CS460/626 : Natural Language Processing/Speech, NLP and the Web (Lecture 1 – Introduction)

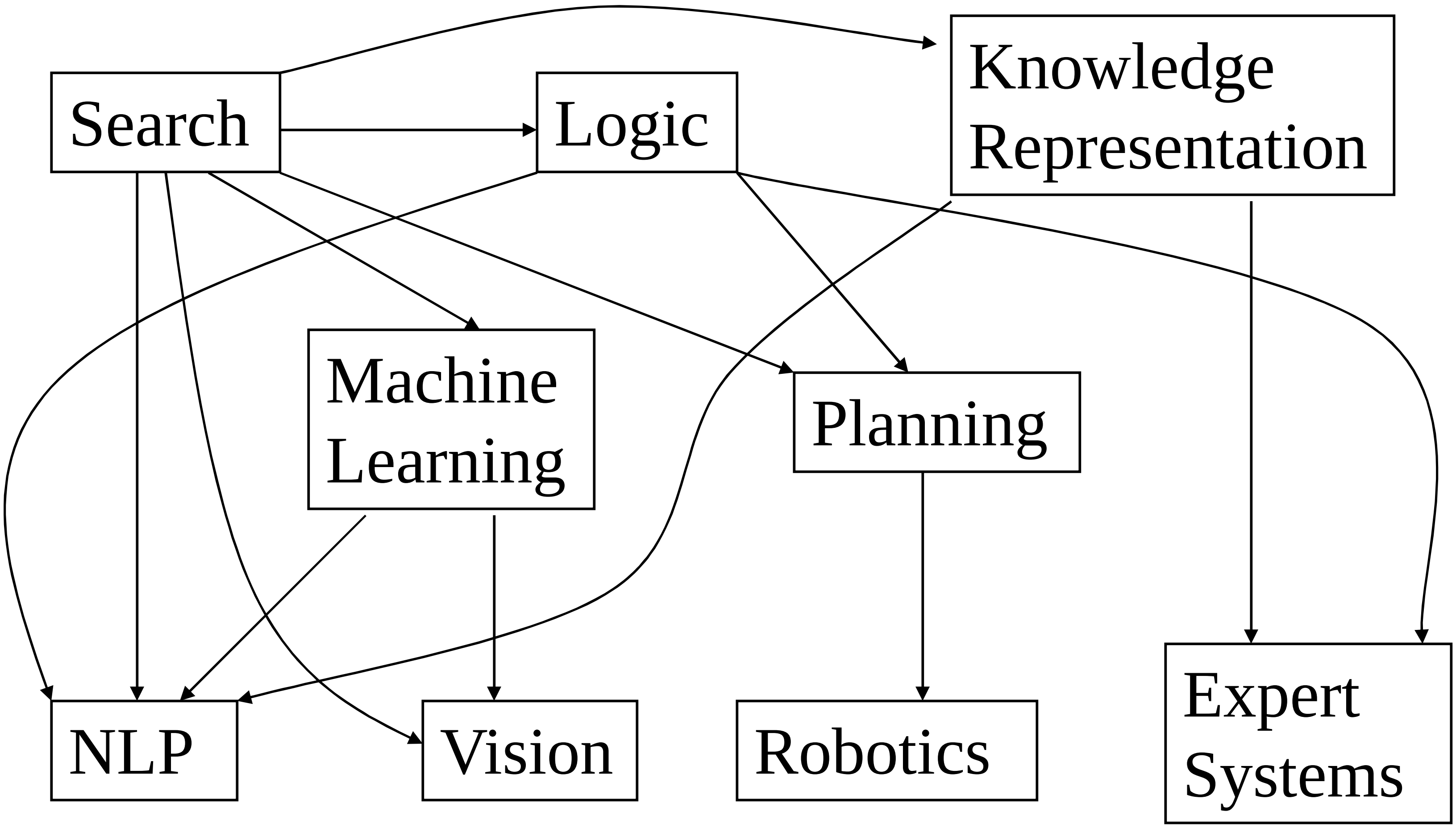
Pushpak Bhattacharyya
CSE Dept.,
IIT Bombay

2nd Jan, 2012

Persons involved

- Faculty instructors: Dr. Pushpak Bhattacharyya (www.cse.iitb.ac.in/~pb)
- TAs: Somya Gupta, Subhabrata Mukherjee {somya, subhabratam}@cse
- Course home page (to be created)
  - http://www.cse.iitb.ac.in/~cs626-460-2012

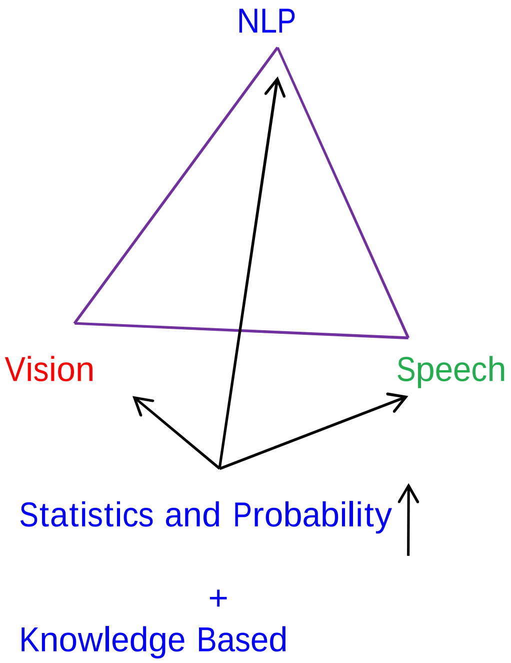
Perpectivising NLP: Areas of AI and their inter-dependencies

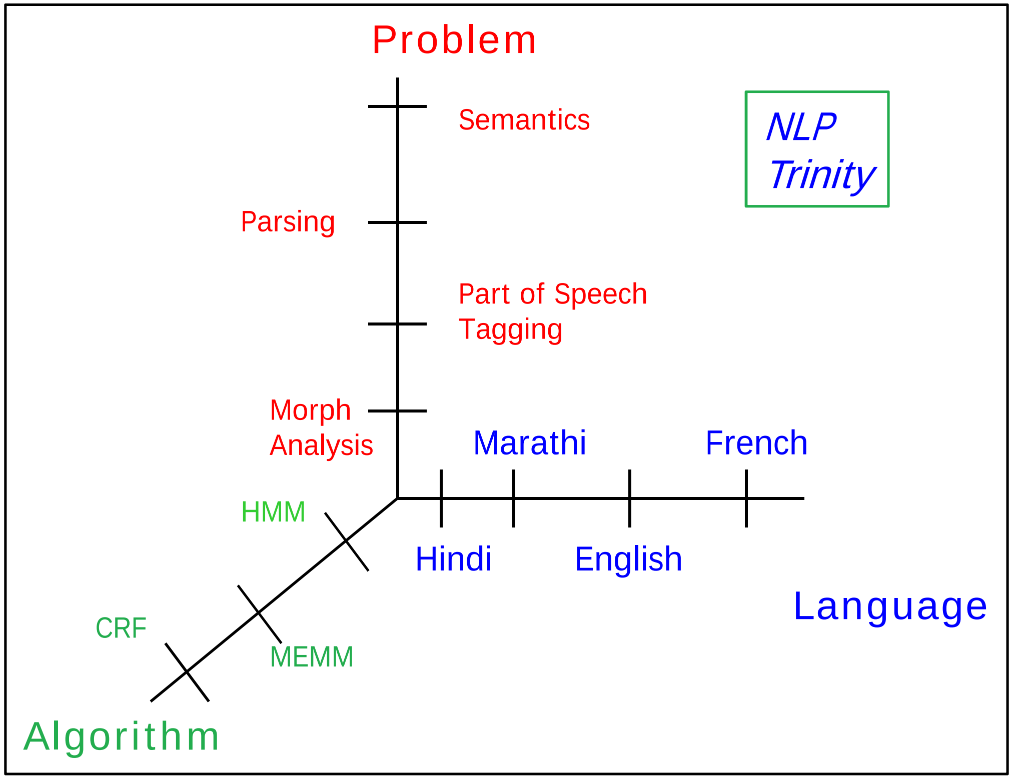


What is NLP

- 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 pictures





Two Views of NLP and the Associated Challenges

1. Classical View
2. Statistical/Machine Learning View

Stages of processing

- 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: *maatraa* 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*

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 come tomorrow*

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

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

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

- Modality 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

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)

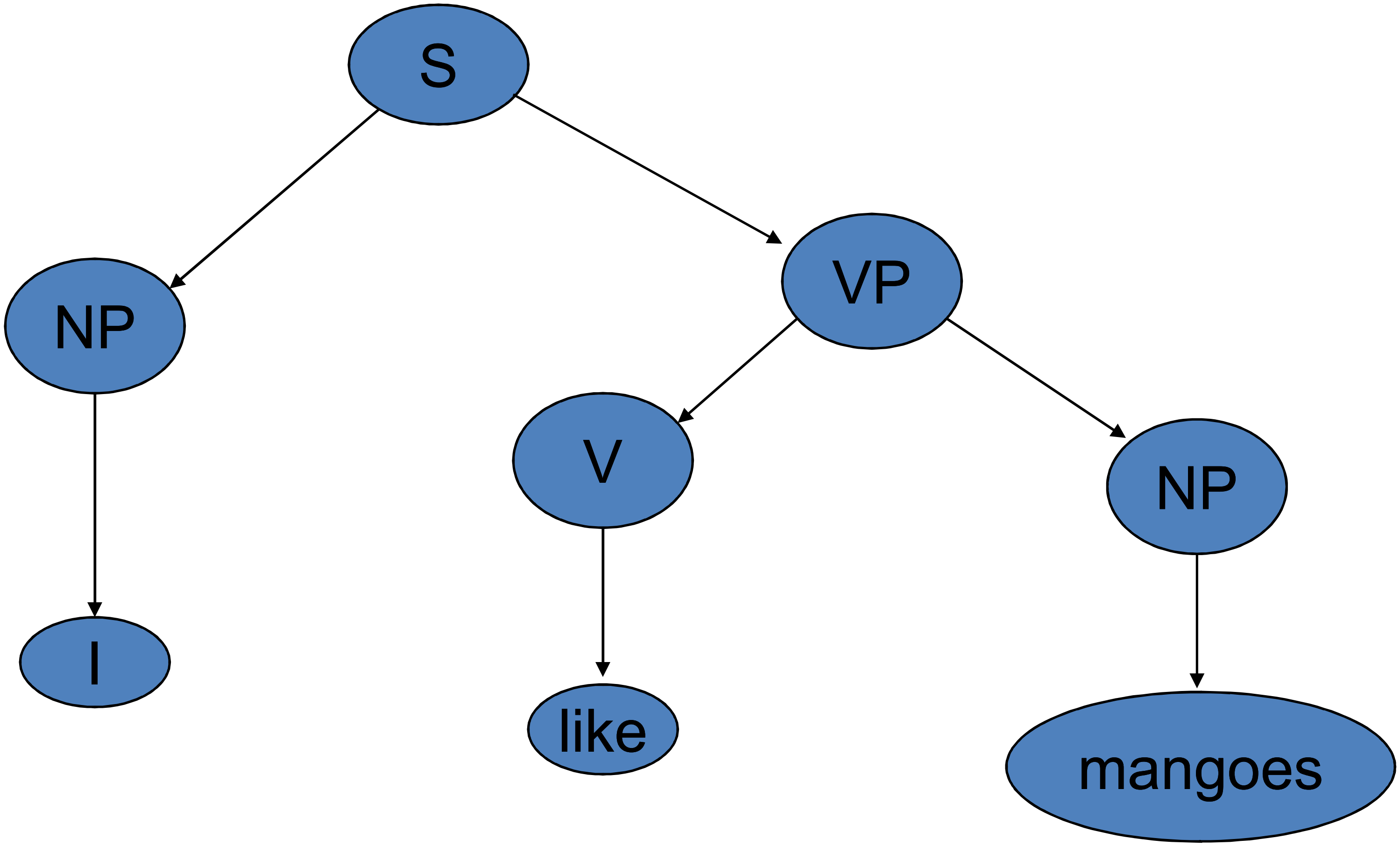
e.g., Ground breaking ceremony/research

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

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"*

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*
- (P.G. Wodehouse, *Ring in Jeeves) Jill had rubbed ointment on Mike the Irish Terrier, taken a look at the goldfish belonging to the cook, which had caused anxiety in the kitchen by refusing its ant's eggs…*
- (Times of India, 26/2/08) *Aid for kins of cops killed in terrorist attacks*

Headache for Parsing: Garden Path sentences

- Garden Pathing
  - *The horse raced past the garden fell.*
  - *The old man the boat.*
  - *Twin Bomb Strike in Baghdad kill 25* (Times of India 05/09/07)

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

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

Part of Speech Tagging

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 Tags

- NN – Noun; e.g. *Dog_NN*
- VM – Main Verb; e.g. *Run_VM*
- VAUX – Auxiliary Verb; e.g. *Is_VAUX*
- JJ – Adjective; e.g. *Red_JJ*
- PRP – Pronoun; e.g. *You_PRP*
- NNP – Proper Noun; e.g. *John_NNP*
- etc.

POS Tag Ambiguity

- In English : I bank$_1$ on the bank$_2$ on the river bank$_3$ for my transactions.
  - Bank$_1$ is verb, the other two banks are noun

- In Hindi :
  - ”Khaanaa” : can be noun (food) or verb (to eat)
  - Mujhe khaanaa khaanaa hai. (first khaanaa is noun and second is verb)

For Hindi

- *Rama achhaa gaata hai*. (hai is VAUX : Auxiliary verb); *Ram sings well*
- *Rama achha ladakaa hai*. (hai is VCOP : Copula verb); *Ram is a good boy*

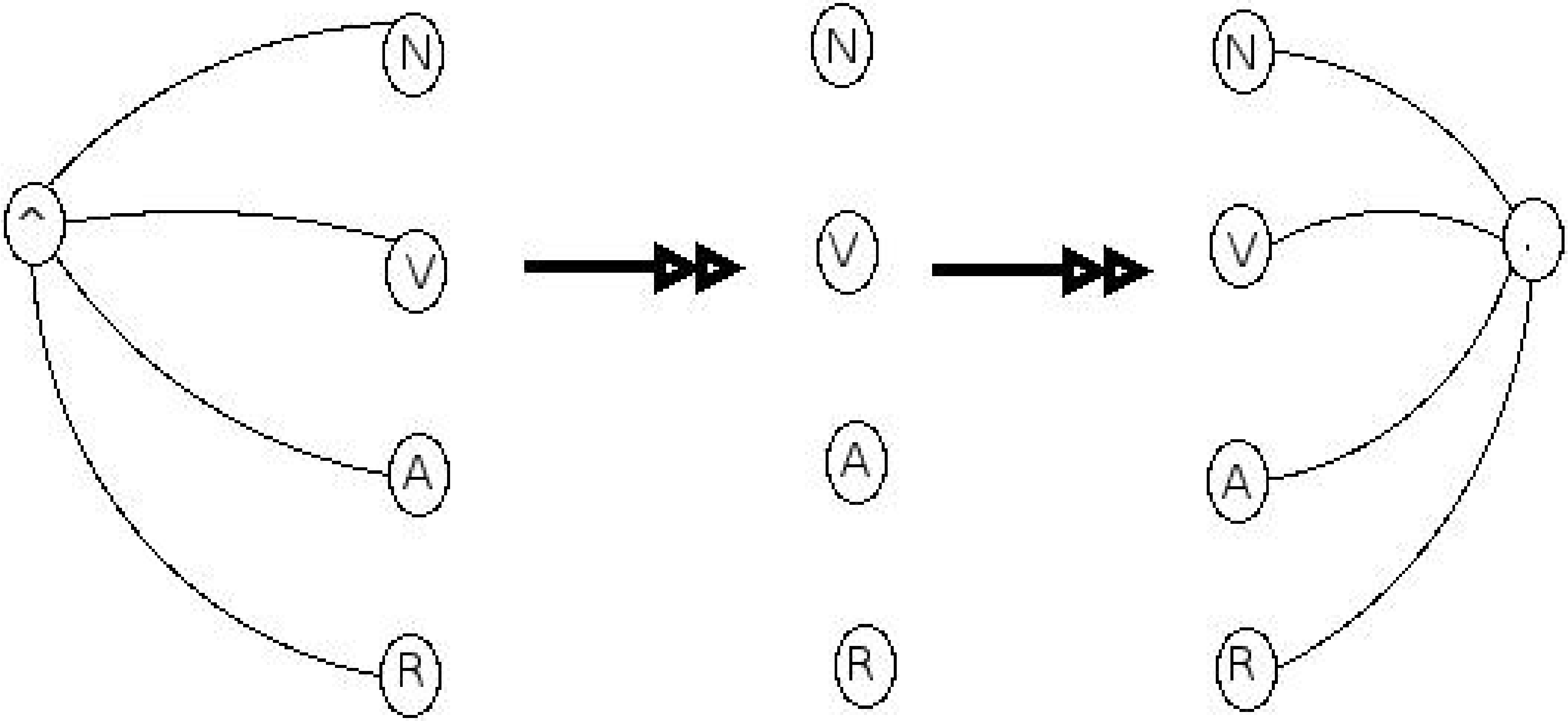
Process

- List all possible tag for each word in sentence.
- Choose best suitable tag sequence.

Example

- ”People jump high”.
- People : Noun/Verb
- jump : Noun/Verb
- high : Noun/Adjective
- We can start with probabilities.

^ People jump high .



Bigram Assumption

Best tag sequence

= T*

= argmax P(T|W)

= argmax P(T)P(W|T) (by Baye's Theorem)

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

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

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

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

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

Lexical Probability Assumption

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

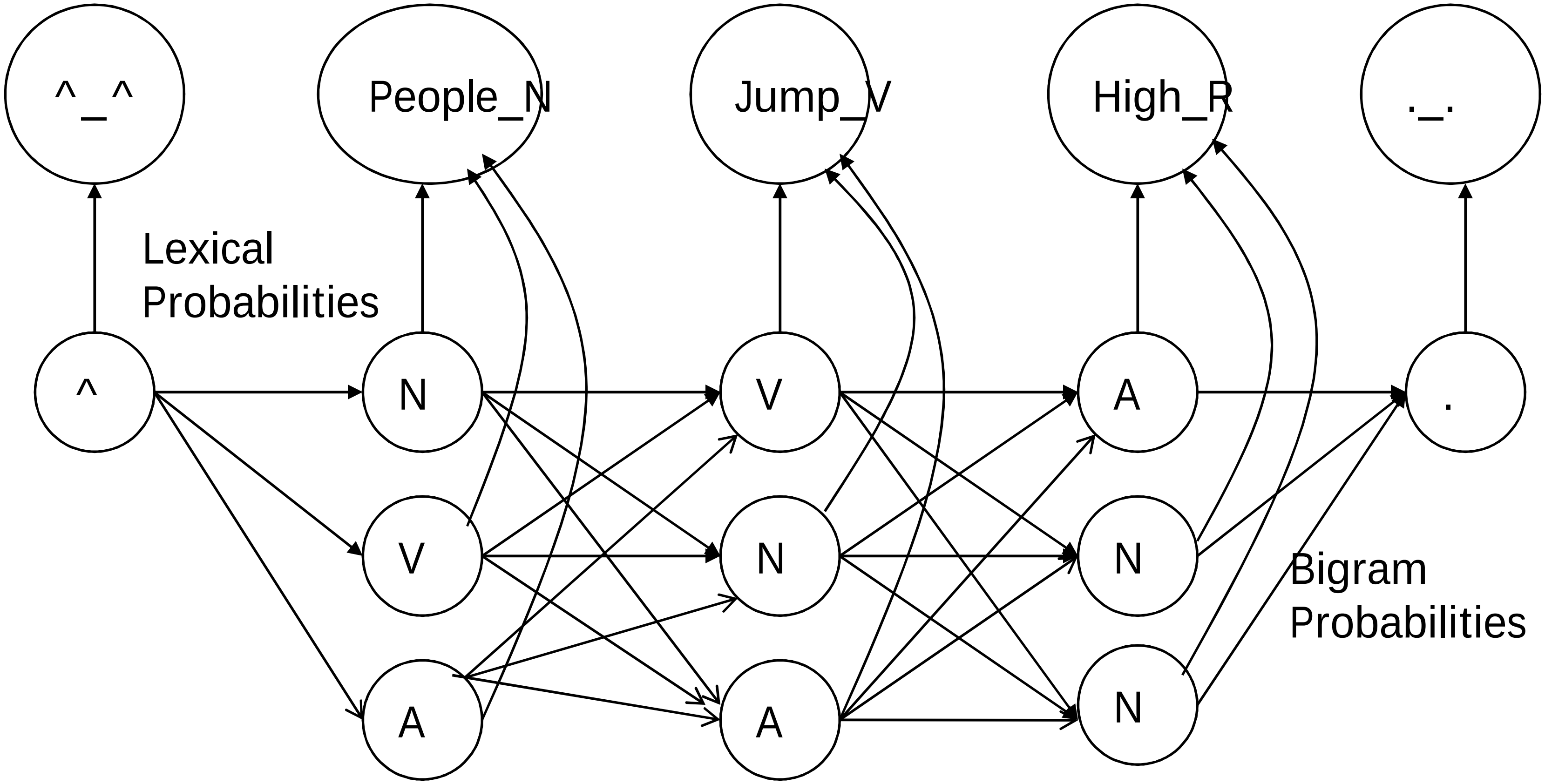
Assumption: A word is determined completely by its tag. This is inspired by speech recognition

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

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

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

Generative Model



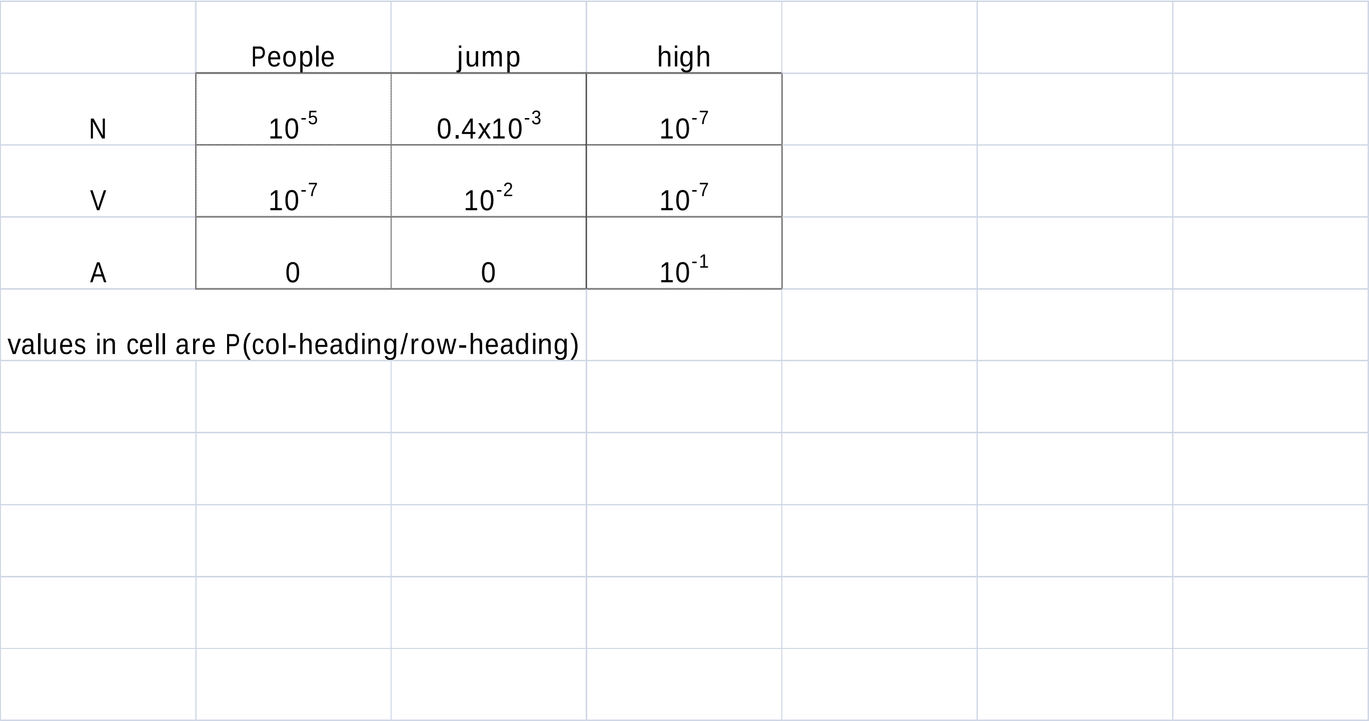
This model is called Generative model.
Here words are observed from tags as states.
This is similar to HMM.

Bigram probabilities

-

	$\boldsymbol{N}$	$\boldsymbol{V}$	$\boldsymbol{A}$
$\boldsymbol{N}$	0.2	0.7	0.1
$\boldsymbol{V}$	0.6	0.2	0.2
$\boldsymbol{A}$	0.5	0.2	0.3

Lexical Probability

- 

	People	jump	high
N	10^{-5}	$0.4x10^{-3}$	10^{-7}
V	10^{-7}	10^{-2}	10^{-7}
A	0	0	10^{-1}

values in cell are P(col-heading/row-heading)

Calculation from actual data

- Corpus
  - ^ *Ram got many NLP books. He found them all very interesting.*
- Pos Tagged
  - ^ *N V A N N . N V N A R A .*

Recording numbers

	^	N	V	A	R	.
^	0	2	0	0	0	0
N	0	1	2	1	0	1
V	0	1	0	1	0	0
A	0	1	0	0	1	1
R	0	0	0	1	0	0
.	1	0	0	0	0	0

Probabilities

	^	N	V	A	R	.
^	0	1	0	0	0	0
N	0	1/5	2/5	1/5	0	1/5
V	0	1/2	0	1/2	0	0
A	0	1/3	0	0	1/3	1/3
R	0	0	0	1	0	0
.	1	0	0	0	0	0

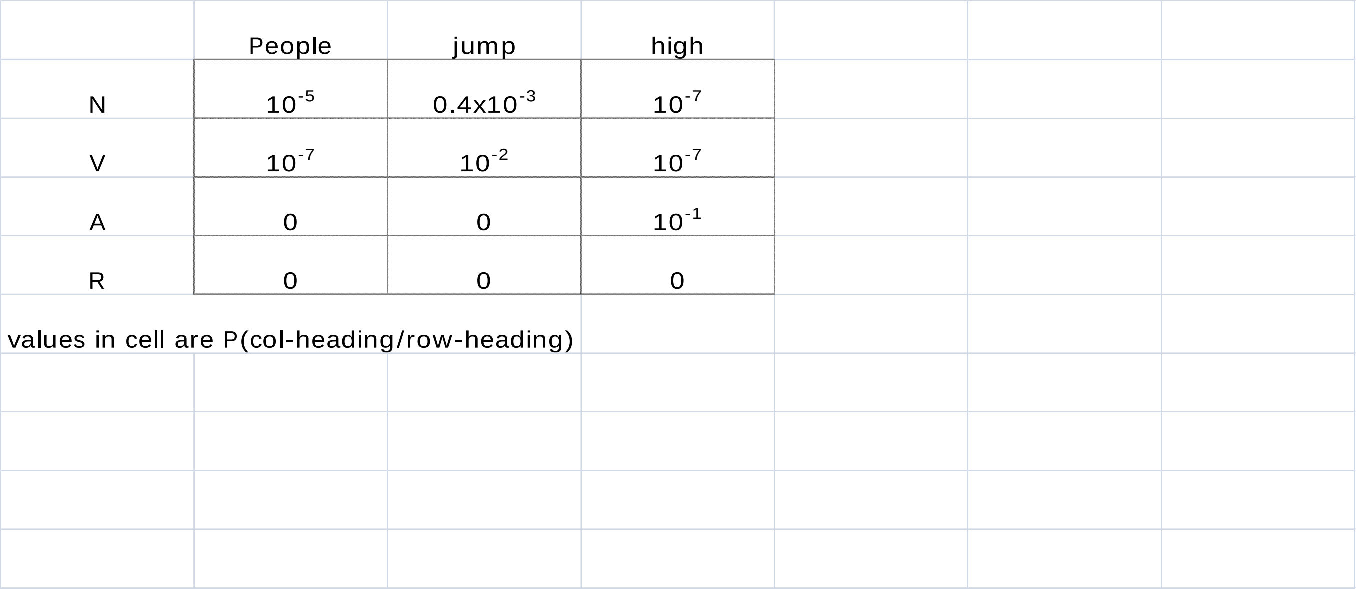
To find

- $T^* = argmax\ (P(T)\ P(W/T))$
- $P(T).P(W/T) = \prod_{i=1 \rightarrow n} P(\ t_i\ /\ t_{i+1}\ ).P(w_i\ /t_i)$
- $P(\ t_i\ /\ t_{i+1}\ )$: *Bigram probability*
- $P(w_i\ /t_i)$: *Lexical probability*

Bigram probabilities

	N	*V*	*A*	*R*
N	0.15	0.7	0.05	0.1
V	0.6	0.2	0.1	0.1
A	0.5	0.2	0.3	0
R	0.1	0.3	0.5	0.1

Lexical Probability

- 

	People	jump	high
N	10^{-5}	$0.4\text{x}10^{-3}$	10^{-7}
V	10^{-7}	10^{-2}	10^{-7}
A	0	0	10^{-1}
R	0	0	0

values in cell are P(col-heading/row-heading)

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, Chaitanya 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