CS626: Speech, NLP and Web

chatGPT giveaways, pramaanas, CLT, LoLN, Dependency Parsing

Pushpak Bhattacharyya

Computer Science and Engineering Department

IIT Bombay

Week 12 of 21st October, 2024

1-slide recap of week of 14th Oct

- Hypothesis Testing: does the conclusion from the sample hold for the population
- VIMP: the NULL hypothesis H_0
- IMP: the confidence interval (usually 95%, 99% and 90%), level of significance (1-confidence in decimal), p-value (the probability of the observation under H_0)
- HT is an exercise similar to proof by contradiction: to show that if H0 is true then the observation is of low probability
- Type-I and Type-II errors: always wrt to H_0
- Type-I: H_0 erroneously rejected; Type-II: H_0 is erroneously accepted

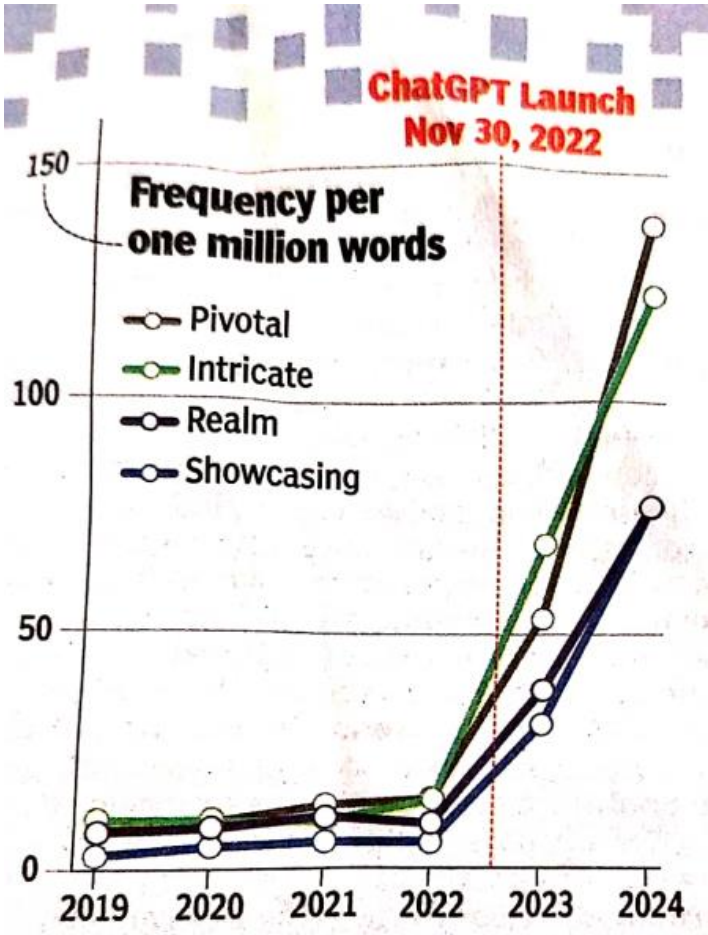
ChatGPT giveaways

ChatGPT, is that you? (ToI 13oct24) (1/2)

- Teachers' difficulty: own writing or result of chatGPT outsourcing?
- Givaways- words chatGPT over-relies on:
- 'delve'- Prof. Jeremy Nguyen, Swinburne Business School, Australia finds the use of this word increase exponentially in PubMed since 2023 when chatGPT came into the scene
- 'additionally', 'nevertheless', 'a testament to…', 'it's important to consider…'- points out the data analyst Margaret Efron
- Overly complex sentence structures
- An out of place formal tone

ChatGPT, is that you? (ToI 13oct24) (2/2)

- Overuse of words like ‘realm’, ‘intricate’, ‘showcasing’, ‘pivotal’- a Stanford study points out that the use of these words exploded since chatGPT, through an analysis of millions of words
- Many AI detection tools online- but are vulnerable



- AI images- hands are giveaways
- Text-image combination nonsensical
- When a picture is too perfect…
- Many AI detection tools online- but are vulnerable

Prompt and Pramaana

“Prompt AI these Pramanas”, Bindra and Saxena, ToI 5oct24 (1/2)

- Many intriguing connections between AI and some of our ancient texts
- **Pramana**- means of acquiring knowledge
- **Pratyaksha** (perception)- direct sensory experience; *fire is hot*
- **Anumana** (inference)- Inference; *seeing smoke and inferring fire*

“Prompt…” (2/2)

- **Upamana (comparison)-** know unfamiliar from familiar; *learn what a zebra is from horse*
- **Sabda (verbal testimony)-** knowledge from trusted sources; e.g. *authentic texts*
- **Arthapatti (postulation)-** explaining something unobserved from observed; *somebody growing fat; must be eating at night despite fasting during day*
- **Anupalabdhi (non-perception)-** knowledge from absence; e.g., *a book missing from a gap in the library rack*

Autonomous Vehicles and Pramaana (1/2)

- "Slows down at zebra crossings, stops at signals, swerves around obstacles, adjusts to unexpected changes"
- **What is AV's source of Pramana**?
- **Pratyaksha** (perception)- *sensors, cameras and LIDAR*
- **Anumana** (inference)- *pedestrian nearing the street → slow down*
- **Upamana (comparison)-** *stop for zebra from the experience of stopping seeing a horse last time*

Autonomous Vehicles and Pramaana (2/2)

- **Sabda (verbal testimony)-** trained on large datasets; learn to recognize stop signs from labelled images
- **Arthapatti (postulation; currently very difficult!)-** explaining something unobserved from observed; *a cop waving the cars around in a scene of accident*
- **Anupalabdhi (non-perception)-** knowledge from absence; e.g., *faded lane, missing traffic light*

End of chatgpt-giveaways, pramaanas

Central Limit Theorem (CLT)

Statement of Central Limit Theorem

- Let $X_1, X_2, X_3, \ldots, X_n$ be n independent random variables, each with mean μ and variance σ^2
- Also let

  $S_n = X_1 + X_2 + X_3, \ldots + X_n$
- Then,

  the following is **standard normal**

$$S_n^* = \frac{S_n - n\mu}{\sigma\sqrt{n}}$$

Mathematical adjustment

$$S_n^* = \frac{S_n - n\mu}{\sigma\sqrt{n}}, \text{ gives}$$

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} = \frac{\frac{S_n}{n} - \mu}{\frac{\sigma\sqrt{n}}{n}} = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Hence, another equivalent statement of CLT

Let $X_1, X_2, X_3, \ldots, X_n$ be n independent random variables forming a sample from a population with mean μ and variance σ^2.

Then the sample mean is normally distributed with mean μ and variance σ^2/n.

MGF

Moment Generating Function

$M_X(t)=E(e^{tX})$, X is a random Variable and $f(x_j)=P(X=x_j)$

for discrete distribution

$$M_X(t)=\sum_{j=1}^{n} e^{tx_j} f(x_j)$$

for continuous distribution

$$M_X(t)=\int_{-\infty}^{+\infty} e^{tx} f(x)dx$$

Significance of MGF (1/2)

- The n^{th} derivatives of the MGF at $t=0$ gives the nth moment of the distribution of the random variable

- Thus the 1^{st} derivative at $t=0$ gives the mean of the distribution

- The 2^{nd} derivative at $t=0$ minus the 1^{st} derivative at $t=0$ gives the variance

Significance of MGF (2/2)

- Similarly the 3rd derivative at *t=0* along with the lower derivatives (combined with proper operators) at *t=0* gives the skewness, i.e., how symmetric is the about the mean

- *4th* derivative at *t=0* similarly can lead to the kurtosis, i.e., how heavily the tails of a distribution differ from the tails of a normal distribution

Proof regarding n^{th} derivative and n^{th} moment

$$M_X'(t) = \frac{d}{dt} E(e^{tX})$$

$$= E\left[\frac{d}{dt}(e^{tX})\right]$$

$$= E[Xe^{tX}]$$

$$= E(X)$$

$$= M_X'(0)$$

$$M_X''(t) = \frac{d}{dt} M_X'(t)$$

$$= \frac{d}{dt} E(Xe^{tX})$$

$$= E\left[\frac{d}{dt}(Xe^{tX})\right]$$

$$= E[X^2 e^{tX}]$$

$$= E(X^2); \quad at \quad t = 0$$

$$\therefore \text{var}(X) = M_X''(0) - [M_X'(0)]^2$$

Uniqueness Theorem

- Suppose X and Y are random variables having moment generating functions $M_X(t)$ and $M_Y(t)$ respectively.

- Then X and Y have the same probability distribution if and only if $M_X(t)=M_Y(t)$ identically.

Standard Normal Distribution, *N(0,1)* and its PDF

Normal:

$$P(y) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y-\mu)^2}{2\sigma} \right)$$

Standard normal:

$$P(Z = y) = \frac{1}{\sqrt{2\pi}} \exp\left(-y^2 / 2 \right)$$

MGF of *N(0,1)*

$$MGF = \int_{-\infty}^{+\infty} e^{ty} \frac{1}{\sqrt{2\pi}} e^{\left(-y^2/2\right)} dy$$

$$= \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{\left(-y^2/2+ty\right)} dy$$

$$= \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y^2-2yt+t^2)} . e^{\frac{t^2}{2}} dy$$

$$= e^{\frac{t^2}{2}} \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y-t)^2} dy$$

$$= e^{\frac{t^2}{2}}$$

Proof of CLT

- To prove that

$$S_n^* = \frac{S_n - n\mu}{\sigma\sqrt{n}}$$

- Is standard normal, we will show that

$$M_{S_n^*}(t) = M_Z(t)$$

- i.e., the moment generating function of S_n* is equal to the moment generating function of standard normal r.v.

Proof: MGF

$$E(e^{tS_n^*}) = E[e^{t(S_n - n\mu)/\sigma\sqrt{n})}]$$

$$= E[e^{t(\sum_{i=1}^{n} X_i - n\mu)/\sigma\sqrt{n})}]$$

$$= E[e^{\sum_{i=1}^{n} t(X_i - \mu)/\sigma\sqrt{n})}]$$

$$= E[\prod_{i=1}^{n} (e^{t(X_i - \mu)/\sigma\sqrt{n})})]$$

$$= \prod_{i=1}^{n} E[e^{t(X_i - \mu)/\sigma\sqrt{n})}]$$

$$= \{E[e^{t(X_i - \mu)/\sigma\sqrt{n})}]\}^n$$

Proof: cntd.

$$\{E[e^{t(X_i-\mu)/\sigma\sqrt{n})}]\}^n$$

now,

$$e^{t(X_i-\mu)/\sigma\sqrt{n})} = [1+\frac{t(X_i-\mu)}{\sigma\sqrt{n}}+\frac{t^2(X_i-\mu)^2}{2\sigma^2 n}+...],$$

by Taylor series exp*ansion*

Proof: working with *E*

$$E[1+\frac{t(X_i-\mu)}{\sigma\sqrt{n}}+\frac{t^2(X_i-\mu)^2}{\sigma^2 n}+...],$$

$$=E(1)+\frac{tE(X_i-\mu)}{\sigma\sqrt{n}}+\frac{t^2E(X_i-\mu)^2}{2\sigma^2 n}+...]$$

$$=1+0+\frac{t^2}{2n}+...$$

As n tends to infinity…

$$E(e^{tS_n^*}) = (1 + \frac{t^2}{2n} + \ldots)^n$$

$$Study\ L_n = (1 + \frac{t^2}{2n} + \ldots)^n,\ as\ n -> \infty$$

$$\log L_n = n \log(1 + \frac{t^2}{2n} + \ldots)$$

$$= \frac{\log(1 + \frac{t^2}{2n} + \ldots)}{1/n}$$

$$Both\ num\ and\ denom \to 0,\quad as\ n -> \infty$$

As n tends to infinity…

take derivative of numerator and numerator as per L'Hospital rule

$$= \frac{(-\frac{t^2}{2n^2})}{(1+\frac{t^2}{2n}+...)/-1/n^2} = \frac{\frac{t^2}{2}}{(1+\frac{t^2}{2n}+...)}$$

$$= \frac{t^2}{2}, \text{ as } n \to \infty$$

same as the mgf of Z

Law of Large Numbers (LoLN)

Intuition for CLT and LoLN

- The larger the number of observations, the more “accurate” the probability value
- E.g., in HMM based POS tagging, suppose we get the transition probability *P(NN|JJ)=0.8*
- This is critical for the POS tagger
- Can we trust this value?
- We can if computed from LARGE amount of data

Estimates from sample → claim about the population, facilitated by
Central Limit Theorem and Law of Large Numbers

Foundation for any probability based work

INTRIGUE of Law of Large Numbers

- I do not (and most often) will not know the population average
- Still I will converge towards it!!
- As the number of observations increases, that trust/confidence grows!!

Weak Law of Large Numbers

(from text book, Sheldon Ross, 2004)

Let $X_1, X_2, \ldots, X_N$ be a sequence of independent and identically distributed random variables, each having mean $E[X_i] = \mu$. Then, for any $\varepsilon > 0$

$$P\left(\left|\frac{X_1 + X_2 + \ldots X_N}{N} - \mu\right| > \varepsilon\right) \to 0, \quad as \quad N \to \infty$$

μ is the population mean.

Strong Law

Let $X_1, X_2, \ldots, X_N$ be a sequence of independent and identically distributed random variables, each having mean $E[X_i] = \mu$.

$$P\left(\left|\frac{X_1 + X_2 + \ldots X_N}{N} - \mu\right| \to 0\right) = 1, \quad as \quad N \to \infty$$

μ is the population mean.

Important implication: note the argument (1/2)

- Informal statement: Since the difference between $(X_1+X_2+\ldots X_N)/N$ and μ can be made as small as possible, by taking more and more N *(*$N \rightarrow$ *infinity)*, the intuition for LARGE DATA requirement is established
- N tends to infinity means large amount of data

Important implication: note the argument (2/2)

- $(X_1+X_2+\ldots X_N)/N \rightarrow \mu$, as $N \rightarrow inf$
- Because of LoLN, I can **TAKE** $(X_1+X_2+\ldots X_N)/N$ as the population mean as N tends to infinity
- IMP: I need not and cannot know μ in almost all cases
- This is surely true about language properties

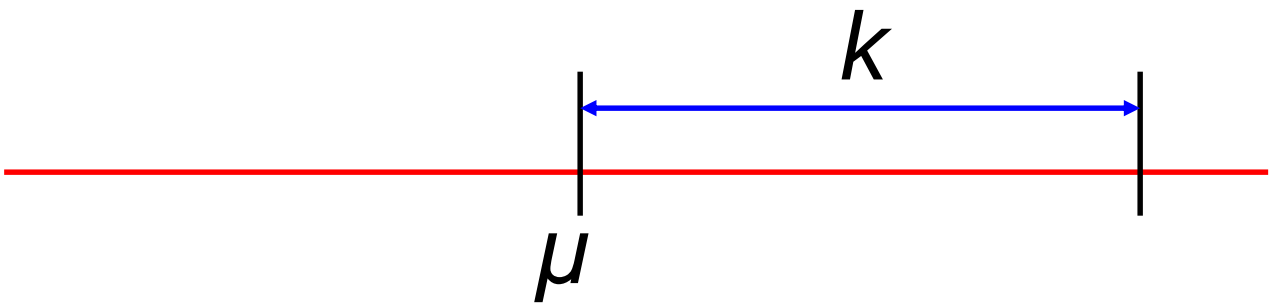
The role of probability

- Strong Law of Large Numbers
  - Demands the **probability to be equal to 1**
- Weak Law of Large Numbers
  - **Uses ε**

Chebyshev's Inequality

If X is a random variable with mean μ and variance σ^2, then for any value k>0

$$P\left[\left|X-\mu\right| \geq k\right] \leq \frac{\sigma^2}{k^2}$$



Markov Inequality

If X is a random variable that takes only non-negative values, then for any *a*>0

$$P(X \geq a) \leq \frac{E(X)}{a}$$

Proof of Chebyshev Inequality

- Use Markov Inequality on $(X-\mu)^2$ with $a=k^2$

$$P\{(X-\mu)^2 \geq k^2\} \leq \frac{E[(X-\mu)^2]}{k^2}$$

- But $(X-\mu)^2>k^2$, *iff* $|X-\mu|>k$, *so*

$$P\{|X-\mu| \geq k\} \leq \frac{E[(X-\mu)^2]}{k^2} = \frac{\sigma^2}{k^2}$$

Another form of Chebyshev's Inequality

If X is a random variable with mean μ and variance σ^2, then for any value k>0

$$P\left[\left|X-\mu\right| \geq k\sigma\right] \leq \frac{1}{k^2}$$

The probability a random variable differs from its mean by more than *k* standard deviations is bounded by $1/k^2$

Completing proof of weak LoLN

Expectation,

$$E\left(\frac{X_1 + X_2 + \ldots X_N}{N}\right) = \mu$$

Variance,

$$V\left(\frac{X_1 + X_2 + \ldots X_N}{N}\right) = \frac{\sigma^2}{N}$$

Now, apply Chebyshev inequality

$$P\left(\left|\frac{X_1 + X_2 + \ldots X_N}{N} - \mu\right| > \varepsilon\right) \leq \frac{\sigma^2}{N\varepsilon^2}$$

As $N \rightarrow$ *infinity*, the above *P(.)* tends to 0

Dependency Parsing

Start of DP

The strongest rain shut down the financial hub of Mumbai

(from: Stanford parser https://nlp.stanford.edu/software/lex-parser.shtml)

Example: POS Tagged sentence

The/DT strongest/JJS rain/NN shut/VBD down/RP the/DT financial/JJ hub/NN of/IN Mumbai/NNP

This has less entropy than the raw sentence, because the POS tags' uncertainty is reduced like for 'rain'

Constituency parse

```
(S
  (NP
      (DT The)
       (JJS strongest)
         (NN rain))
   )

 (VP
     …
```

```
(VP
  (VP
     (VBD shut)
     (PRT (RP down))
  (NP
     (NP
        (DT the) (JJ financial)
        (NN hub))
     (PP (IN of)
          (NP (NNP Mumbai)))))
```

Parse further reduces entropy by, for example, reducing the structural ambiguity, like that of attaching the PP '*of Mumbai*'

Dependency Parse

root(ROOT-0, shut-4)

nsubj(shut-4, rain-3)

prt(shut-4, down-5)

det(rain-3, the-1)

amod(rain-3, strongest-2)

dobj(shut-4, hub-8)

det(hub-8, the-6)

amod(hub-8, financial-7)

prep(hub-8, of-9)

pobj(of-9, Mumbai-10)

Note: dependency parsing chooses to remain shallow; prepositions are NOT Disambiguated wrt their semantic roles.

Examples to illustrate difference between DP and Semantic Role Labeling (SRL)

Sentence	Shallow relation from Dependency Parsing	Deeper relation from Semantic Role Labeling
John broke the window	nsubj	Agent
The stone broke the window	nsubj	Instrument
The window broke	nsubj	Object
1947 saw the freedom of India	nsubj	Time
Delhi saw bloodshed when Nadir Shah attacked Delhi	nsubj	Place

Disambiguation is needed to convert shallow DP relations to semantic roles.

Hindi vs. English (1/2)

Hindi translations uncover different semantic roles:

- *जॉन ने खिड़की तोड़ दी; jon ne khidakee tod dee*
- *पत्थर से खिड़की टूट गयी; patthar se khidakee toot gayee*
- *खिड़की टूट गयी; khidakee toot gayee*
- *1947 में भारत को आज़ादी मिली; 1947 mein bhaarat ko aazaadee milee*
- *जब नादिर शाह ने दिल्ली पर हमला किया तो दिल्ली में खून-खराबा हुआ; jab naadir shaah ne dillee par hamala kiya to dillee mein khoon-kharaaba hua*

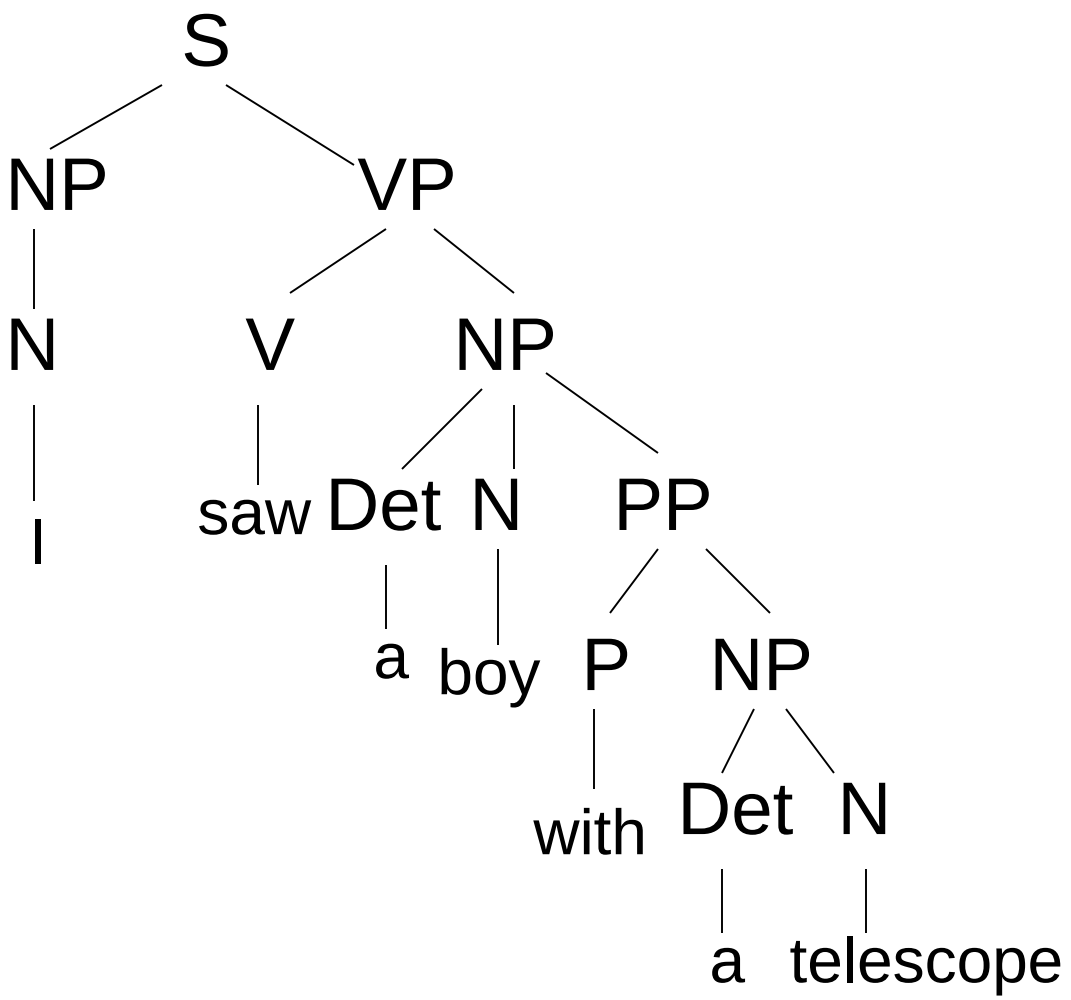
Hindi vs. English (2/2)

- Hindi has signals for semantic difference through case markers
- English is more ambiguous
- But English sentences are more metaphorical
- Ambiguity needed for more colourful and complex linguistic constructs

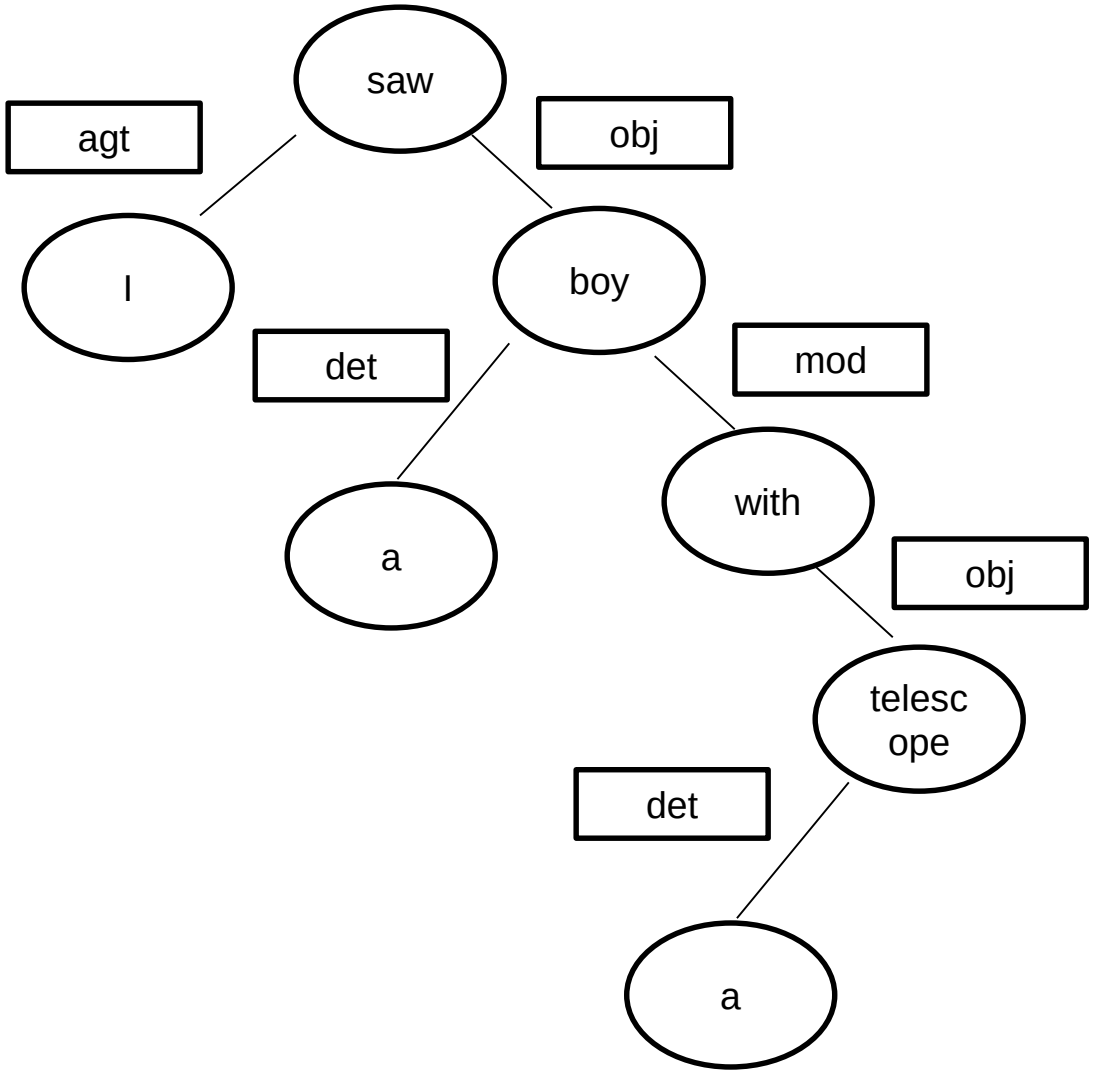
Two kinds of parse representations: Constituency Vs. Dependency

- Penn Constituency Treebank
  - http://www.cis.upenn.edu/~treebank/
- Prague Dependency Treebank
  - http://ufal.mff.cuni.cz/pdt2.0/

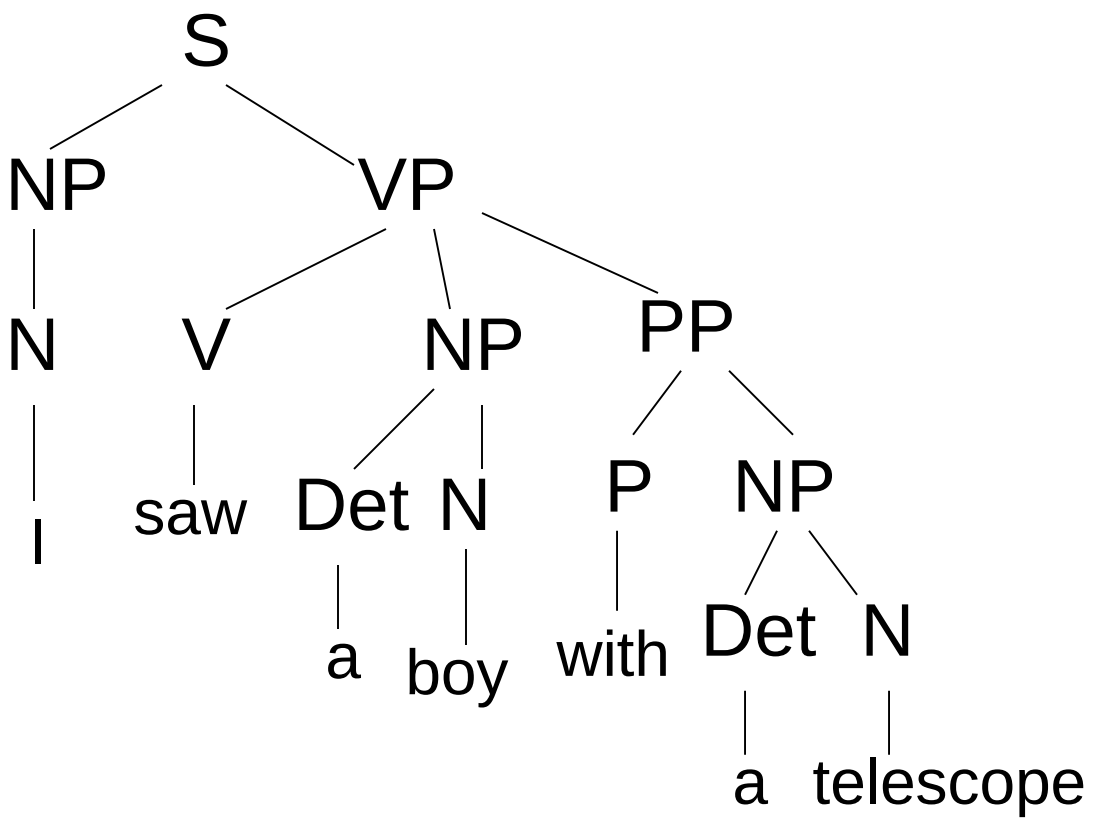
“I saw the boy with a telescope”: Constituency parse-1: *telescope with boy*



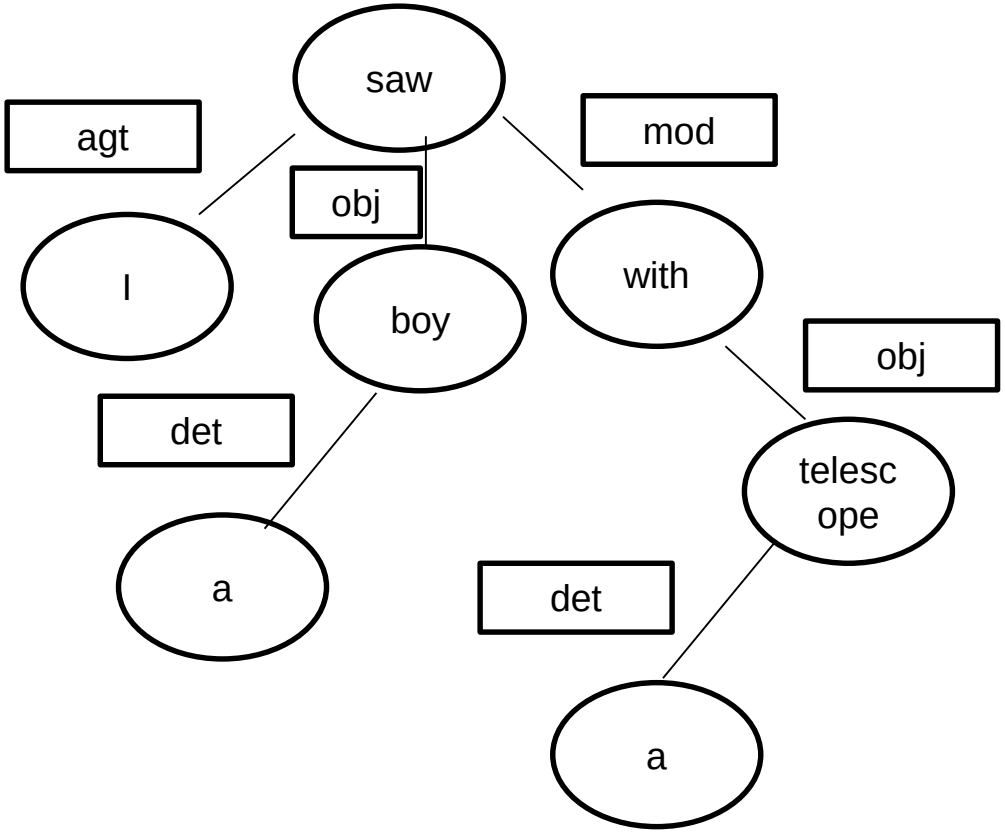
“I saw the boy with a telescope”: Dependency Parse Tree-1



Constituency Parse Tree-2: *telescope with me*



Dependency Parse Tree-2



Advantage of DP over CP

- Related entities are closer in DP than in CP: in terms of path length
- Free word order does not affect DP; CP needs additional rules
- Additional rules may overgeneralize!!

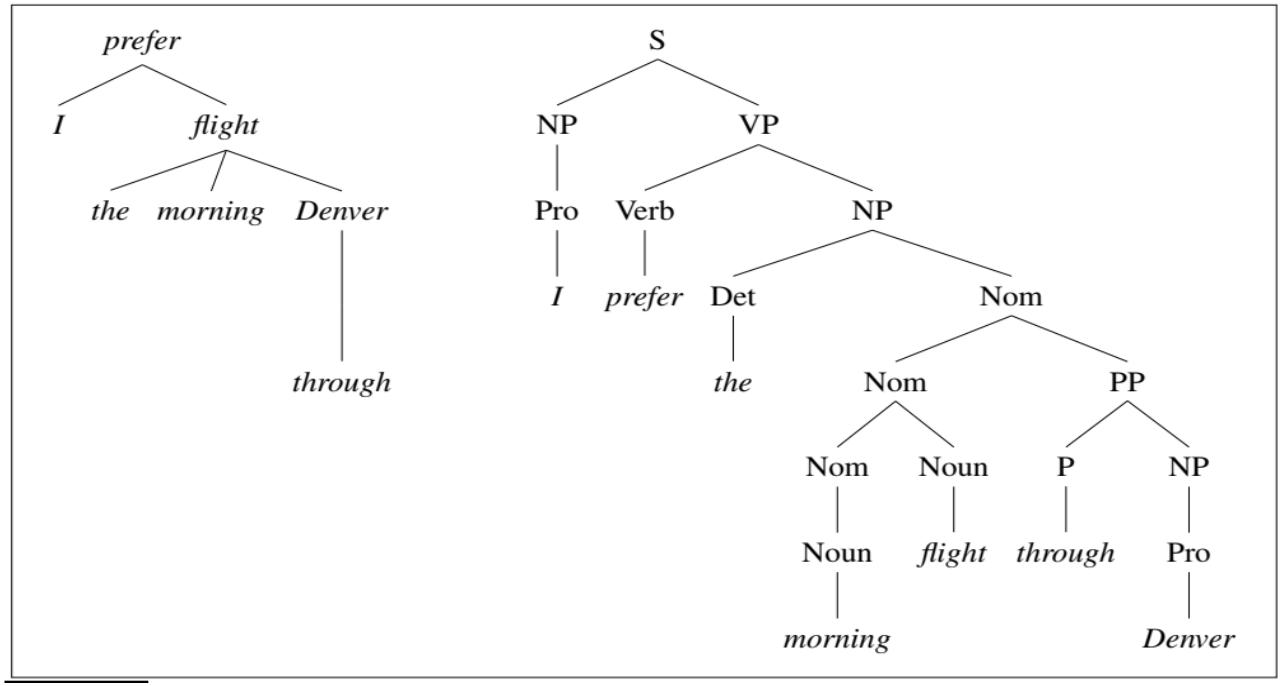
…CP needs additional rules

- *I saw the boy with a telescope*
  - *S → NP VP*
  - *VP → VBD NP PP*
- *With a telescope I saw the boy*
  - *S → NP VP*
  - *S → PP NP VP ???*

Impact of free word order on constituency parsing

- Constituency parse fundamentally uses adjacency information.
- Word order disturbs the adjacency
- Chomsky normal form demands that
  - The deduction should happen by linking together two adjacent entities.
- Example:
  - राम ने श्याम को देखा | (Ram ne Shyam ko dekha)
    - श्याम को देखा =VP
  - श्याम को राम ने देखा | (Shyam ko Ram ne dekha)
    - VP is discontinuous
    - Constituency parsing fails here
  - The agent and object are reversed in the above example
  - CP needs additional rules

Arguments are immediately linked



J & M, Chapter 15,
3rd Edition

Prefer: who prefers? "*I*"; what is preferred?: "*flight*".

On other hand, phrases are like *suitcases* that put all related things **at one place**: "The morning flight through Denver"

Subset of Dependency Relations: from Universal Dependency Project (Nivre et all 2016)

Clausal Argument Relations	**Description**
NSUBJ	Nominal subject
DOBJ	Direct object
IOBJ	Indirect object
CCOMP	Clausal complement
XCOMP	Open clausal complement
Nominal Modifier Relations	**Description**
NMOD	Nominal modifier
AMOD	Adjectival modifier
NUMMOD	Numeric modifier
APPOS	Appositional modifier
DET	Determiner
CASE	Prepositions, postpositions and other case markers
Other Notable Relations	**Description**
CONJ	Conjunct
CC	Coordinating conjunction

Examples to illustrate Dependency Relations

- NSUBJ, DOBJ, IOBJ- "*Ram gave a book to Shyam*"
  - Main Verb (MV): *gave*
  - NSUBJ: *Ram;* DOBJ: *book;* IOBJ: *Shyam*
- CCOMP, XCOMP: "I said that he should go", "I told him to go"
  - CCOMP: *said→go*
  - XCOMP: *told→go*

A note on CCOMP and XCOMP

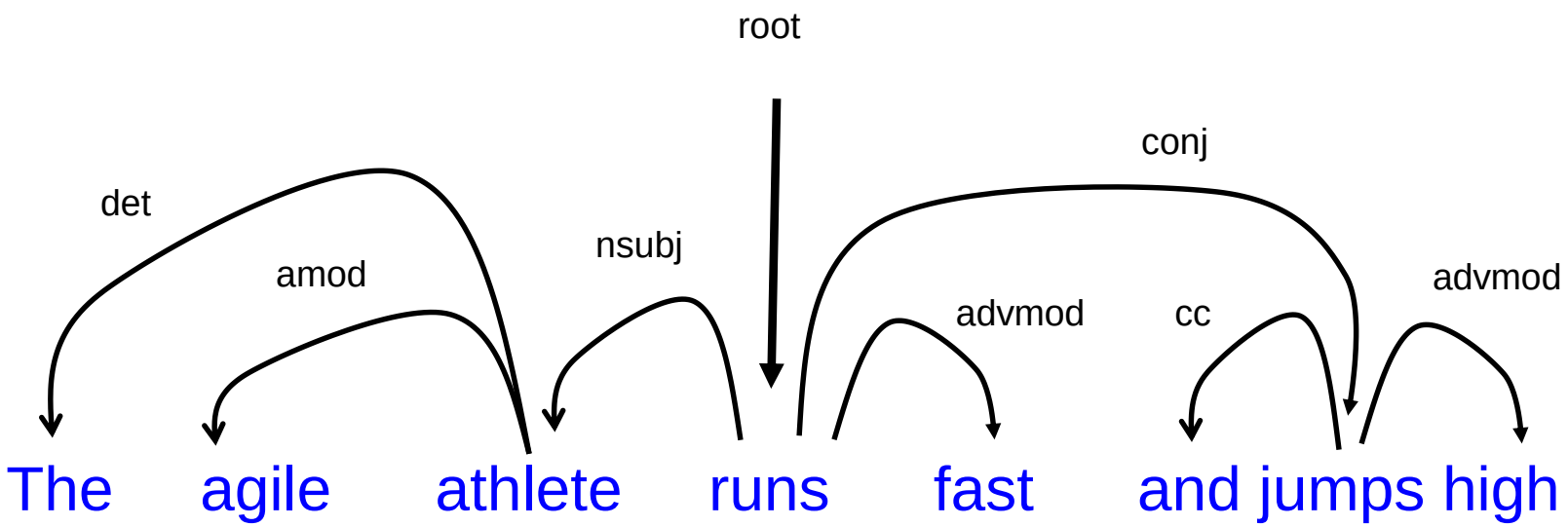
- CCOMP links the main verb with the finite verb
- XCOMP links main verb with an infinite verb
- Finite verb means: “takes GNPTAM marking”
- Infinite verb: remains in lemma form
- E.g. “told him to *go”:* ‘go’ will not change form (infinite form)
- “said he should go/be_going”: ‘go’ can change form

Illustration of DRs cntd.

- NMOD (nominal modifier), AMOD (adjective modifier), NUMMOD (numerical modifier), APPOS (appositional modifier)
  - NMOD: *The bungalow of the Director: Director ← bungalow*
  - AMOD: *The large bungalow: large ← bungalow*
  - NUMMOD: *Three cups: three ← cups*
  - APPOS: *covid19, the pandemic: covid19 ← pandemic*

Illustration of DRs cntd.

- DET (determiner), CASE (preposition, postposition and other case markers), CONJ (conjunct), CC (coordinating conjuct)
  - DET: *The bungalow: The←bungalow*
  - CASE: *The bungalow of Director: of→Director*
  - CONJ: *He is sincere and honest: sincere→honest*
  - CC: *He is sincere and honest: honest→and*


root
conj
det
amod
nsubj
advmod
cc
advmod
The
agile
athlete
runs
fast
and
jumps
high

Dependency Tree

- (1) There is a single designated root node that has no incoming arcs.
- (2) With the exception of the root node, each vertex has exactly one incoming arc.
- (3). There is a unique path from the root node to each vertex in *V*.