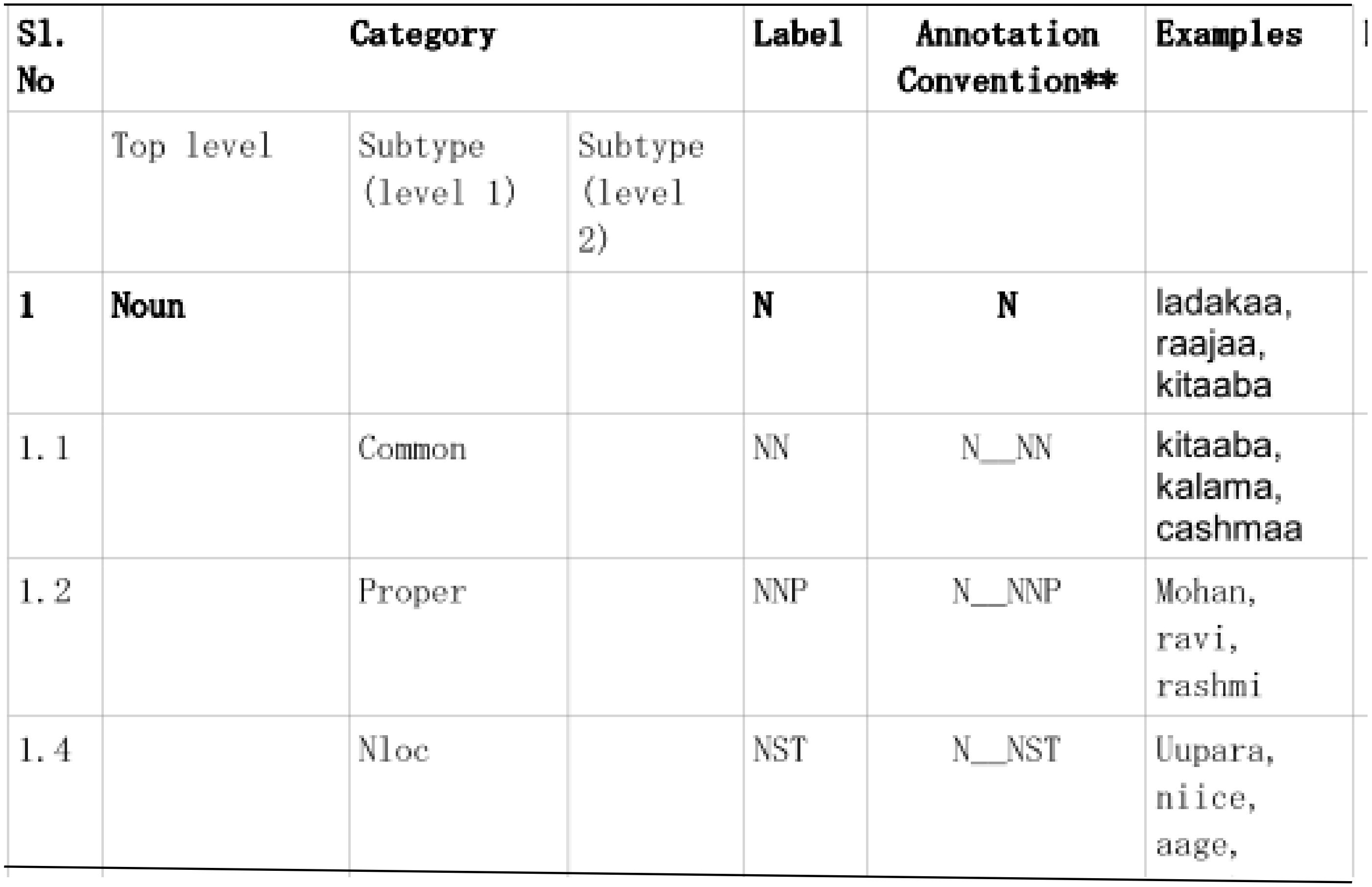
CC	Coord Conjuncn	*and,but,or*	NN	Noun, sing. or mass	*dog*
CD	Cardinal number	*one,two*	NNS	Noun, plural	*dogs*
DT	Determiner	*the,some*	NNP	Proper noun, sing.	*Edinburgh*
EX	Existential there	*there*	NNPS	Proper noun, plural	*Orkneys*
FW	Foreign Word	*mon dieu*	PDT	Predeterminer	*all, both*
IN	Preposition	*of,in,by*	POS	Possessive ending	*'s*
JJ	Adjective	*big*	PP	Personal pronoun	*I,you,she*
JJR	Adj., comparative	*bigger*	PP$	Possessive pronoun	*my,one's*
JJS	Adj., superlative	*biggest*	RB	Adverb	*quickly*
LS	List item marker	*1,One*	RBR	Adverb, comparative	*faster*
MD	Modal	*can,should*	RBS	Adverb, superlative	*fastest*

Penn tagset (2/2)



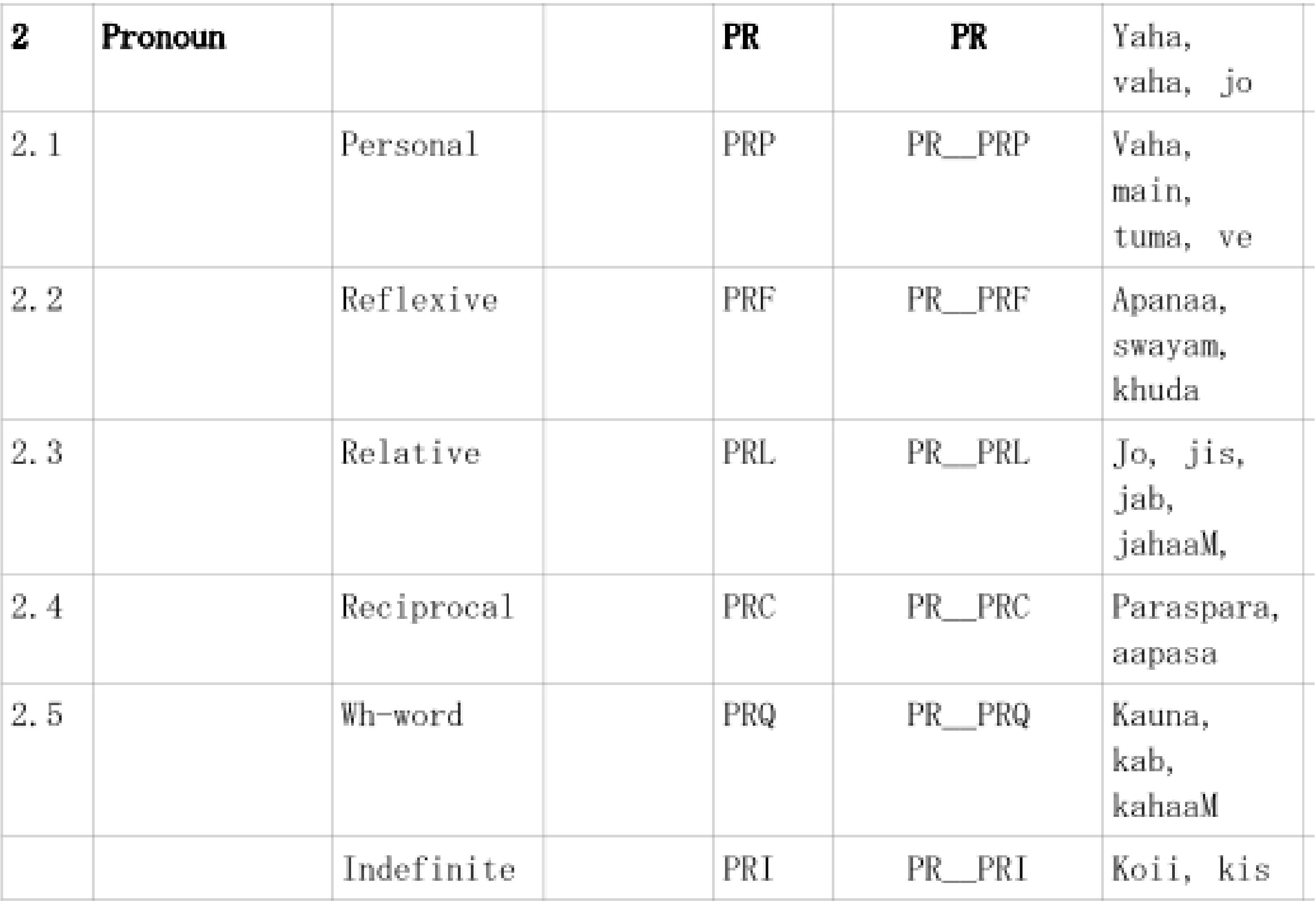
RP	Particle	*up,off*	WP$	Possessive-Wh	*whose*
SYM	Symbol	+,%,&	WRB	Wh-adverb	*how,where*
TO	"to"	*to*	$	Dollar sign	$
UH	Interjection	*oh, oops*	#	Pound sign	#
VB	verb, base form	*eat*	“	Left quote	‘ , “
VBD	verb, past tense	*ate*	”	Right quote	’ , ”
VBG	verb, gerund	*eating*	(	Left paren	(
VBN	verb, past part	*eaten*	)	Right paren	)
VBP	Verb, non-3sg, pres	*eat*	,	Comma	,
VBZ	Verb, 3sg, pres	*eats*	.	Sent-final punct	. ! ?
WDT	Wh-determiner	*which,that*	:	Mid-sent punct.	: ; — ...
WP	Wh-pronoun	*what,who*			

Indian Language Tagset: Noun



Sl. No	Category			Label	Annotation Convention**	Examples
	Top level	Subtype (level 1)	Subtype (level 2)			
1	**Noun**			**N**	**N**	ladakaa, raajaa, kitaaba
1.1		Common		NN	N__NN	kitaaba, kalama, cashmaa
1.2		Proper		NNP	N__NNP	Mohan, ravi, rashmi
1.4		Nloc		NST	N__NST	Uupara, niice, aage,

Indian Language Tagset: Pronoun



2	**Pronoun**			**PR**	**PR**	Yaha, vaha, jo
2.1		Personal		PRP	PR__PRP	Vaha, main, tuma, ve
2.2		Reflexive		PRF	PR__PRF	Apanaa, swayam, khuda
2.3		Relative		PRL	PR__PRL	Jo, jis, jab, jahaaM,
2.4		Reciprocal		PRC	PR__PRC	Paraspara, aapasa
2.5		Wh-word		PRQ	PR__PRQ	Kauna, kab, kahaaM
		Indefinite		PRI	PR__PRI	Koii, kis

Challenge of POS tagging

Example from Indian Language

Tagging of *jo*, *vaha*, *kaun* and their inflected forms in Hindi and their equivalents in multiple languages

DEM and PRON labels

- ***Jo_DEM*** *ladakaa kal aayaa thaa, vaha cricket acchhaa khel letaa hai*

- ***Jo_PRON*** *kal aayaa thaa, vaha cricket acchhaa khel letaa hai*

Disambiguation rule-1

- ***If***
  - ***Jo is followed by noun***
- ***Then***
  - ***DEM***
- ***Else***
  - ***…***

False Negative

- When there is arbitrary amount of text between the *jo* and the noun
- *Jo_??? **bhaagtaa huaa, haftaa huaa, rotaa huaa, chennai academy a koching lenevaalaa** ladakaa kal aayaa thaa, vaha cricket acchhaa khel letaa hai*

False Positive

- *Jo_DEM* (wrong!) ***duniyadarii samajhkar chaltaa hai, …***
- *Jo_DEM/PRON? manushya manushyoM ke biich ristoM naatoM ko samajhkar chaltaa hai, …* (ambiguous)

Morphology: syncretism

Languages that are poor in Morphology (Chinese, English) have Role Ambiguity or **Syncretism** (fusion of originally different inflected forms resulting in a reduction in the use of inflections)

Eg: *You/They/He/I will <u>come</u> tomorrow*

Here, just by looking at the verb '*come*' its syntactic features aren't apparent i.e.

Gender, Number, Person, Tense, Aspect, Modality (GNPTAM)

-<u>Aspect</u> tells us how the event occurred; whether it is completed, continuous, or habitual. Eg: *John came, John will be coming*

- <u>Modality</u> indicates possibility or obligation. Eg: *John can arrive / John must arrive*

Contrast this with the Hindi Translation of *'I will come tomorrow'*

मैं Main (I) *कल kal(tomorrow)* *आउंगा aaunga (will come)*

आउंगा aaunga – GNPTAM: Male, Singular, First, Future

आओगे (Aaoge) – has number ambiguity, but still contains more information than '*come*' in English

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 proposed 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 Speech Processing
  - Phonology and Phonetics
  - Statistical Approach
  - Automatic Speech Recognition and Speech Synthesis
- Deep Parsing
  - Classical Approaches: Top-Down, Bottom-UP and Hybrid Methods
  - Chart Parsing, Earley Parsing
  - Statistical Approach: Probabilistic Parsing, Tree Bank Corpora

Topics proposed 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
  - Paper-reading/Seminar

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

Conclusions

- Both Linguistics and Computation needed
- Linguistics is the eye, Computation the body
- Phenomenon→ Fomalization→Technique→Experimentation→Evaluation→Hypothesis Testing
    - has accorded to NLP the prestige it commands today
- Natural Science like approach
- Neither Theory Building nor Data Driven Pattern finding can be ignored