CS626: Speech, NLP and Web

Dependency Parsing, Projectivity, Algo, NER

Pushpak Bhattacharyya

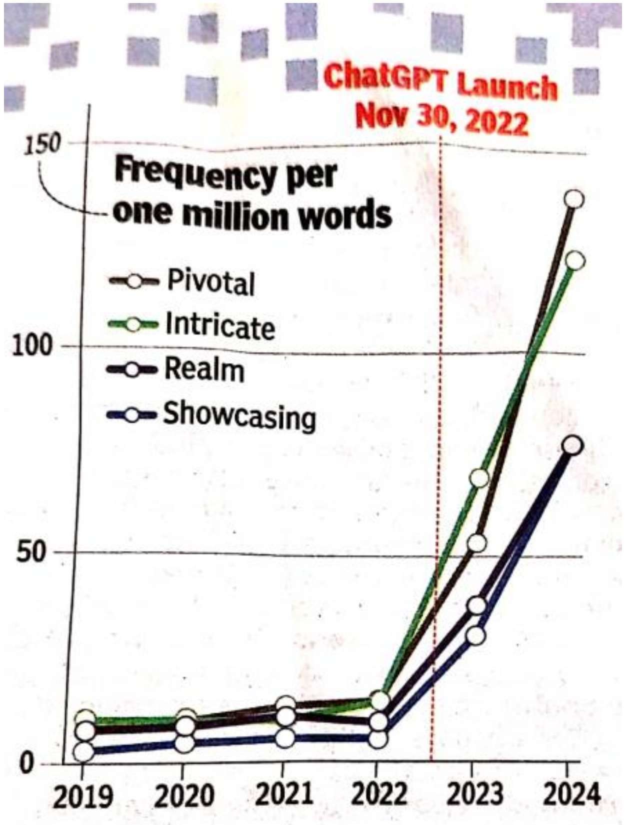
Computer Science and Engineering Department

IIT Bombay

Week 13 of 28th October, 2024

1-slide recap of week of 21st Oct

- ChatGPT give-aways: ‘delve’, ‘additionally’, ‘additionally’, ‘nevertheless’, ‘a testament to…’
- **Pramana**- means of acquiring knowledge: **Pratyaksha** (perception), **Anumana** (inference), **Upamana** (comparison



- **Sabda** (verbal testimony**), Arthapatti** (postulation), **Anupalabdhi** (non-perception)
- CLT, LoLN

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)

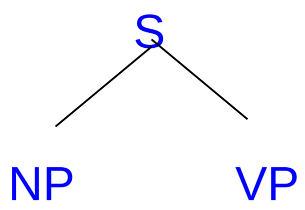
- Hindi has signals for semantic difference through case markers

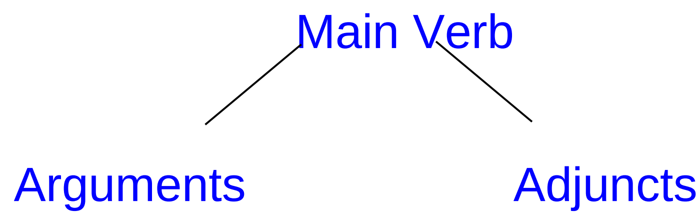
- English is more ambiguous

- But English sentences are more metaphorical

- Ambiguity needed for more colourful and complex linguistic constructs

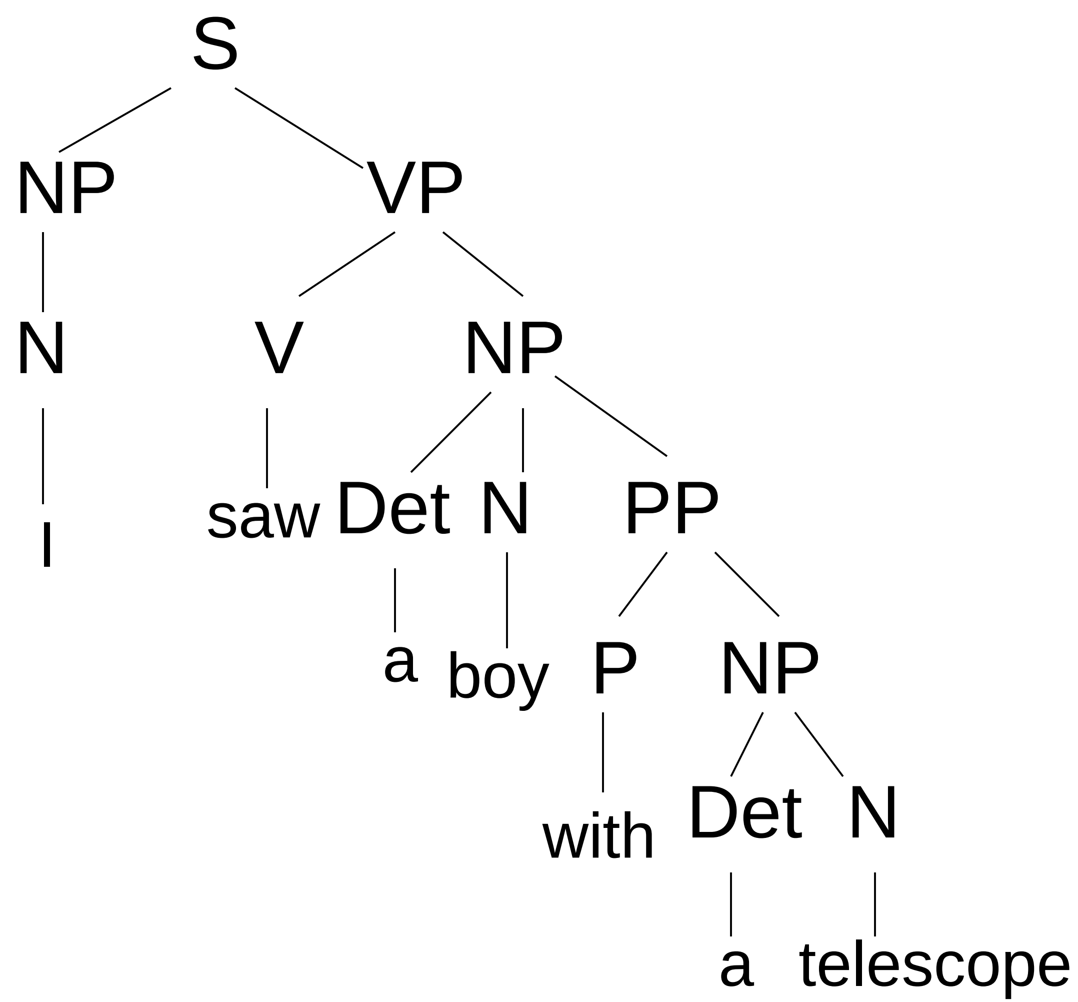
Two kinds of parse representations: Constituency Vs. Dependency



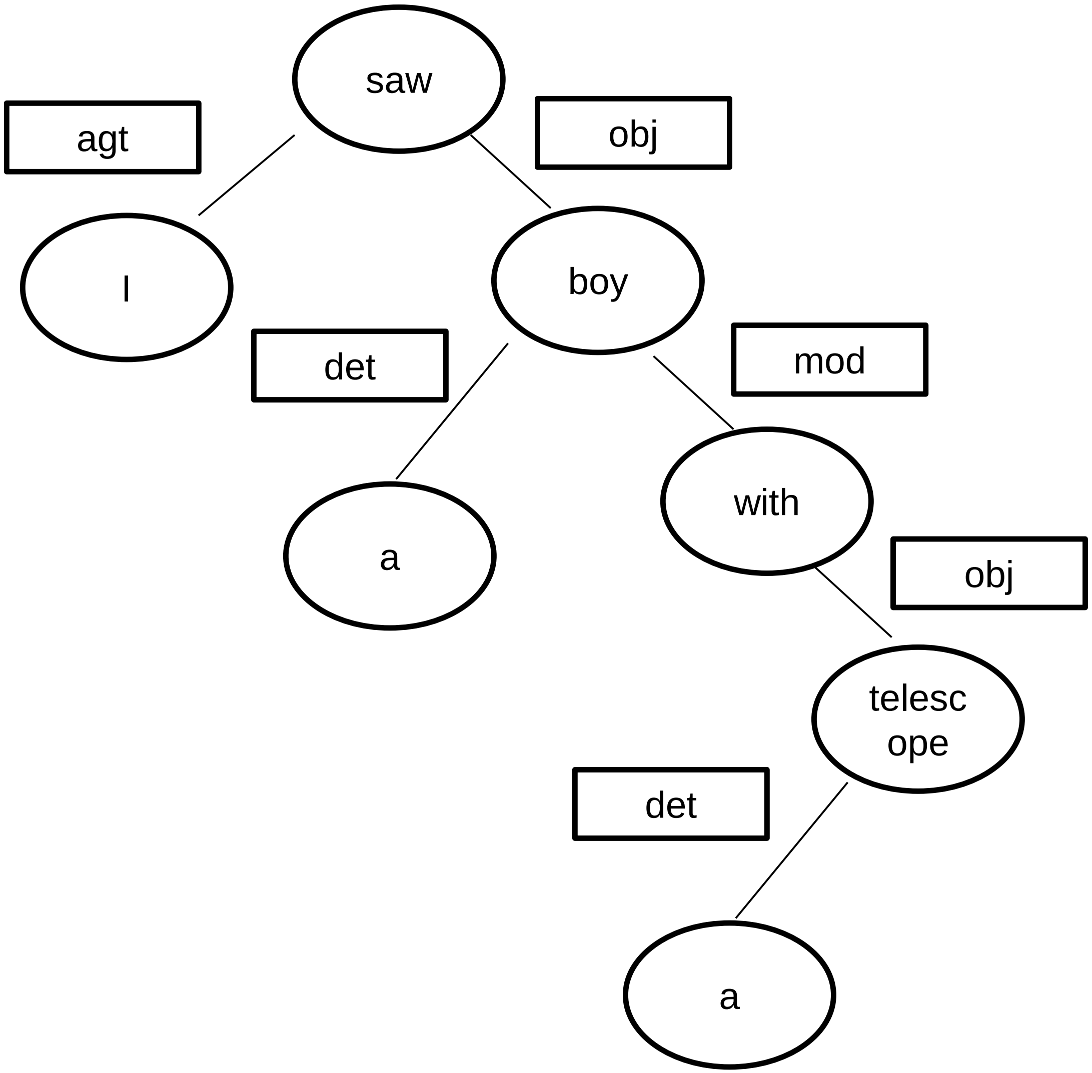


- Penn Constituency Treebank
  - http://www.cis.upenn.edu/~treebank/
- Prague Dependency Treebank
  - http://ufal.mff.cuni.cz/pdt2.0/

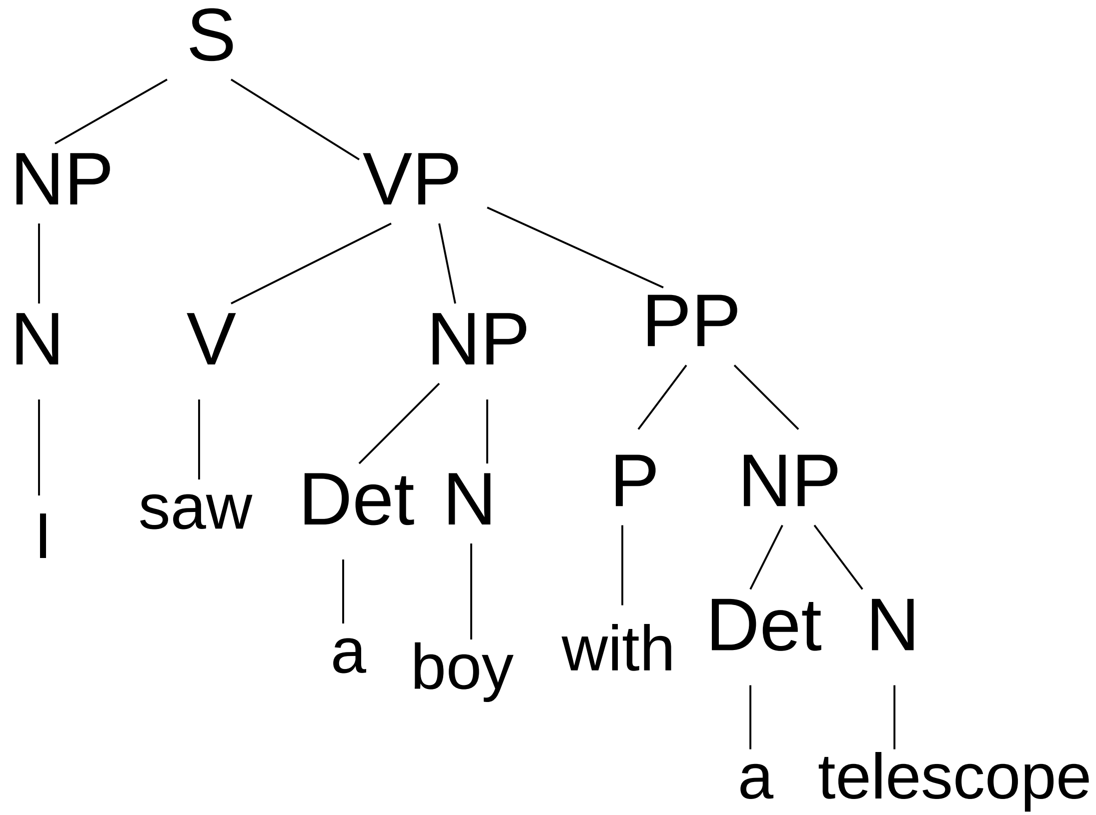
“I saw the boy with a telescope”: Constituency parse-1: *telescope with boy*



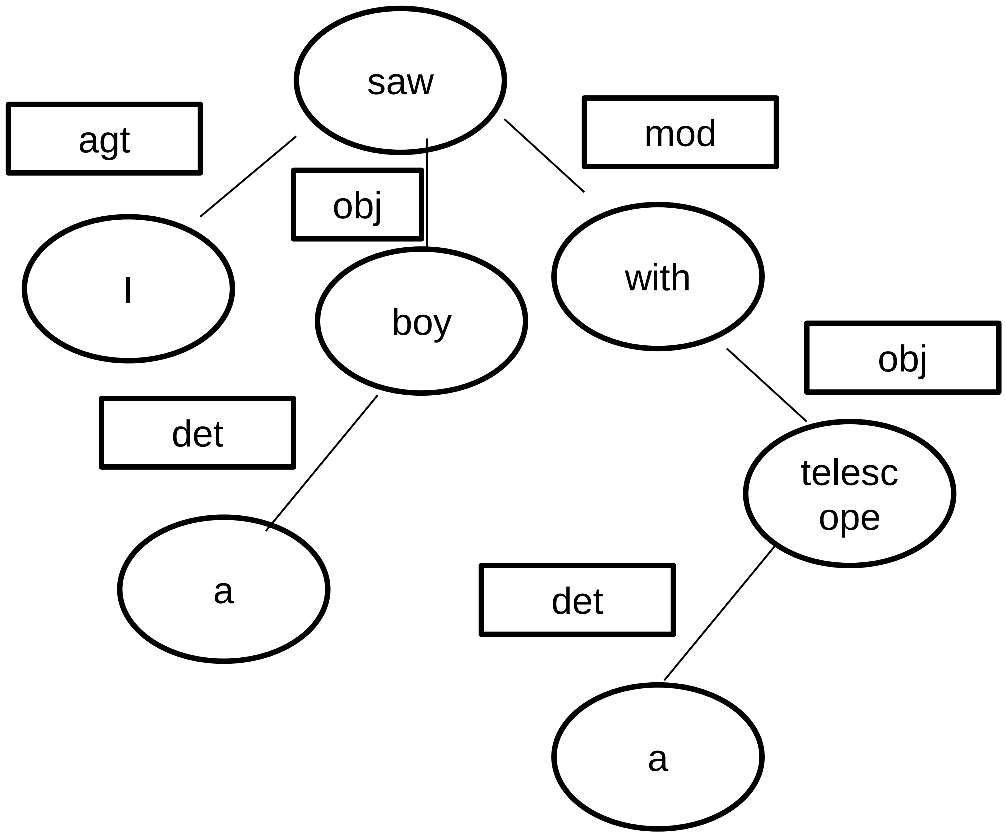
"I saw the boy with a telescope": Dependency Parse Tree-1



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!!

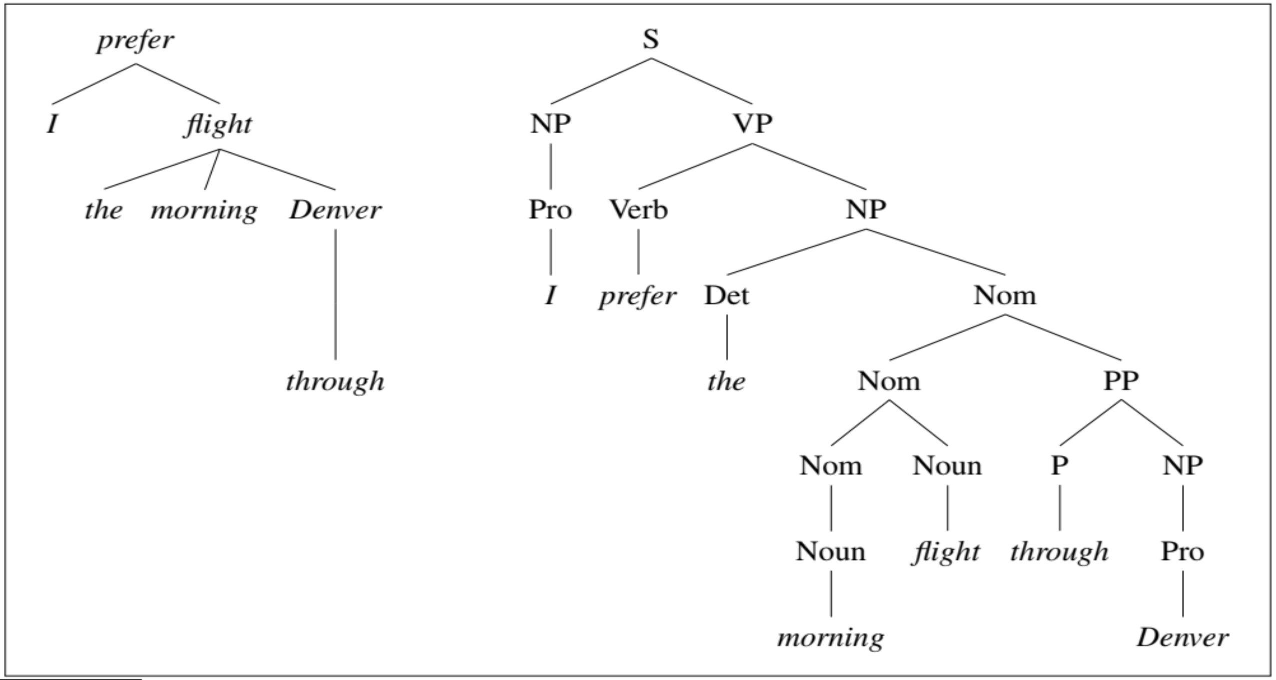
…CP needs additional rules

- *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

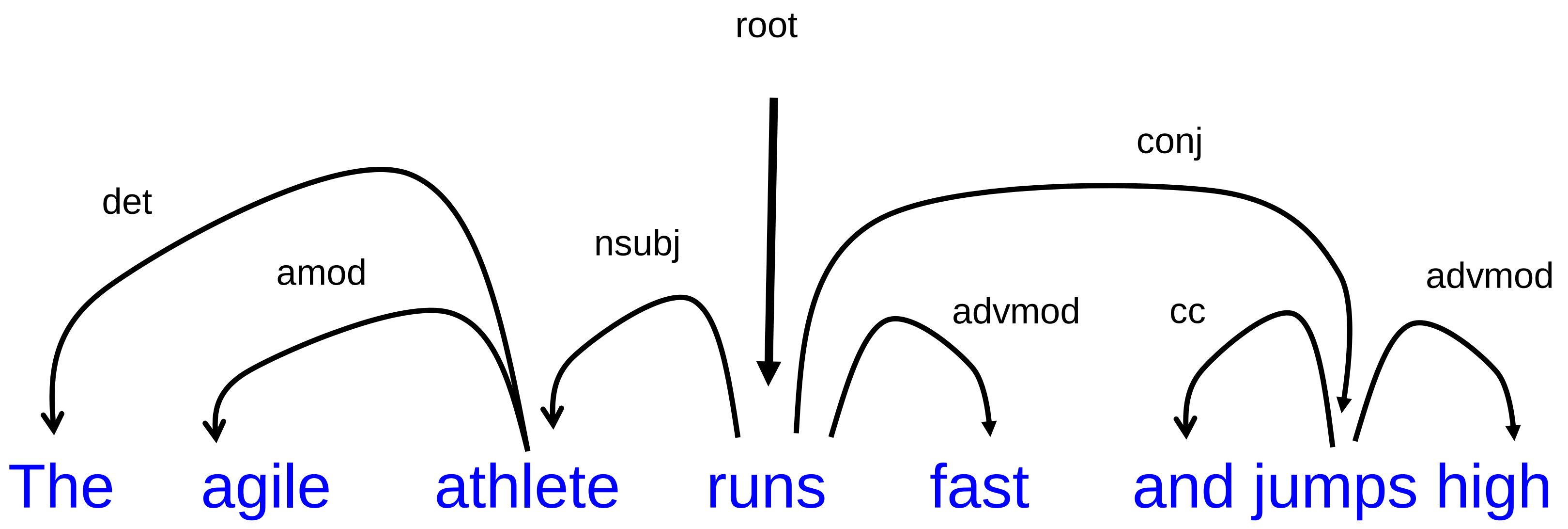
- CCOMP links the main verb with the finite verb
- XCOMP links main verb with an infinite verb
- Finite verb means: "takes GNPTAM marking"
- Infinite verb: remains in lemma form
- E.g. "told him to *go": '*go' will not change form (infinite form)
- "said he should go/be_going": 'go' can change form

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
nsubj
amod
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.

Projectivity

Definition

- An arc from a head to a modifier is said to be projective if there is a path from the head to every word that lies between the head and the modifier in the sentence
- A dependency tree is then said to be projective if all the arcs that make it up are projective
- ***Intuition***- the dependency graph can be drawn on a plane w/o crossing of arcs (condition: all arc must be on ONE side of the sentence: upper or lower, but not both)

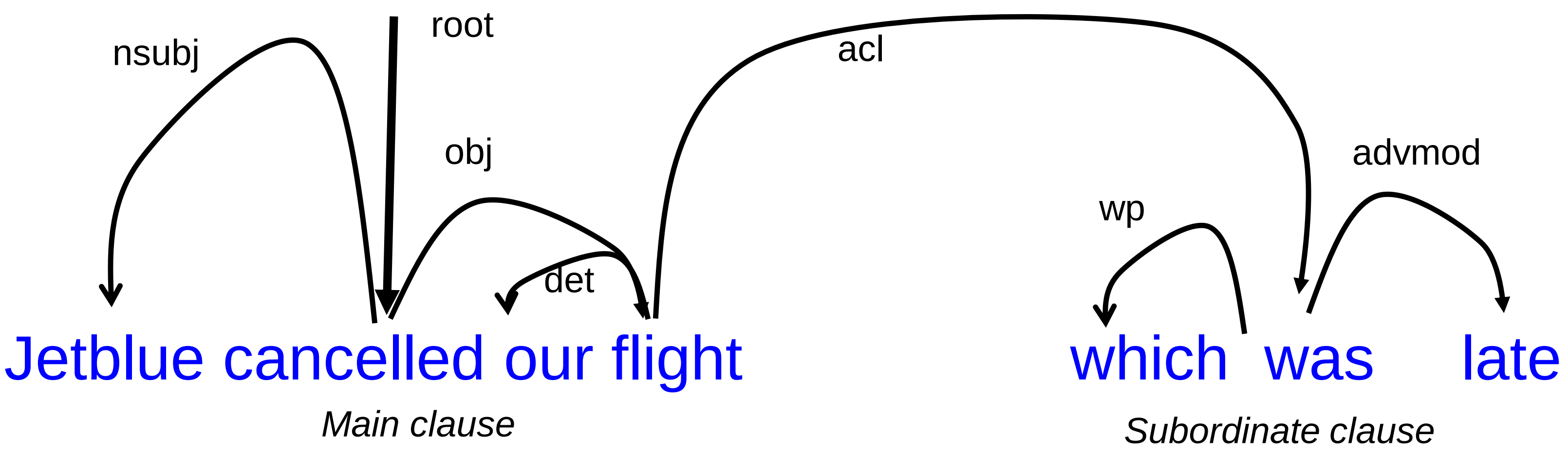
Conditions for projective dependency tree

1. All arcs are on ONE side (above or below) of the sentence.

2. There is NO crossing of arcs.

Equivalent: for EVERY Head-Modifier pair in the sentence, there is a path from the said Head to EVERY word in between the said Head and the said modifier.

Example (from J & M, 2019)



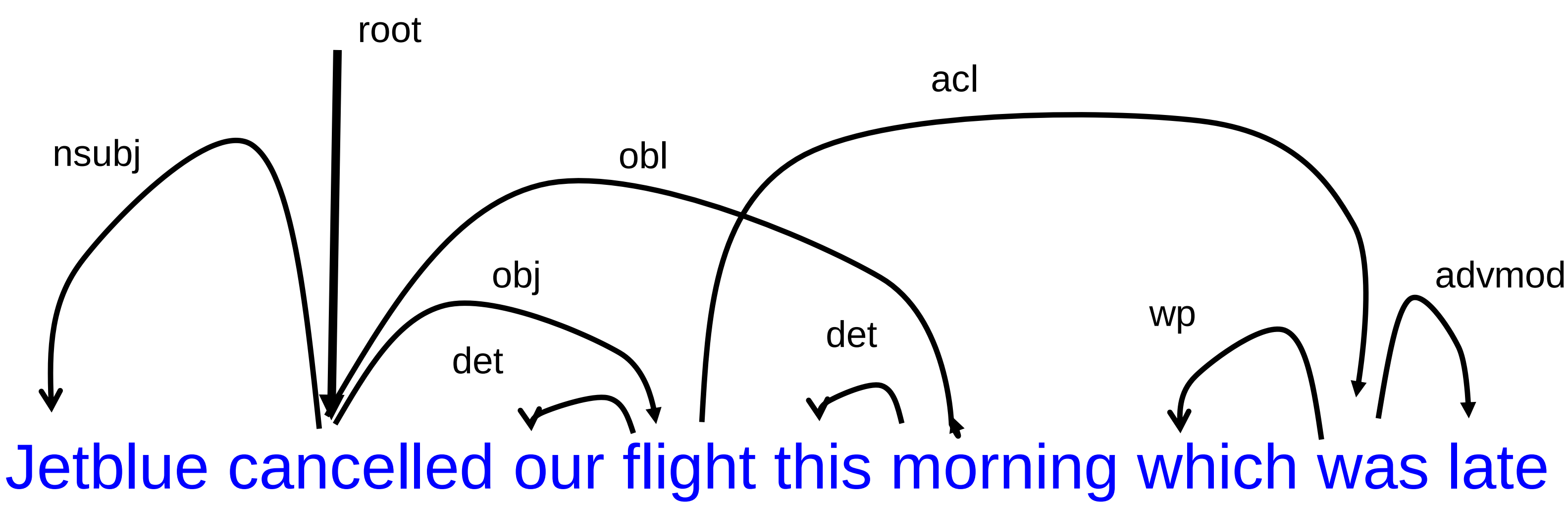
Uses Universal Dependency

wp- relative pronoun

acl- clausal modifier of noun

The head of the *acl* relation is the noun that is modified, and the dependent is the head of the clause that modifies the noun

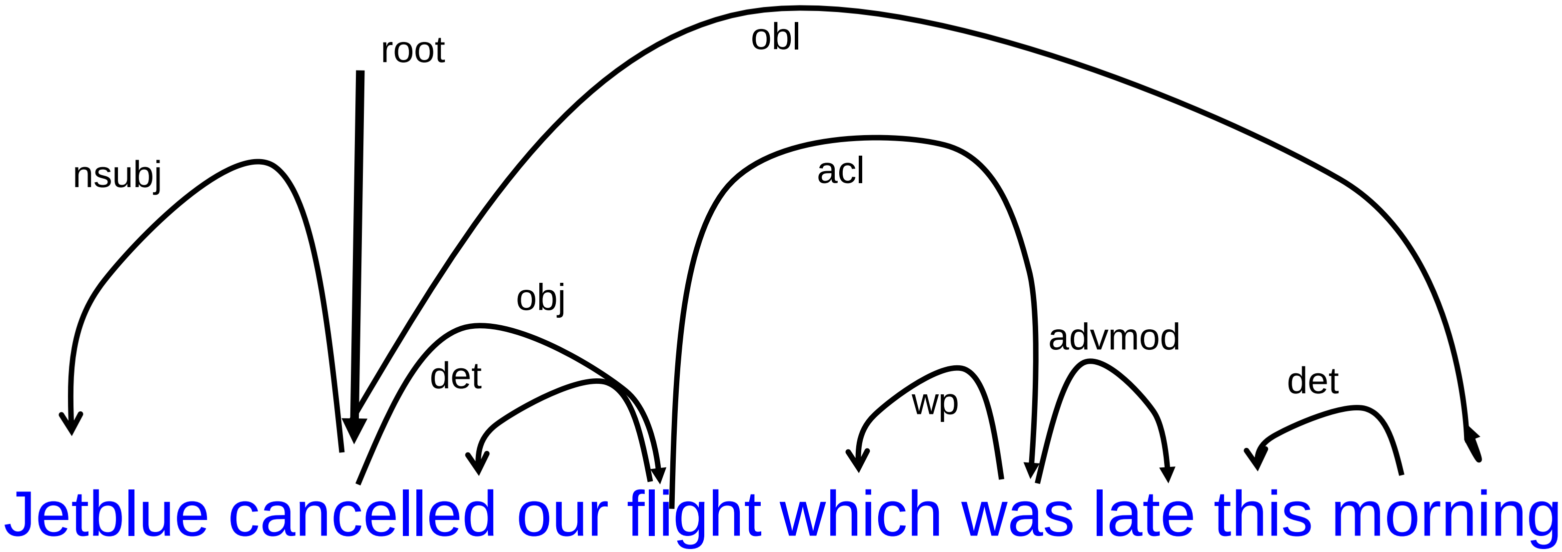
Example cntd. (from J & M, 2019): *insert "this morning"*

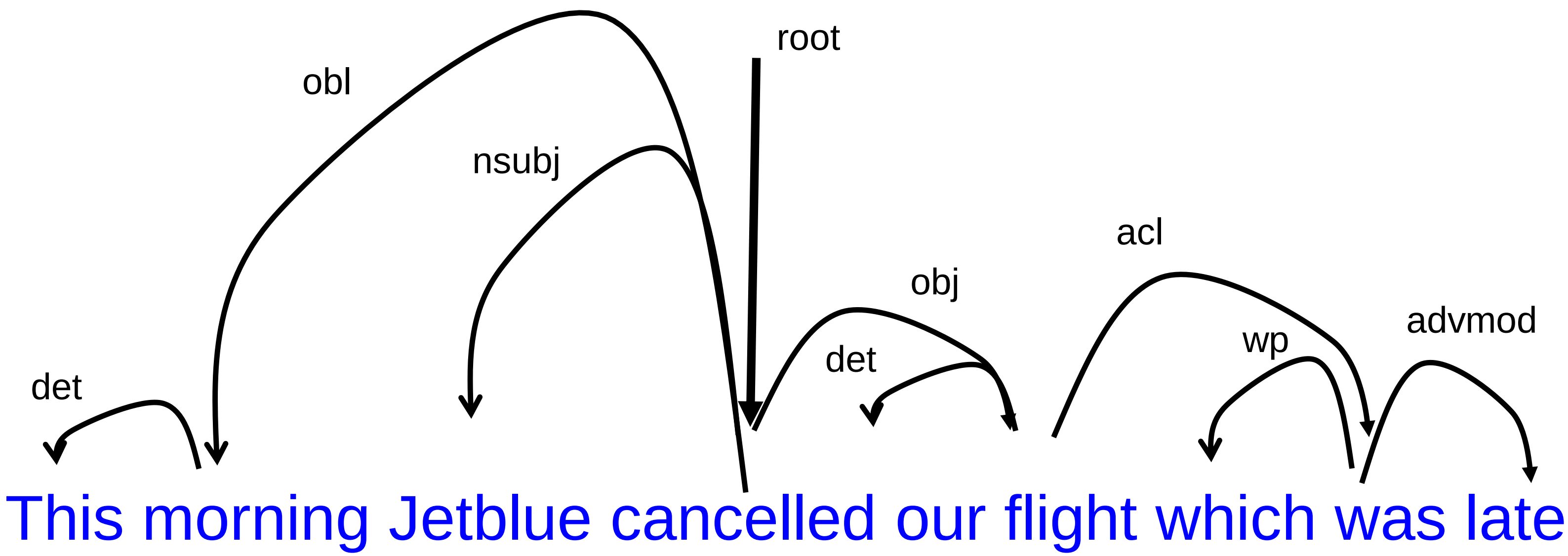


The obl relation is used for a nominal dependent (noun, pronoun, noun phrase) of a verb (or main predicate). This concerns different cases of prepositional groups, and more generally, it corresponds to an adverbial attaching to a verb, adjective or other adverb.

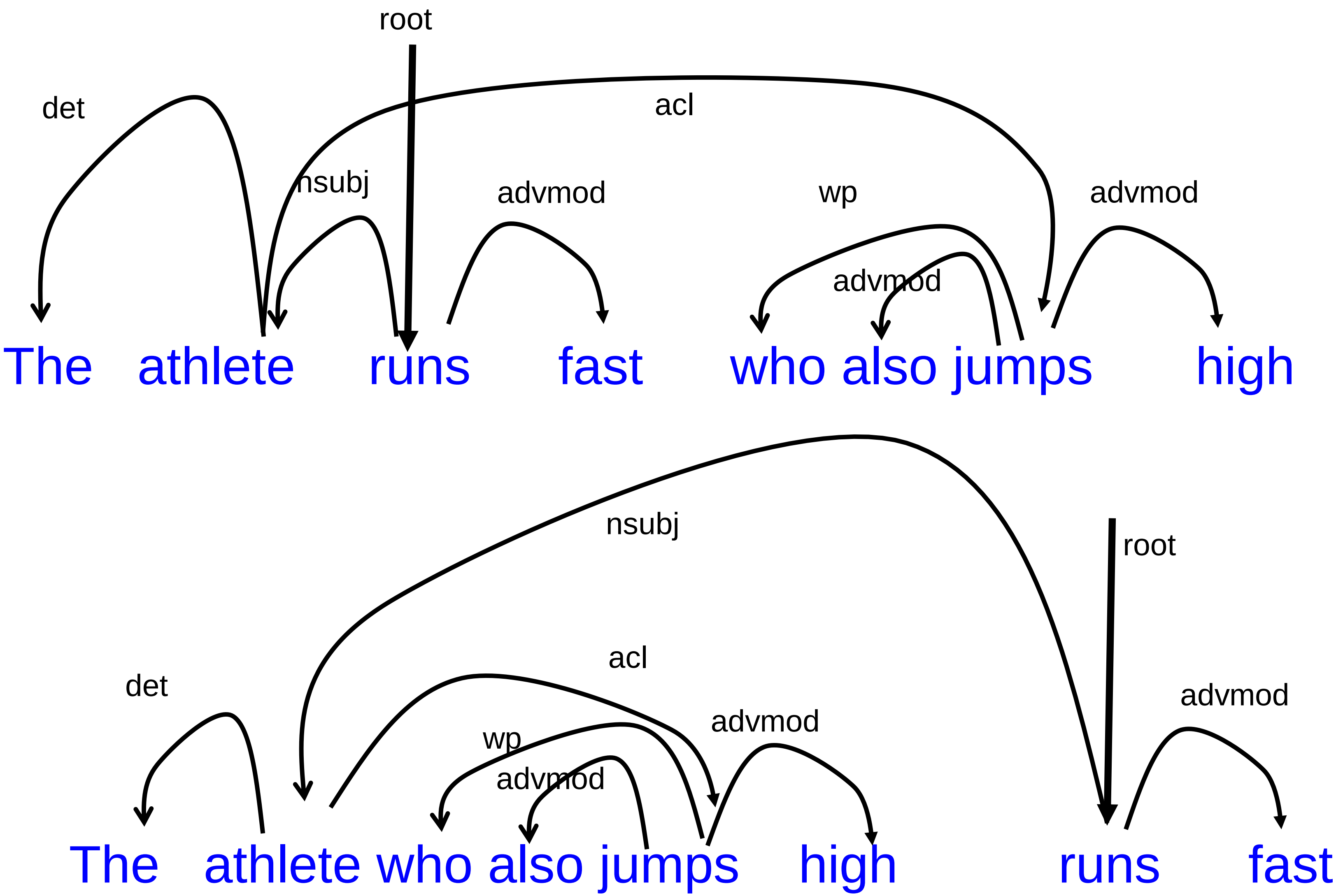
https://universaldependencies.org/

Move around “this morning”

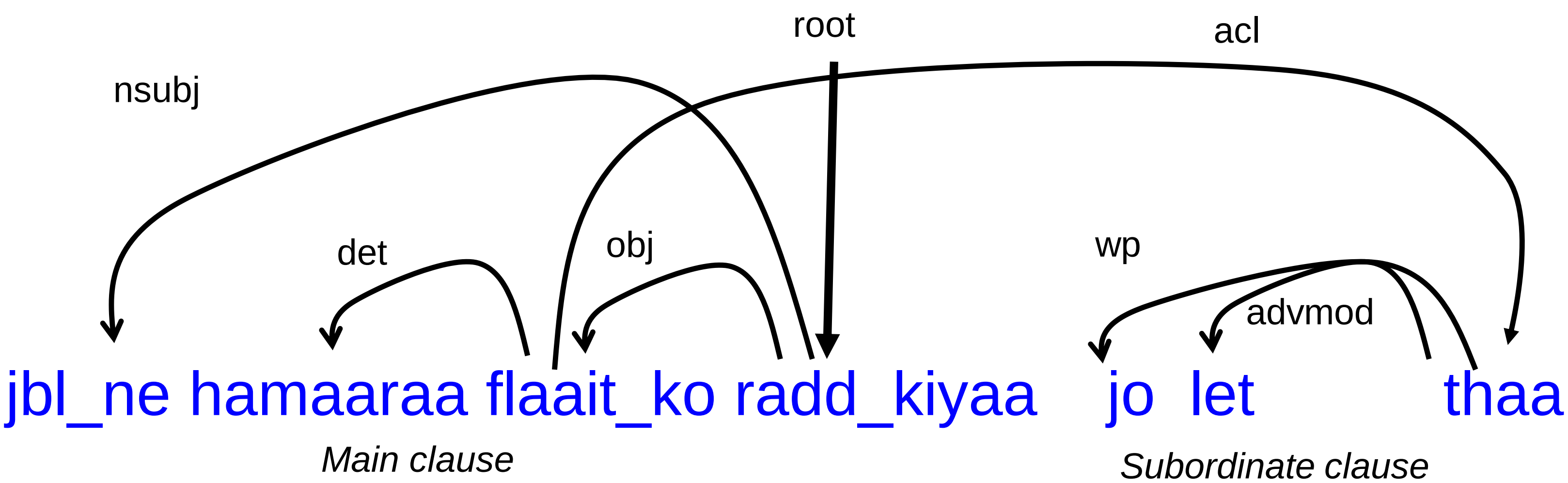




Another Example

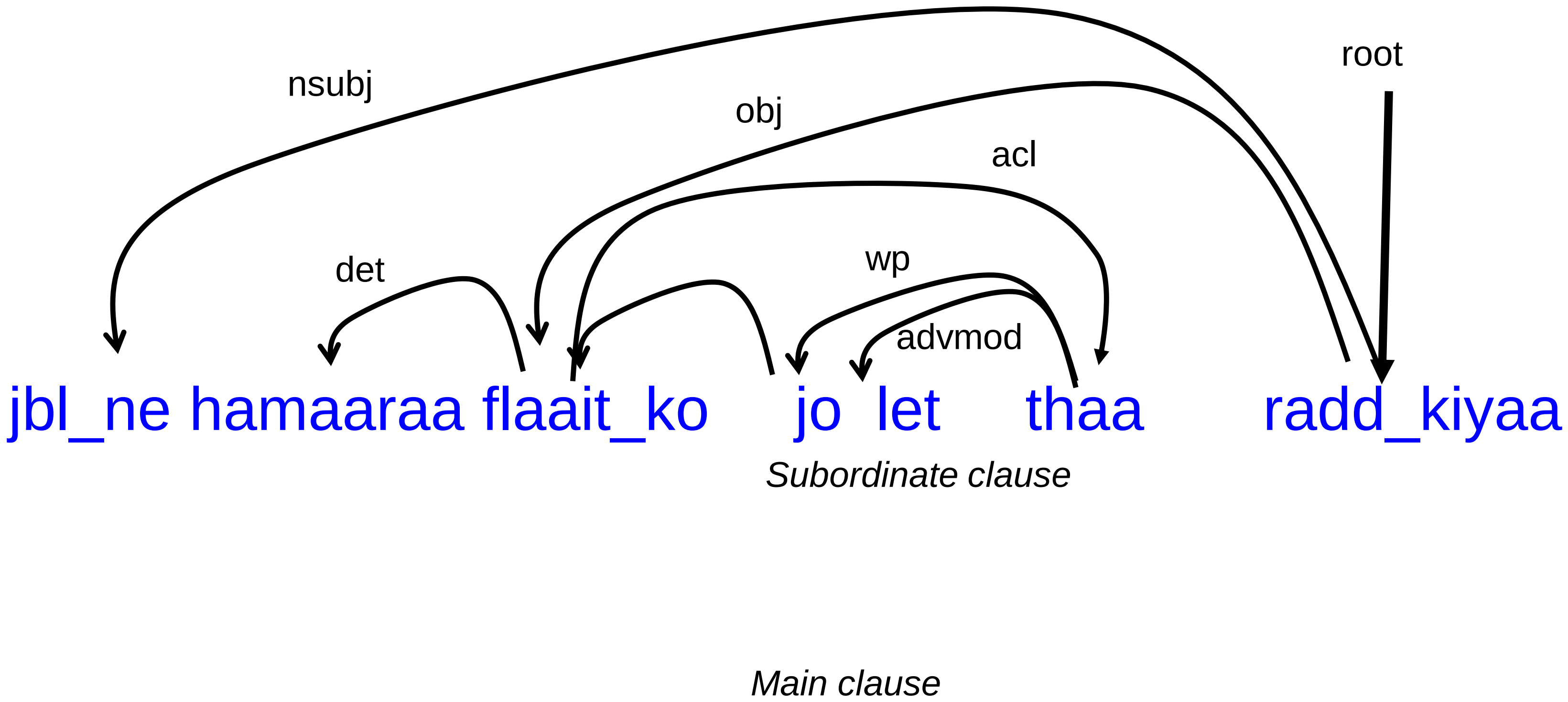


DP for other languages (Indian languages)



Uses Universal Dependency

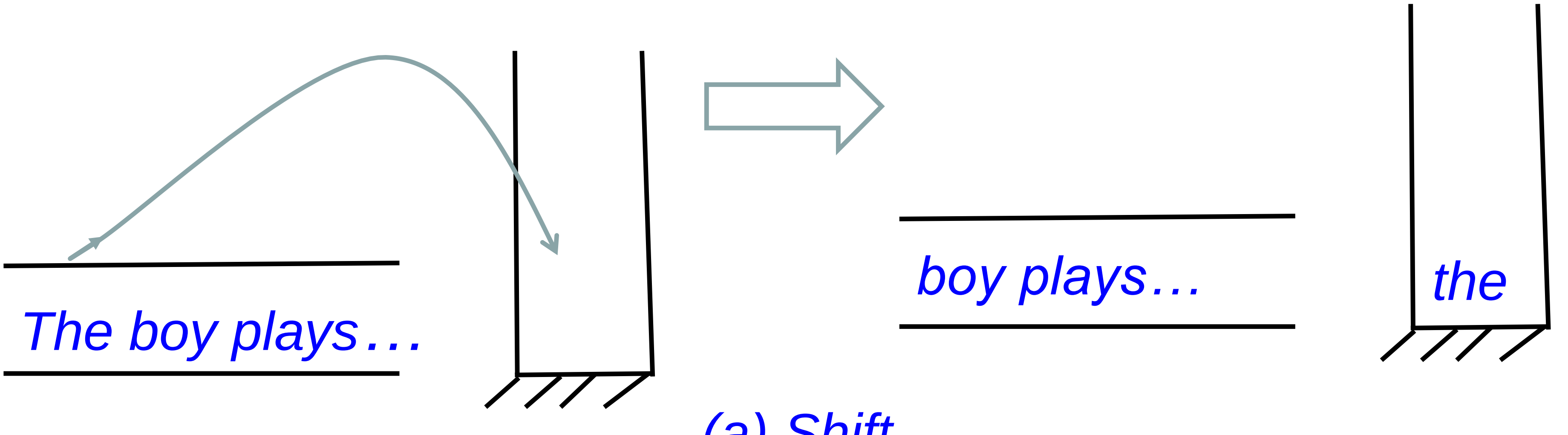
DP for other languages (Indian languages)



Uses Universal Dependency

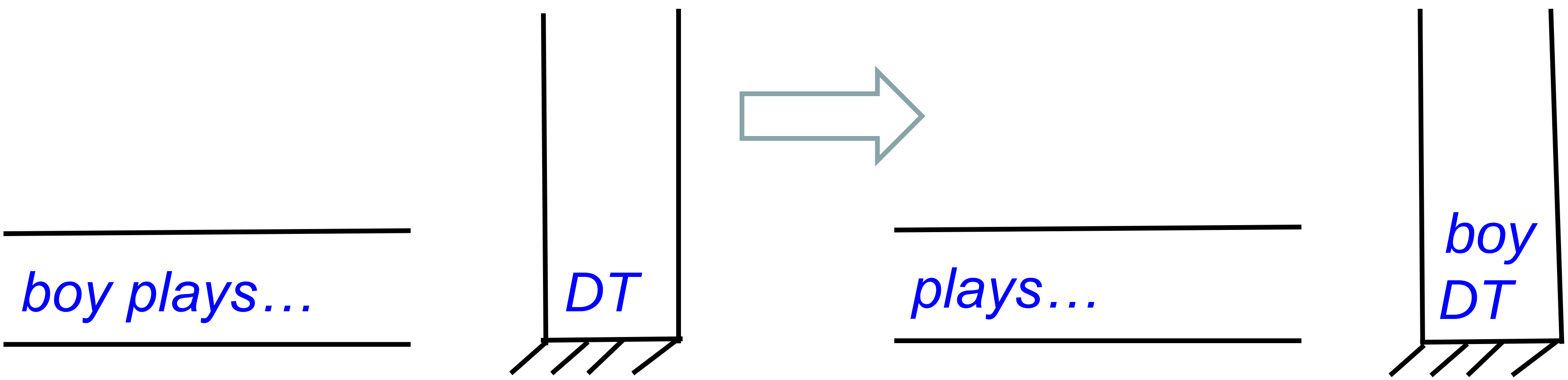
DP Algo

Shift Reduce (1/3)



(a) Shift

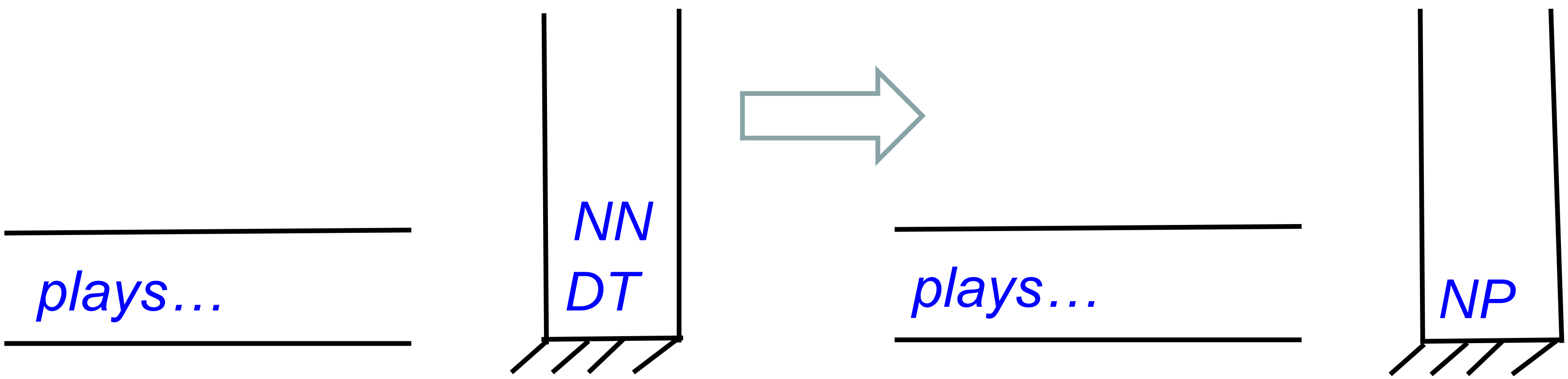
Shift Reduce (2/3)



(b) Reduce

(c) Shift

Shift Reduce (3/3)



(d) Reduce

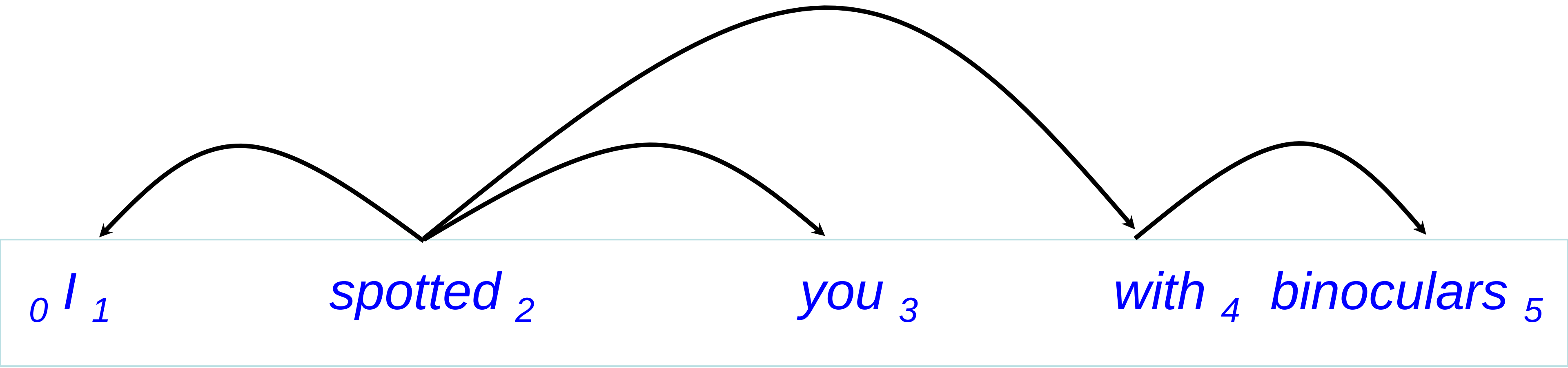
(e) Reduce

“I spotted you with binoculars”.

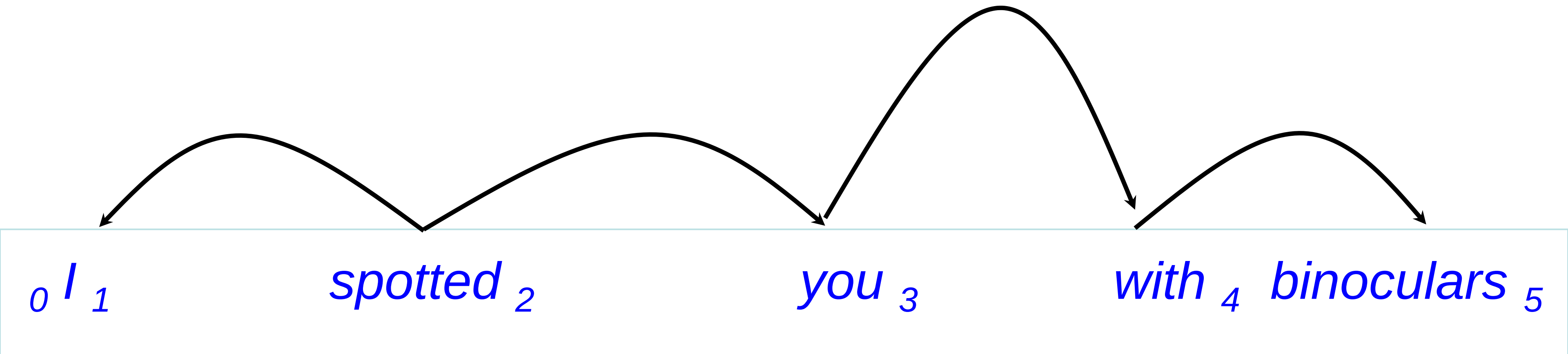
$_0$ I $_1$ spotted $_2$ you $_3$ with $_4$ binoculars $_5$

- Has two meanings
- I have the binoculars OR
- You have the binoculars

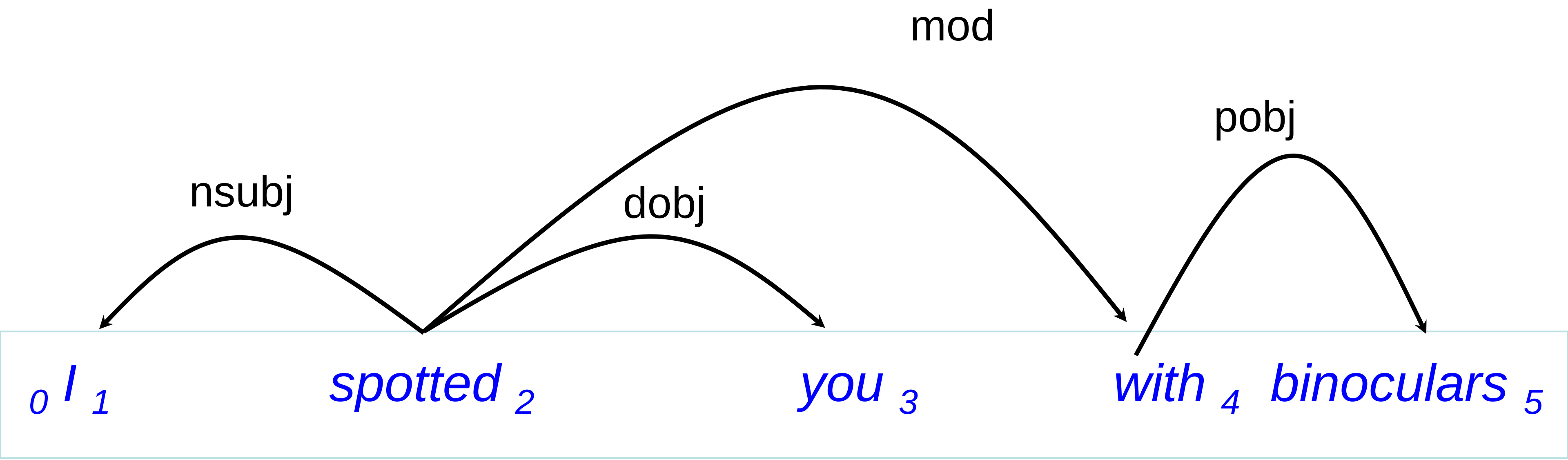
Unlabeled Dependency Tree-1



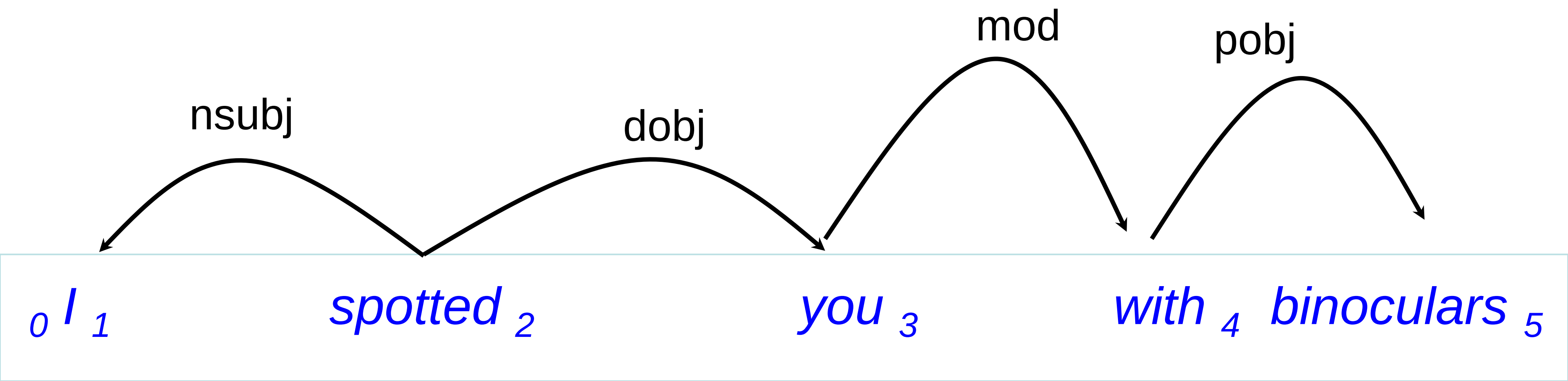
Unlabeled Dependency Tree-2



Labeled Dependency Tree-1



Labeled Dependency Tree-2



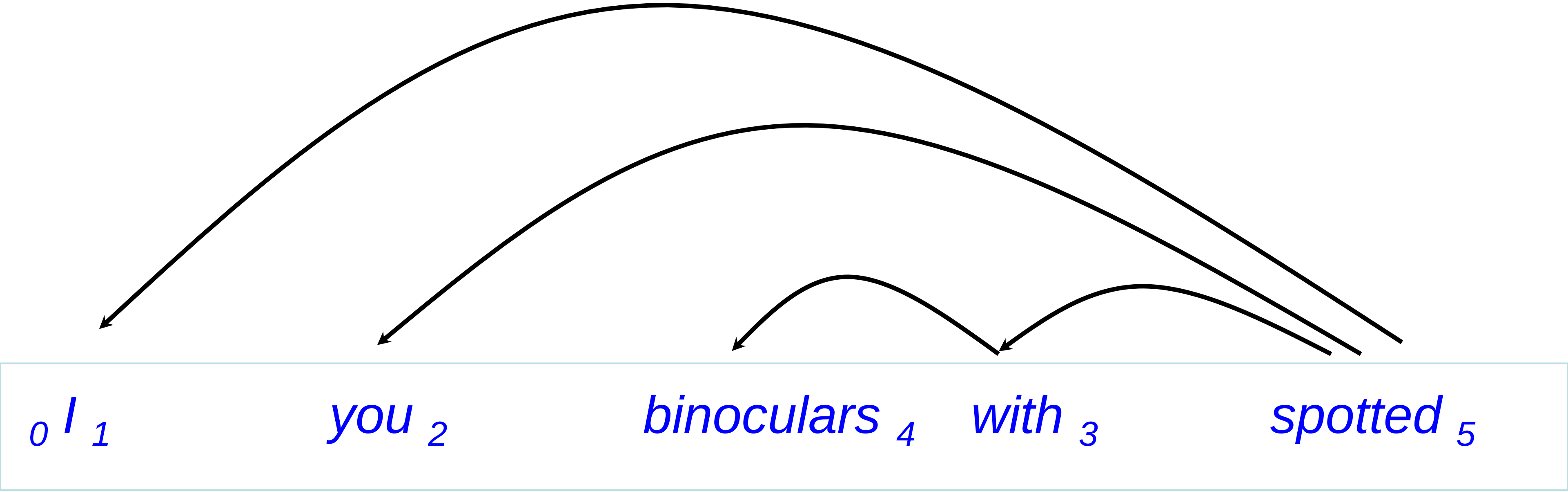
For SOV Syntax

“I you binoculars with spotted”

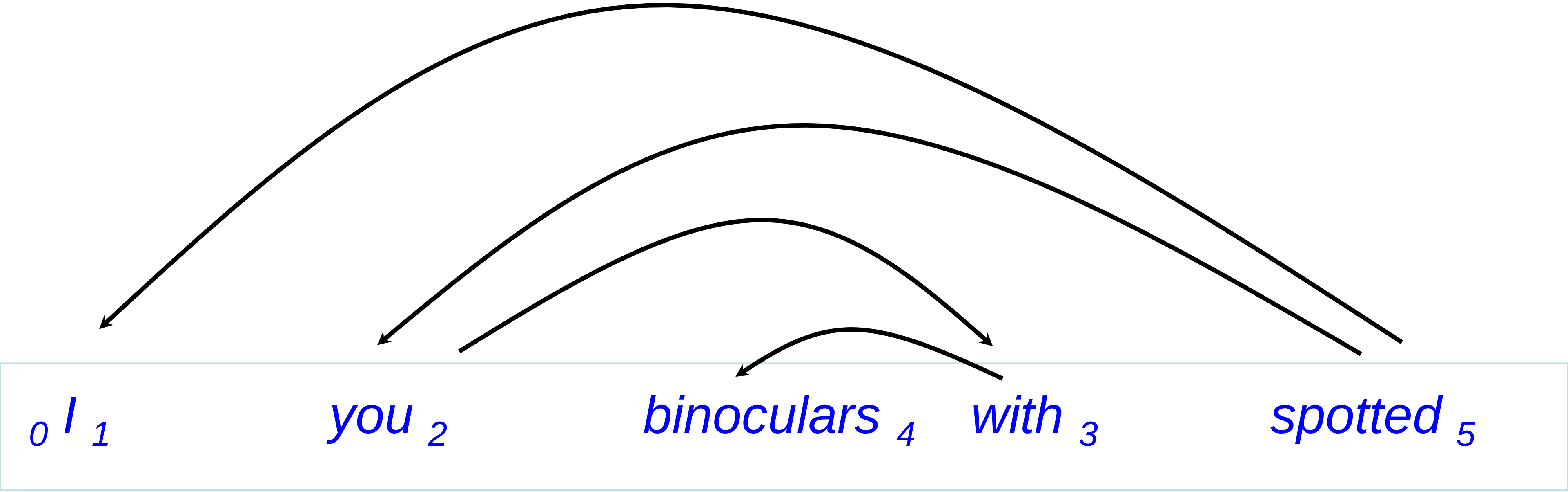
$_0$ *I* $_1$ *spotted* $_2$ *you* $_3$ *with* $_4$ *binoculars* $_5$

- Has two meanings
- I have the binoculars OR
- You have the binoculars

Unlabeled Dependency Tree-1



Unlabeled Dependency Tree-2



Dependency Parsing Example

- Transition based parsing
- Shift and Reduce

Parse-1

1. [root]	[I spotted you with binoculars]	shift	no-relation-added
2. [root I]	[spotted you with binoculars]	shift	no-relation-added
3. [root I spotted]	[you with binoculars]	left-arc	I→spotted
4. [root spotted]	[you with binoculars]	shift	no-relation-added
5. [root spotted you]	[with binoculars]	right-arc	spotted→you
6. [root spotted]	[with binoculars]	shift	no-relation-added
7. [root spotted with]	[binoculars]	shift	no-relation added
8. [root spotted with binoculars]	[]	right-arc	with→binoculars
9. [root spotted with]	[]	right-arc	spotted→with
10. [root spotted]	[]	right-arc	root→spotted
11. [root]	[]	parsing ends	

Parse-2

1. [root]	[I spotted you with binoculars]	shift	no-relation-added
2. [root I]	[spotted you with binoculars]	shift	no-relation-added
3. [root I spotted]	[you with binoculars]	left-arc	I←spotted
4. [root spotted]	[you with binoculars]	shift	no-relation-added
5. [root spotted you]	[with binoculars]	shift	no-relation-added
6. [root spotted you with]	[binoculars]	shift	no-relation-added
7. [root spotted you with binoculars]	[]	right-arc	with→binoculars
8. [root spotted you with]	[]	right-arc	you→with
9. [root spotted you]	[]	right-arc	spotted→you
10. [root spotted]	[]	right-arc	root→spotted
11. [root]	[]	parsing ends	

Essence of DP

- Cannot pop a *head* out of the stack if any of its dependents remains on the stack

- The above works if the sentence's semantics is consistent with projectivity

Recall: CYK Algo

- $S \rightarrow NP\ VP$ 1.0
- $NP \rightarrow DT\ NN$ 0.5
- $NP \rightarrow NNS$ 0.3
- $NP \rightarrow NP\ PP$ 0.2
- $PP \rightarrow P\ NP$ 1.0
- $VP \rightarrow VP\ PP$ 0.6
- $VP \rightarrow VBD\ NP$ 0.4

- $DT \rightarrow$ the 1.0
- $NN \rightarrow$ gunman 0.5
- $NN \rightarrow$ building 0.5
- $VBD \rightarrow$ sprayed 1.0
- $NNS \rightarrow$ bullets 1.0

CYK based DP: 1/7

The gunman sprayed the building with bullets (unlabeled DT)

Head→ Modifier	**The**	**gunman**	**sprayed**	**the**	**buildi ng**	**with**	**bullets**
The		L					
gunman							
sprayed							
the							
building							
with							
bullets							

CYK based DP: 2/7

The gunman sprayed the building with bullets (unlabeled DT)

Head→ Modifier	The	gunman	**sprayed**	**the**	**buildi ng**	**with**	**bullets**
The		L					
gunman			L				
sprayed							
the							
building							
with							
bullets							

CYK based DP: 3/7

The gunman sprayed the building with bullets (unlabeled DT)

Head→ Modifier	The	gunman	**sprayed**	the	**buildi ng**	**with**	**bullets**
The		L					
gunman			L				
sprayed							
the					L		
building							
with							
bullets							

CYK based DP: 4/7

The gunman sprayed the building with bullets (unlabeled DT)

Head→ Modifier	The	gunman	**sprayed**	the	buildin g	**with**	**bullets**
The		L					
gunman			L				
sprayed					R		
the					L		
building							
with							
bullets							

CYK based DP: 5/7

The gunman sprayed the building with bullets (unlabeled DT)

Head→ Modifier	The	gunman	**sprayed**	the	buildin g	**with**	bullets
The		L					
gunman			L				
sprayed					R		
the					L		
building							
with							R
bullets							

CYK based DP: 6/7

The gunman sprayed the building with bullets (unlabeled DT)

Head→ Modifier	The	gunman	**sprayed**	the	buildin g	**with**	bullets
The		L					
gunman			L				
sprayed					R		
the					L		
building							
with							R
bullets							

CYK based DP (7/7)

The gunman sprayed the building with bullets (unlabeled DT)

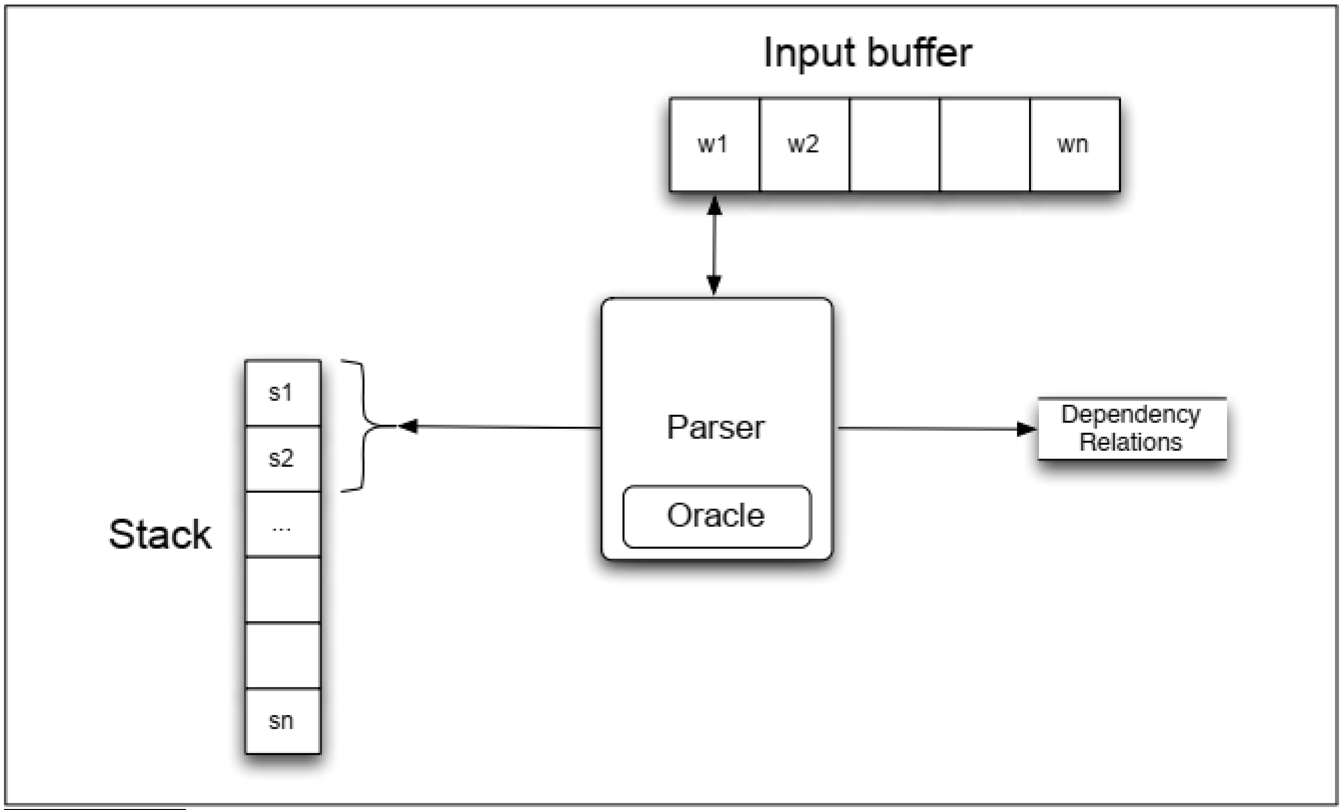
Head→ Modifier	The	gunman	**sprayed**	the	buildin g	with	bullets
The		L					
gunman			L				
sprayed					R	R	
the					L		
building							
with							R
bullets							

Data Driven Algorithms for Dependency Tree Construction

Two Data Driven Approaches

- Transition-based
  - State machine for mapping a sentence to its dependency graph
  - Inducing a model for predicting the next transition, given the current state and the transition history so far.
- Graph-based
  - Induce a model for assigning scores to the candidate dependency graphs for a sentence
  - Find the maximum-scoring dependency Tree
  - Maximum spanning tree (MST) parsing

Basic Transition Based DP



Examines top two elements of the stack and selects an action based on consulting an oracle that examines the current configuration.

Speech and Language Processing, Jurafksy & Martin, Ch-15, 2019

Example: transition based

Step	Stack	Word List	Action	Relation Added
0	[root]	[book, me, the, morning, flight]	SHIFT	
1	[root, book]	[me, the, morning, flight]	SHIFT	
2	[root, book, me]	[the, morning, flight]	RIGHTARC	(book $\rightarrow$ me)
3	[root, book]	[the, morning, flight]	SHIFT	
4	[root, book, the]	[morning, flight]	SHIFT	
5	[root, book, the, morning]	[flight]	SHIFT	
6	[root, book, the, morning, flight]	[]	LEFTARC	(morning $\leftarrow$ flight)
7	[root, book, the, flight]	[]	LEFTARC	(the $\leftarrow$ flight)
8	[root, book, flight]	[]	RIGHTARC	(book $\rightarrow$ flight)
9	[root, book]	[]	RIGHTARC	(root $\rightarrow$ book)
10	[root]	[]	Done	

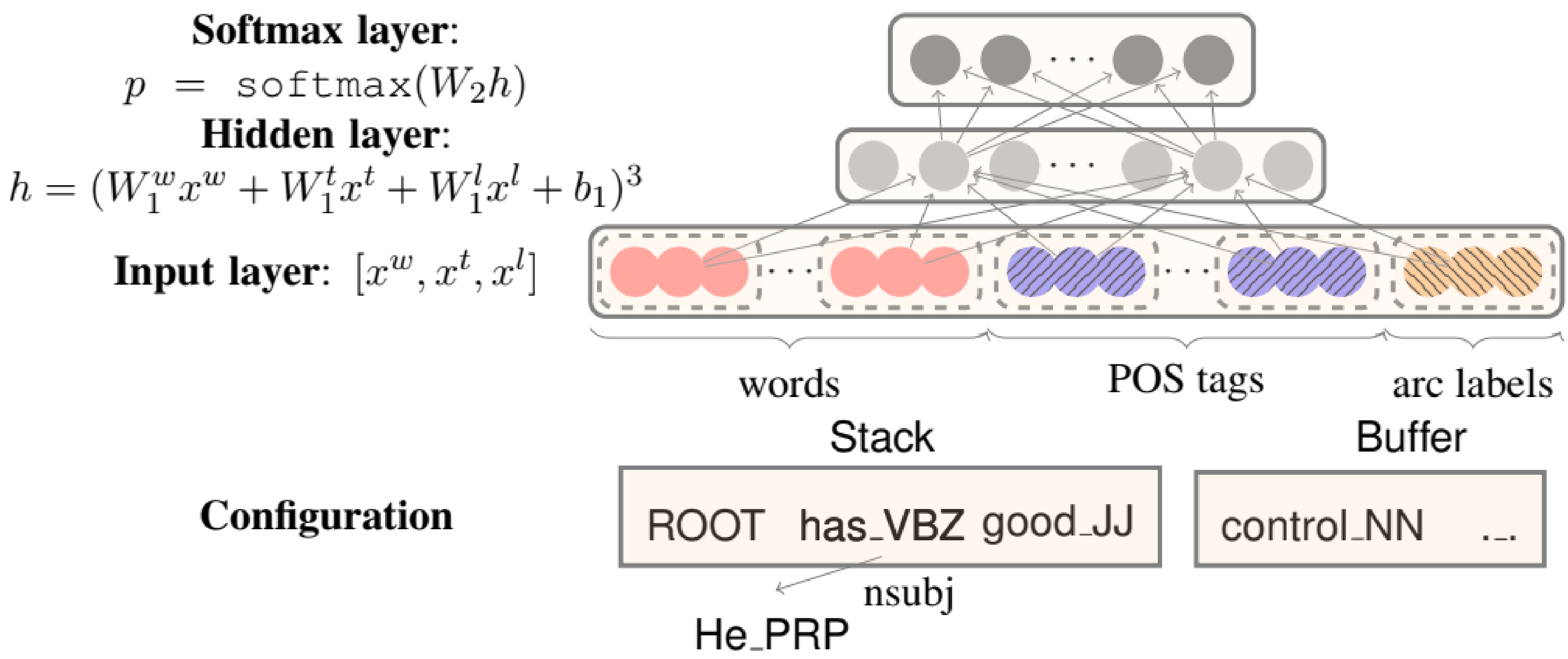
Trace of a transition-based parse

Speech and Language Processing, Jurafksy & Martin, Ch-15, 2019

ORACLE

- Decides whether to “shift” or to “reduce”
- If “reduce”, whether to set up “rightarc” or to set up “leftarc”
- Can be controlled by DEPENDENCY GRAMMAR RULES
- Or, by rules learnt from data
- Or, by a neural network

A neural transition based parser
(chen and Manning 2014)

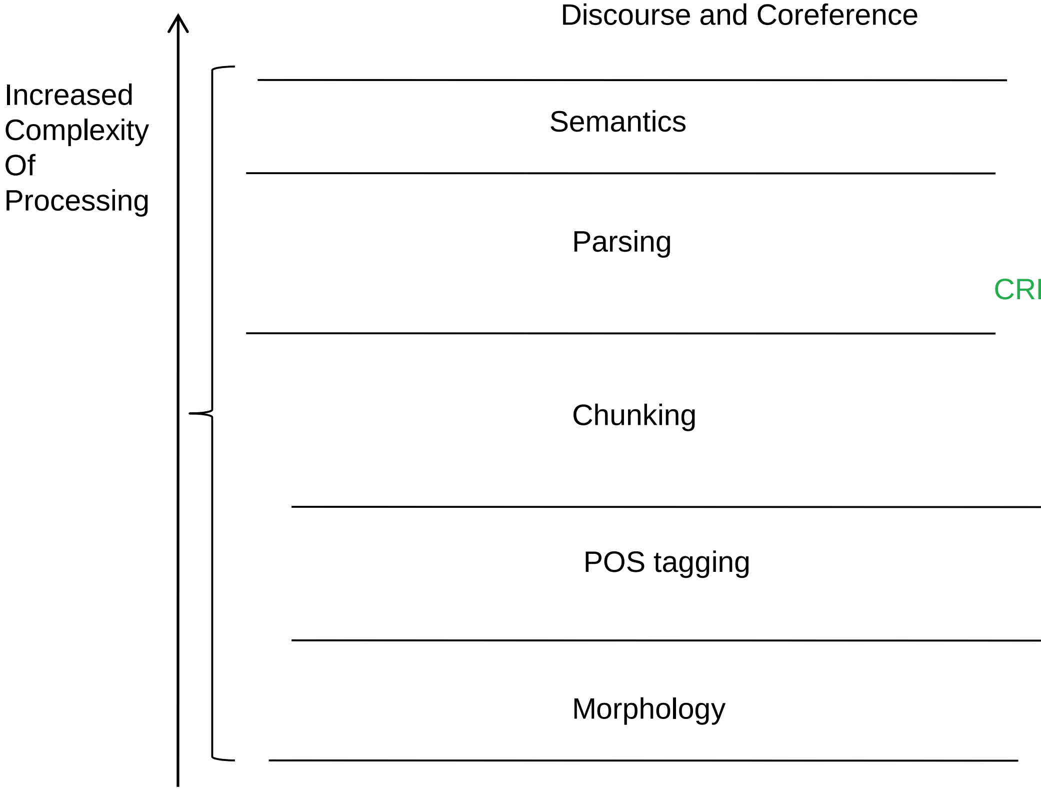


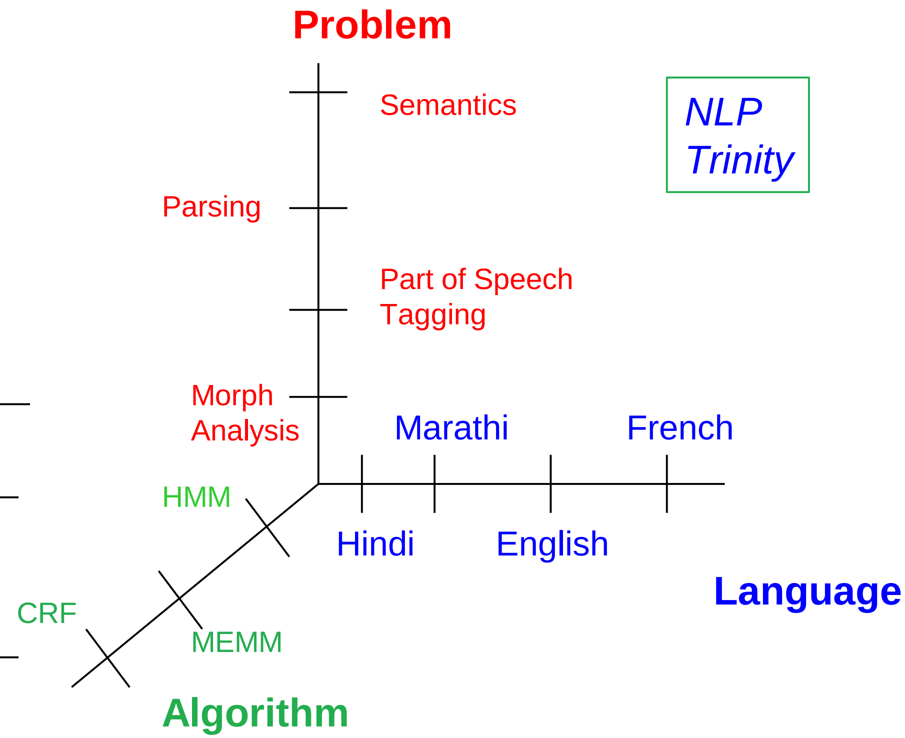
How to get the Neural Net Trained and how to get the training data

- The training data will come from Dependency Trees
- For example, given “the morning flight” and “the←flight”, “morning←flight”, it is possible to ***simulate*** the parser generate training data (next slide)
- Such trees come from ***Prague Dependency Tree Bank***

Named Entity Recognition

NLP Perspective





Inherent resilience of the structure called LANGUAGE

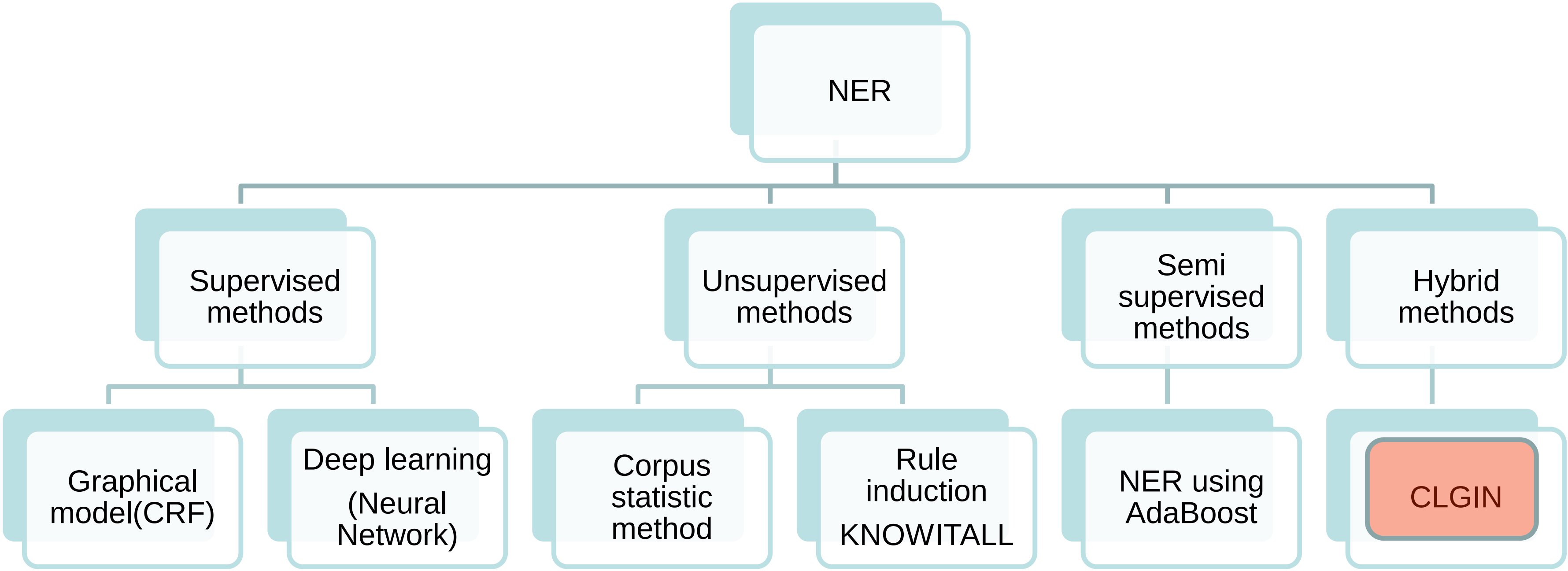
- Example: *Apple increased its laptop production*
- *I know that Apple increased its laptop production*
- *I know apples are costly (fruit apple): plural 's' disambiguates*
- *An apple a day keeps the doctor away: article disambiguates*
- *I bought an Apple for my accounting work: capital 'A' disambiguates*
- *Vulnerability: He has an apple: looseness in capitalization makes disambiguation impossible*

Multilingual Named Entity Recognition

- If a named entity is recognized in one language, it can be added to lexicon (gazetteer list) and used for processing tasks in other languages

- *Extract Once, Use Many times*

Approaches


NER
Supervised methods
Unsupervised methods
Semi supervised methods
Hybrid methods
Graphical model(CRF)
Deep learning (Neural Network)
Corpus statistic method
Rule induction KNOWITALL
NER using AdaBoost
CLGIN

Background:
Information Extraction

Definition: Information Extraction

- To extract information that fits pre-defined schemas or templates
- IE Definition
  - Entity: an object of interest such as a person or organization
  - Attribute: A property of an entity such as name, type
  - Relation: A relationship that holds between two or more entities such as Position of Person in a Company

Information Extraction

- **Note** the difference between Named Entity Identification and Named Entity Recognition

- **Named Entity Identification** is a binary classification problem which classifies whether a given token is a named entity or not

- **Named Entity Recognition** involves detection and categorization of named entities

Goal of IE

As a task: **Filling slots from text- unstructured→structured**

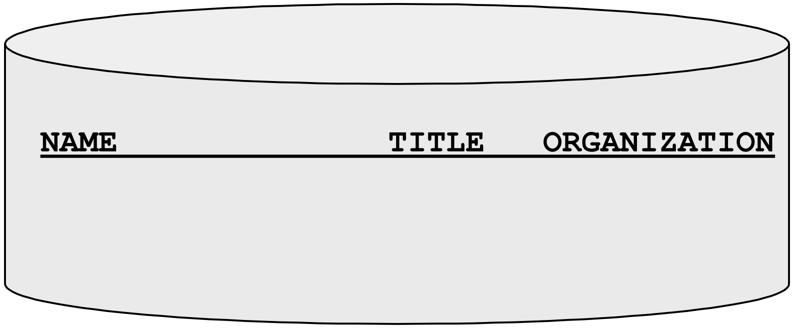
October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access.“

Richard Stallman, founder of the Free Software Foundation, countered saying…



Courtesy of William W. Cohen

Unstructured→structured

As a task:

October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access.“

Richard Stallman, founder of the Free Software Foundation, countered saying…

IE

NAME	TITLE	ORGANIZATION
Bill Gates	CEO	Microsoft
Bill Veghte	VP	Microsoft
Richard Stallman	founder	Free Soft..

Courtesy of William W. Cohen

Named Entity Identification (NEI) and Named Entity Recognition (NER)

Definition

NERC – Named Entity Recognition and Classification (NERC) involves identification of proper names in texts, and classification into a set of pre-defined categories of interest as:

- Person names (names of people)
- Organization names (companies, government organizations, committees, etc.)
- Location names (cities, countries etc)
- Miscellaneous names (Date, time, number, percentage, monetary expressions, number expressions and measurement expressions)

NEI and NER

- **Note** the difference between Named Entity Identification and Named Entity Recognition
- **Named Entity Identification** is a binary classification problem which classifies whether a given token is a named entity or not
- **Named Entity Recognition** involves detection and categorization of named entities

Challenge of NEI and NER

- Variation of NEs – e.g. *Prof. Manning, Chris Manning, Dr. Chris Manning*
- Ambiguity of NE types:
  - *Washington (location vs. person)*
  - *May (person vs. month)*
  - *Ford (person vs organization)*
  - *1945 (date vs. time)*
- Ambiguity with common words, e.g. “Kabita“

  - Name of person vs. poem

More complex problems in NER

- Issues of style, structure, domain, genre etc.
- Punctuation, spelling, spacing, formatting, ... all have an impact:

Dept. of Computing and Maths

Manchester Metropolitan University

Manchester

United Kingdom

Many to Many Relationship

Rows are labelled with **entity ids** and columns contain **names**

	NAMES			
ENTITIES	E353	Manning	Prof_Manning	Chris_Manning
	E201	Oxygen	O_2	
	E356	Kolkata	Calcullta	West Bengal Capital
	E404	IIT	Indian_Institute_of_Technology	Indian_Institute_of_Tech.

E.g.: E353 to represent the specific person entity ‘Chris Manning'

What would be skyline NER performance

- Human performance is considered to be the ultimate goal to be reached by the m/c.; measured by IAA (Inter Annotator Agreement)

- IAA gives the skyline

- E.g., WSD IAA is around 85% which translates to skyline performance of percentage in the vicinity of 80s

Applications

- Intelligent document access
  - News
  - Scientific articles, e.g, MEDLINE abstracts
- Information retrieval and extraction
  - Augmenting a query NE information→ more refined information extraction
- Machine translation
  - Translation vs. transliteration
  - *Indira Gandhi Open Unievsity→ Indiraa Gandhi mukta vishwavidyaalay*
- Automatic Summarization
  - Paragraphs containing more NEs are most likely to be included into the summary

Applications

- Question-Answering Systems
  - NEs are important to retrieve the answers of particular questions (*Who is the PM of India/Which country is Modi PM of?)*
- Speech Related Tasks
  - NER is important for identifying the number format, telephone number and date format
  - In speech rhythm- necessary to provide a short break after the name of person
  - Solving Out Of Vocabulary words is important in speech recognition

Corpora, Annotation

- MUC-6 and MUC-7 corpora - English
- CONLL shared task corpora
  - http://cnts.uia.ac.be/conll2003/ner/: NEs in English and German
  - http://cnts.uia.ac.be/conll2002/ner/: NEs in Spanish and Dutch
- ACE – English - http://www.ldc.upenn.edu/Projects/ACE/
- TIDES surprise language exercise (NEs in Hindi)
- NERSSEAL shared task- NEs in Bengali, Hindi, Telugu, Oriya and Urdu (http://ltrc.iiit.ac.in/ner-ssea-08/index.cgi?topic=5)

Corpora, Annotation

- Biomedical and Biochemical corpora
  - BioNLP-04 shared task
  - BioCreative shared tasks
  - AiMed

Tag set

Text is tagged (1/2)

Identifying and classifying elements in text into predefined categories such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc.

Text is tagged (2/2)

Example:

Jim bought 300 shares of ABC Corp. in 2006.

<ENAMEX TYPE="PERSON"> Jim </ENAMEX> bought <NUMEX

TYPE = "QUANTITY"> 300 </NUMEX> shares of <ENAMEX TYPE =

"ORGANIZATION"> ABC Corp. </ENAMEX> in <TIMEX TYPE = "

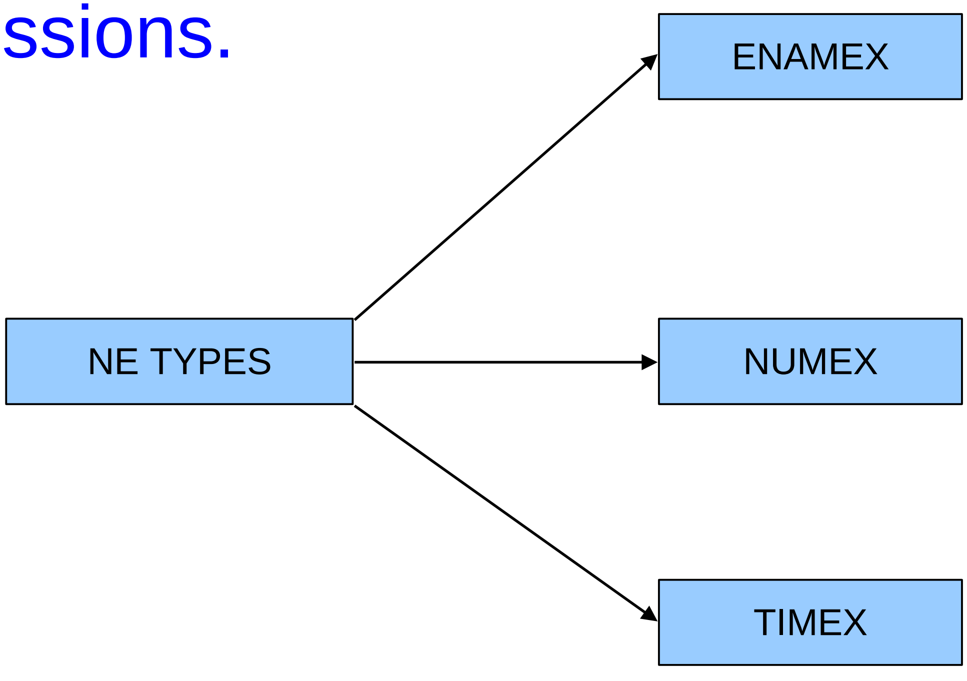
DATE "> 2006 </TIMEX>.

MUC TAGSET

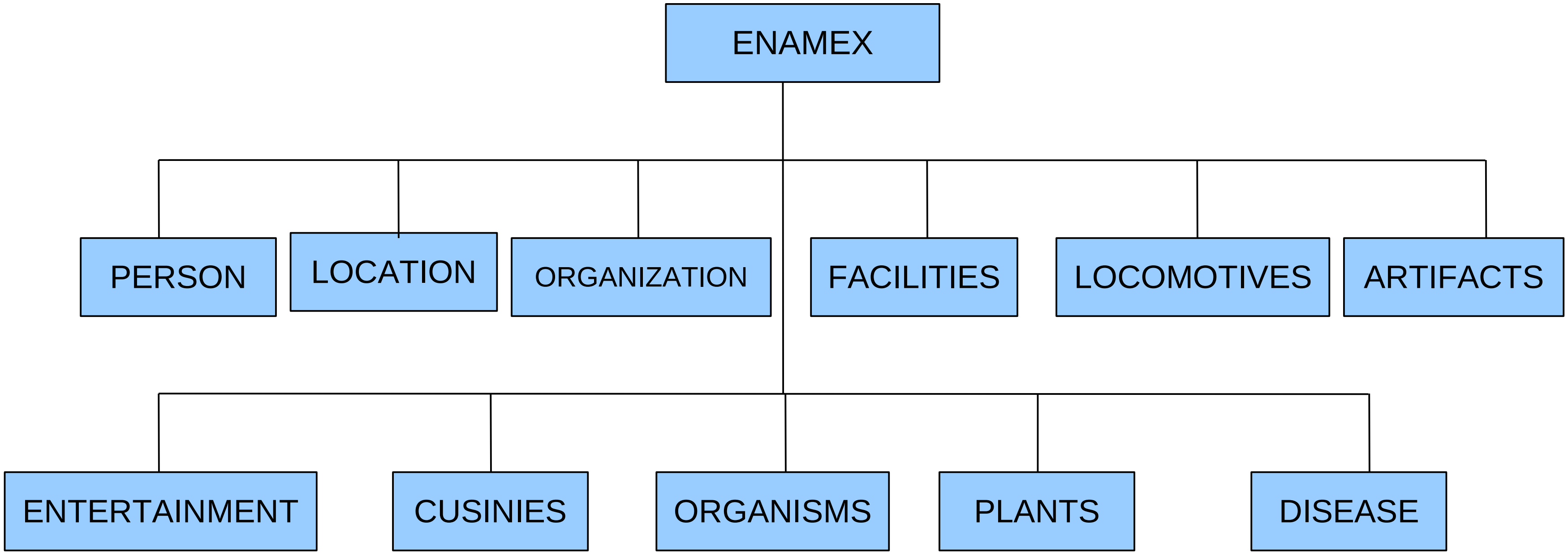
The Named entity hierarchy contains 106 tags

It is divided into three major classes Entity Name, Time and Numerical

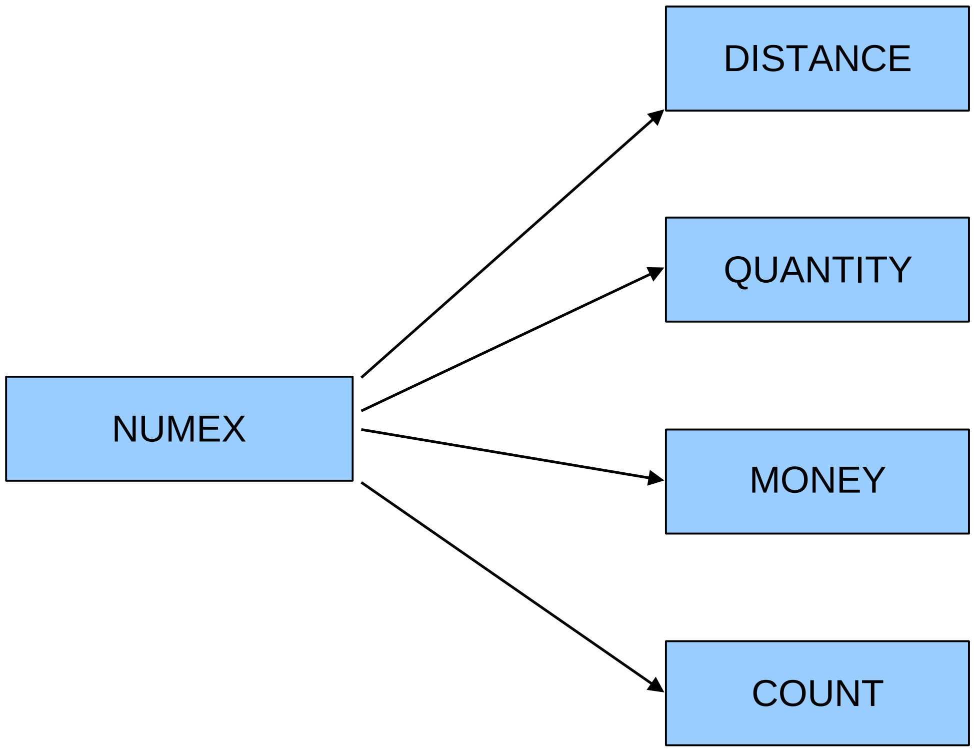
expressions.



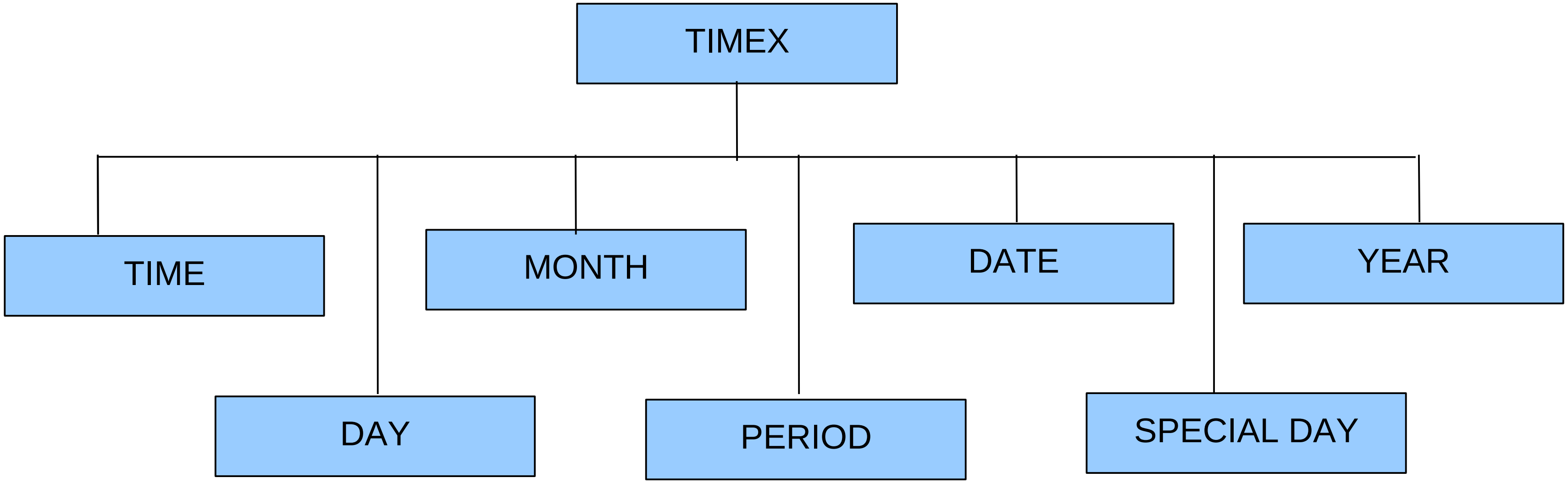
ENAMEX TYPES



NUMEX TYPES



TIMEX Types



PERSON

- Person
  - Individual
    - Family name
    - Title
  - Group

✓Persons are entities limited to humans. A person may be a single individual or a group.

✓Individual refer to names of each individual person, also includes names of fictional characters found in stories/novels etc.

✓Individual name occurs with family name

✓an individual is often referred with a title such as Mr., Mrs., Ms., Dr., etc along with their name

✓GROUP refers to set of individual

PERSON: Example

Mr. Chandrababu Naidu is the President of Telugu Desam Party

Chandrababu : Person => Individual
Naidu : Person => Family Name
Mr. : Person => Title

2)Apolo Hospital doctors : Person => GROUP

From MUC-7 Corpus

The shuttle, with an international crew of < NUMEX TYPE=“NUMBER”>six</NUMEX>, was set to blast off

from

<ENAMEX TYPE="ORGANIZATION|LOCATION">

Kennedy Space Center</ENAMEX>

on

<TIMEX TYPE="DATE">Thursday</TIMEX>

at <TIMEX TYPE="TIME">4:18 a.m. EST</TIMEX>.

Computation

NER solution Approaches

- Handcrafted systems
  - Knowledge (rule) based
    - Patterns
    - Gazetteers
- Automatic systems
  - Machine learning-*Supervised*, *Semi-supervised*, *Unsupervised*
  - Deep Learning based
- Hybrid systems

Corpora, Annotation

Some NE Annotated Corpora

- MUC-6 and MUC-7 corpora - English
- CONLL shared task corpora
  - http://cnts.uia.ac.be/conll2003/ner/ : NEs in English and German
  - http://cnts.uia.ac.be/conll2002/ner/ : NEs in Spanish and Dutch
- ACE – English - http://www.ldc.upenn.edu/Projects/ACE/
- TIDES surprise language exercise (NEs in Hindi)
- NERSSEAL shared task- NEs in Bengali, Hindi, Telugu, Oriya and Urdu (http://ltrc.iiit.ac.in/ner-ssea-08/index.cgi?topic=5)

Corpora, Annotation

- Biomedical and Biochemical corpora
  - BioNLP-04 shared task
  - BioCreative shared tasks
  - AiMed

Performance Evaluation

- Evaluation metric – mathematically defines how to measure the system's performance against a human-annotated, gold standard

- Scoring program–implements the metric and provides performance measures
  - For each document and over the entire corpus
  - For each type of NE

The Evaluation Metric

Precision = correct answers/answers produced

Recall = correct answers/total possible correct answers

Trade-off between precision and recall

F-Measure = $(\beta^2 + 1)PR / \beta^2R + P$

β reflects the weighting between precision and recall, typically β=1

The Evaluation Metric (2)

$$\text{Precision} = \frac{\text{Correct} + \frac{1}{2}\,\text{Partially correct}}{\text{Correct} + \text{Incorrect} + \text{Partial}}$$

$$\text{Recall} = \frac{\text{Correct} + \frac{1}{2}\,\text{Partially correct}}{\text{Correct} + \text{Missing} + \text{Partial}}$$

NE boundaries are often misplaced, so some partially correct results

Pre-processing for NER

- Format detection
- Word segmentation (for languages like Chinese)
- Tokenisation
- Sentence splitting
- Part-of-Speech (PoS) tagging

NER–automatic approaches

- Learning of statistical models or symbolic rules
  - Use of annotated text corpus
    - Manually annotated
    - Automatically annotated

- ML approaches frequently break down the NE task in two parts:
  - Recognising the entity boundaries
  - Classifying the entities in the NE categories

NER – automatic approaches

- Tokens in text are often coded with the IOB scheme
  - O – outside, B-XXX – first word in NE, I-XXX – all other words in NE

e.g.

Argentina	B-LOC
played	O
with	O
Del	B-PER
Bosque	I-PER

- Probabilities:
  - Simple:
    - P(tag i | token i)
  - With external evidence:
    - P(tag i | token i-1, token i, token i+1)

NER–automatic approaches

- Decision trees
  - Tree-oriented sequence of tests in every word
    - Determine probabilities of having a IOB tag
  - Use training data
  - Viterbi, ID3, C4.5 algorithms
    - Select most probable tag sequence
  - SEKINE et al (1998)
  - BALUJA et al (1999)
    - F-measure: 90%

NER – automatic approaches

- HMM-*Generative model*
  - Markov models, Viterbi
  - Works well when large amount of data is available Nymble (1997) / IdentiFinder (1999)

- Maximum Entropy (ME)-*Discriminative model*
  - Separate, independent probabilities for every evidence (external and internal features) are merged multiplicatively
  - MENE (NYU-1998)
    - Capitalization, many lexical features, type of text
    - F-Measure: 89%

ML features

- The choice of features
  - Lexical features (words)
  - Part-of-speech
  - Orthographic information
  - Affixes (prefix and suffix of any word)
  - Gazetteers

- External, unmarked data is useful to derive gazetteers and for extracting training instances

IdentiFinder [Bikel et al 99]

- Based on Hidden Markov Models
- 7 regions of HMM–one for each *MUC type*, *not-name*, *begin-sentence* and *end-sentence*

- Features
  - Capitalisation
  - Numeric symbols
  - Punctuation marks
  - Position in the sentence
  - 14 features in total, combining above info, e.g., containsDigitAndDash (09-96), containsDigitAndComma (23,000.00)

IdentiFinder (2)

- Evaluation: MUC-6 (English) and MET-1(Spanish) corpora

- Mixed case English
  - IdentiFinder - 94.9% F-measure
  - Best rule-based – 96.4% F-measure
- Spanish mixed case
  - IdentiFinder – 90% F-measure
  - Best rule-based - 93% F-measure
  - Lower case names, noisy training data, less training data

- Impact of size of data- Trained with 650,000 words, but similar performance with half of the data. Less than 100,000 words reduce the performance to below 90% on English

MENE [Borthwick et al 98]

- Rule-based NE + ML based NE- achieve better performance
- Tokens tagged as: XXX_start, XXX_continue, XXX_end, XXX_unique, other (non-NE), where XXX is an NE category
- Uses Maximum Entropy (ME)
  - One only needs to find the best features for the problem
  - ME estimation routine finds the best relative weights for the features

MENE (2)

- Features
  - Binary features–"token begins with capitalised letter", "token is a four-digit number"
  - Lexical features–dependencies on the surrounding tokens (window ±2) e.g., "Mr" for people, "to" for locations
  - Dictionary features–equivalent to gazetteers (first names, company names, dates, abbreviations)
  - External systems–whether the current token is recognised as a NE by a rule-based system

MENE (3)

- MUC-7 formal run corpus
  - MENE – *84.2%* F-measure
  - Rule-based systems– *86% - 91 %* F-measure
  - MENE + rule-based systems – *92%* F-measure

- Learning curve
  - 20 docs – 80.97% F-measure
  - 40 docs – 84.14% F-measure
  - 100 docs – 89.17% F-measure
  - 425 docs – 92.94% F-measure

NE Performance for Various Indian Languages (CRF based)

S.no	Language	Precision	Recall	F-measure
1	English	77.43	72.24	74.45
2	Tamil	78.48	64.31	68.26
3	Punjabi	73.01	64.92	68.74
4	Bengali	70.54	56.03	62.42
5	Marathi	64.43	64.74	64.61
6	Telugu	45.9	34.21	41.12

HMM based

Hidden Markov Model for NER

- One of the earliest successful method to solve NER was HMM proposed by Bikel et.al (1999)

- HMM is a generative model that tries to maximize

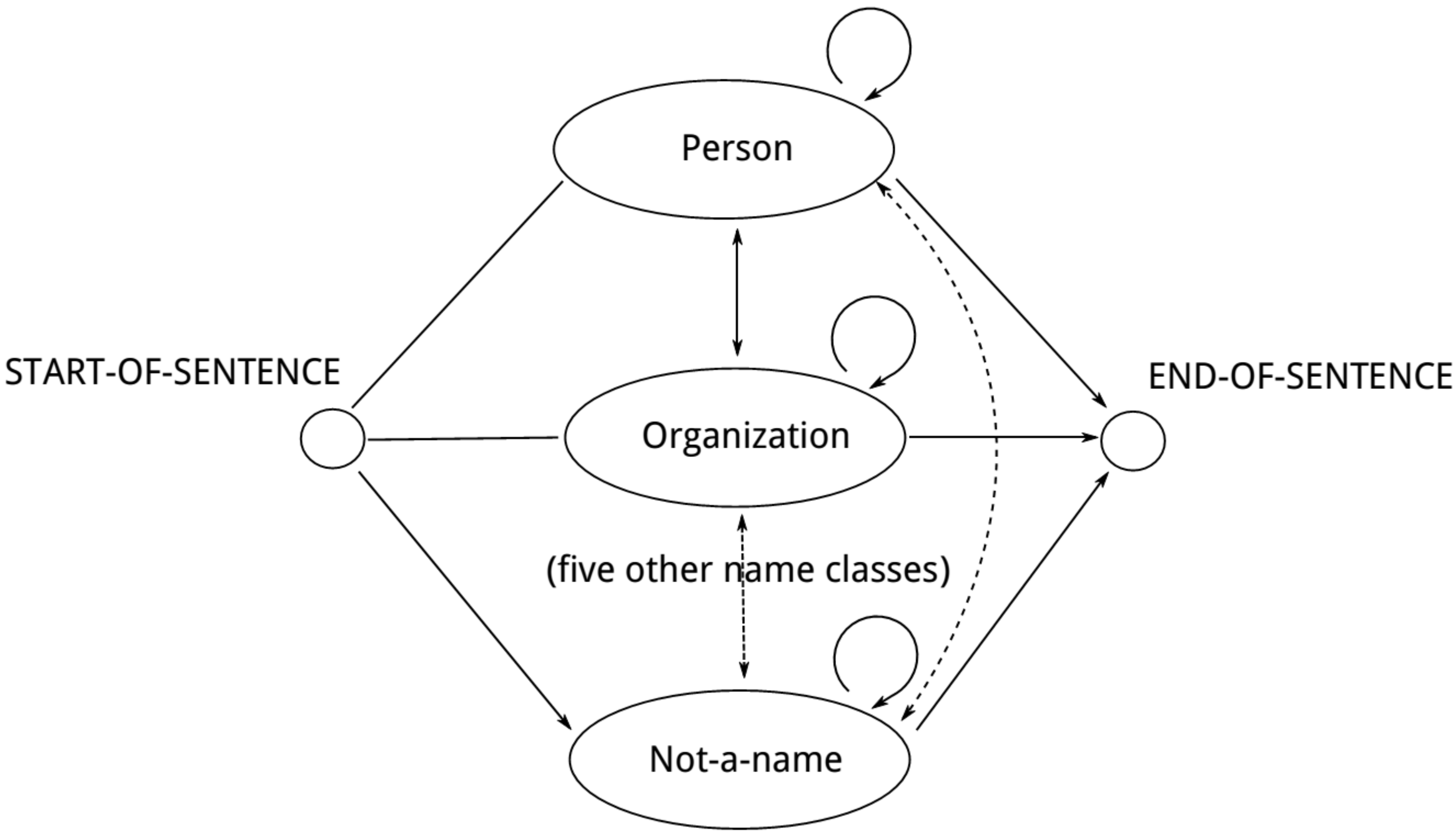
$$\max \Pr(NC \mid W)$$

$$\Pr(NC \mid W) = \frac{\Pr(NC, W)}{\Pr(W)}$$

  - Where NC= named entity class sequence and W= word sequence

- Just like POS tagging, the Viterbi algorithm is used to maximize Pr(NC,W) through the entire space of all possible name-class assignments

Model


Person
START-OF-SENTENCE
Organization
END-OF-SENTENCE
(five other name classes)
Not-a-name

Generation Step (HMM)

- Joint probability distribution of named class and words can be broken down as
  - Select a name-class nc, conditioned on the previous name-class and previous word.
  - Generate the first word inside the name-class, conditioning on the current and previous name-classes.

$$\Pr(nc \mid nc_{-1}, w_{-1}).\Pr(< w, f >_{first} \mid nc, nc_{-1})$$

  - Generate all subsequent words inside the current name-class, where each subsequent word is conditioned on its immediate predecessor

$$\Pr(< w, f > \mid < w, f >_{-1}, nc)$$

Example

Mr. Jones eats.

Mr. <ENAMEX TYPE="PERSON">Jones</ENAMEX> eats

Pr(NOT-A-NAME | START-OF-SENTENCE, +end+) *
Pr(“Mr.” | NOT-A-NAME, START-OF-SENTENCE) *
Pr(+end+ | "Mr.", NOT-A-NAME) *
Pr(PERSON | NOT-A-NAME ,"Mr.")*
Pr("Jones" | PERSON, NOT-A-NAME) *
Pr(+end+ | "Jones" PERSON) *
Pr(NOT-A-NAME | PERSON, "Jones") *
Pr("eats" | NOT-A-NAME, PERSON) *
Pr("." | \eats", NOT-A-NAME) *
Pr(+end+ | ".", NOT-A-NAME)*
Pr(END-OF-SENTENCE | NOT-A-NAME ".")

Lexical Knowledge Network, also called KNOWLEDGE GRAPHS

Use of Lexical Knowledge Networks

- Proliferation of ML based methods
- Still the need for deep knowledge based methods is acutely felt.
- ML methods capture surface phenomena and can do limited inference.
- Deeper knowledge needed for hard tasks like Text Entailment, Question Answering and other high end NLP tasks.

Consider the following problems

- How do you disambiguate 'web' in '*the spider spun a web*' from '*go surf the web*'?
- How do you summarise a long paragraph?
- How do you automatically construct language phrasebooks for tourists?
- Can a search query such as *"a game played with bat and ball"* be answered as *"cricket"*?

Some foundational points

Syntagmatic and Paradigmatic Relations

- Syntagmatic and paradigmatic relations
  - Lexico-semantic relations: synonymy, antonymy, hypernymy, mernymy, troponymy etc. **CAT is-a ANIMAL**
  - Coccurence: **CATS MEW**
- Wordnet: primarily paradigmatic relations
- ConceptNet: primarily Syntagmatic Relations

Selectional Preferences (Indian Tradition)

- "Desire" of some words in the sentence ("aakaangksha").
  - *I **saw** the boy with long hair.*
  - *The verb **"saw"*** and the noun ***"boy"** desire an object here.*
- "Appropriateness" of some other words in the sentence to fulfil that desire ("yogyataa").
  - *I saw the boy with long hair.*
  - *The PP **"with long hair"** can be appropriately connected only to **"boy"** and not **"saw".***
- In case, the ambiguity is still present, "proximity" ("sannidhi") can determine the meaning.
  - *E.g. I saw the boy with a telescope.*
  - *The PP **"with a telescope"** can be attached to both **"boy"** and **"saw"**, so ambiguity still present. It is then attached to*

Selectional Preference

- There are words which demand arguments, like, verbs, prepositions, adjectives and sometimes nouns. These arguments are typically nouns.
- Arguments must have the property to fulfil the demand. They must satisfy selectional preferences.
  - Example
    - Give (verb)
      - » agent – animate
      - » obj – direct
      - » obj – indirect
    - *I **gave** him the **book***
    - *I **gave** him the **book** (yesterday in the school) -> adjunct*
- How does this help in WSD?
  - One type of contextual information is the information about the type of arguments that a word takes.

Verb Argument frame

- Structure expressing the desire of a word is called the *Argument Frame*
- Selectional Preference
  - Properties of the “Supply Words” meeting the desire of the previous set

Argument frame (example)

Sentence: *I am fond of X*

Fond

{

Arg1: *Prepositional Phrase (PP)*

{

PP: *<u>of</u> NP*

{

N: *somebody/something*

}

}

}

Verb Argument frame (example)

Verb: *give*

Give

{

agent: *<the give>$_{animate}$*

direct object: *<the thing given>*

indirect object: *<beneficiary>$_{animate/organization}$*

}

[I]$_{\textbf{agent}}$ gave a [book]$_{\textbf{dobj}}$ to [Ram]$_{\textbf{iobj}}$.

Parameters for Lexical Knowledge Networks

1. Domains addressed by the structure

2. Principles of Construction

3. Methods of Construction

4. Representation

5. Quality of database

6. Applications

7. Usability mechanisms for software applications and users: APIs, record structure, User interfaces

8. Size and coverage

Wordnet

https://www.cfilt.iitb.ac.in/indowordnet/

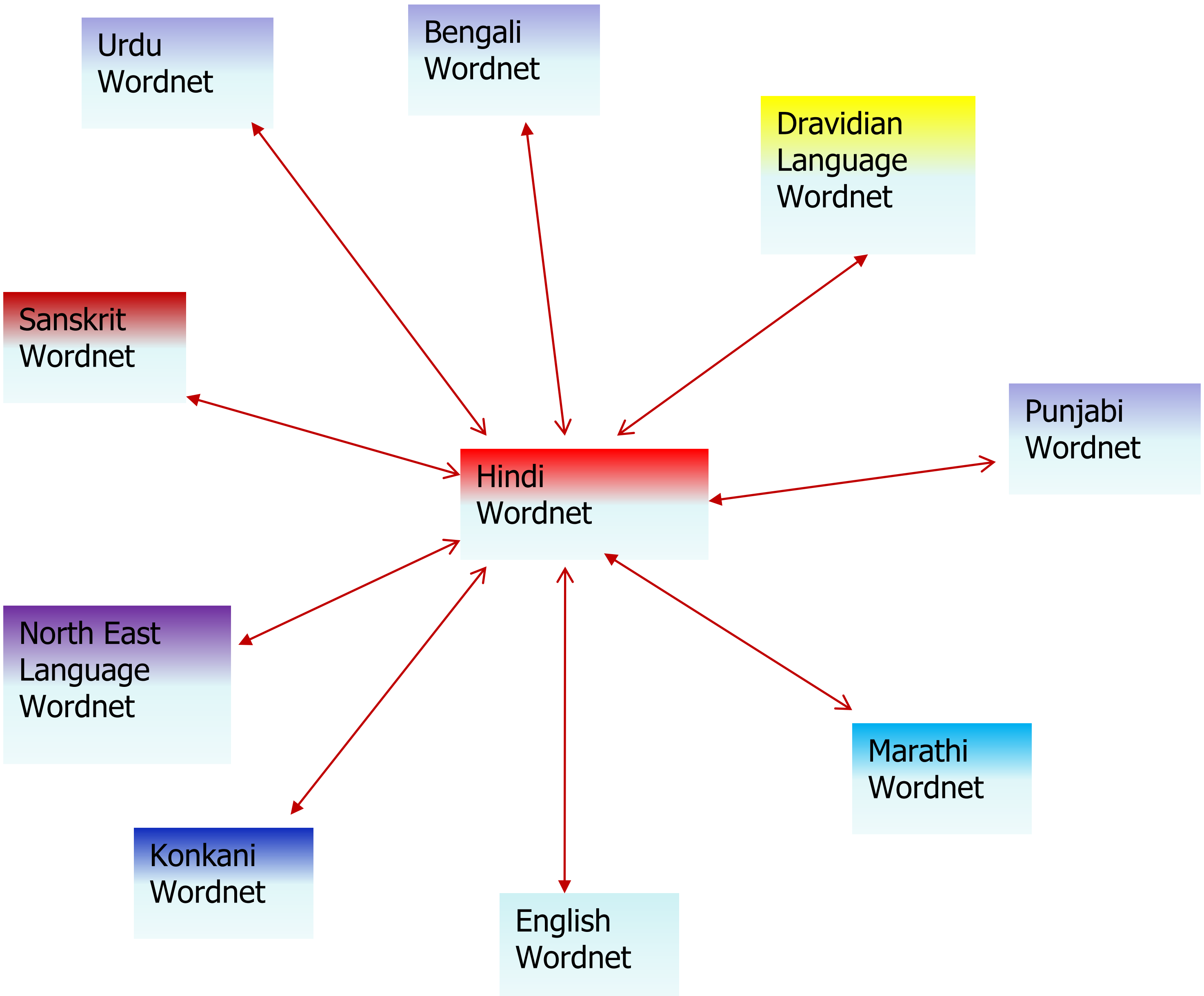
Wordnet: main purpose

- Disambiguation: **Sense Disambiguation**
- Main instrument: Relational Semantics
- Disambiguate a *by other words*
  - *{house}: "house" as a kind of "physical structure"*
  - *{family, house}: "family" as an abstract concept*
  - *{house, astrological position}: "astrological place" as a concept*

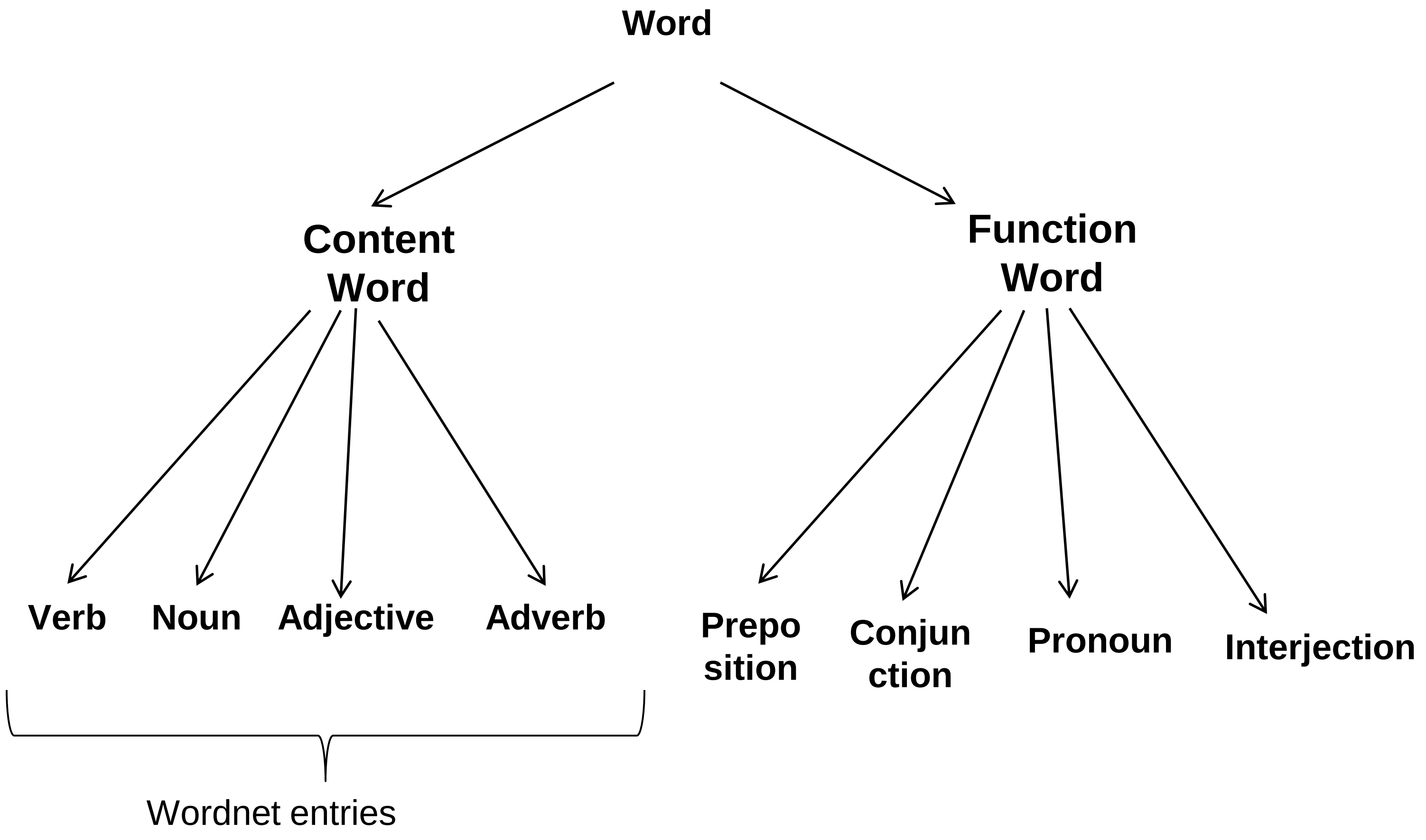
Wordnet - Lexical Matrix (with examples)

Word Meanings	Word Forms				
	F_1	**F_2**	**F_3**	**…**	**F_n**
M_1	(*depend*) $E_{1,1}$	(*bank*) $E_{1,2}$	(rely) $E_{1,3}$		
M_2		*(bank)* $E_{2,2}$		(*embankment*) $E_{2,\ldots}$	
M_3		*(bank)* $E_{3,2}$	$E_{3,3}$		
…				…	
M_m					$E_{m,n}$

INDOWORDNET



Classification of Words



Sense tagged corpora (task: sentiment analysis)

- I have enjoyed_21803158 #LA#_18933620 every_42346474 time_17209466 I have been_22629830 there_3110157 , regardless_3119663 if it was for work_1578942 or pleasure_11057430.
- I usually_3107782 fly_21922384 into #LA#_18933620, but this time_17209466 we decided_2689493 to drive_21912201 .
- Interesting_41394947, to say_2999158 the least_3112746 .

Senses of “pleasure”

The noun pleasure has 5 senses (first 2 from tagged texts)

1. (21) pleasure, pleasance -- (a fundamental feeling that is hard to define but that people desire to experience; "he was tingling with pleasure")
2. (4) joy, delight, pleasure -- (something or someone that provides pleasure; a source of happiness; "a joy to behold"; "the pleasure of his company"; "the new car is a delight")
3. pleasure -- (a formal expression; "he serves at the pleasure of the President")
4. pleasure -- (an activity that affords enjoyment; "he puts duty before pleasure")

Basic Principle

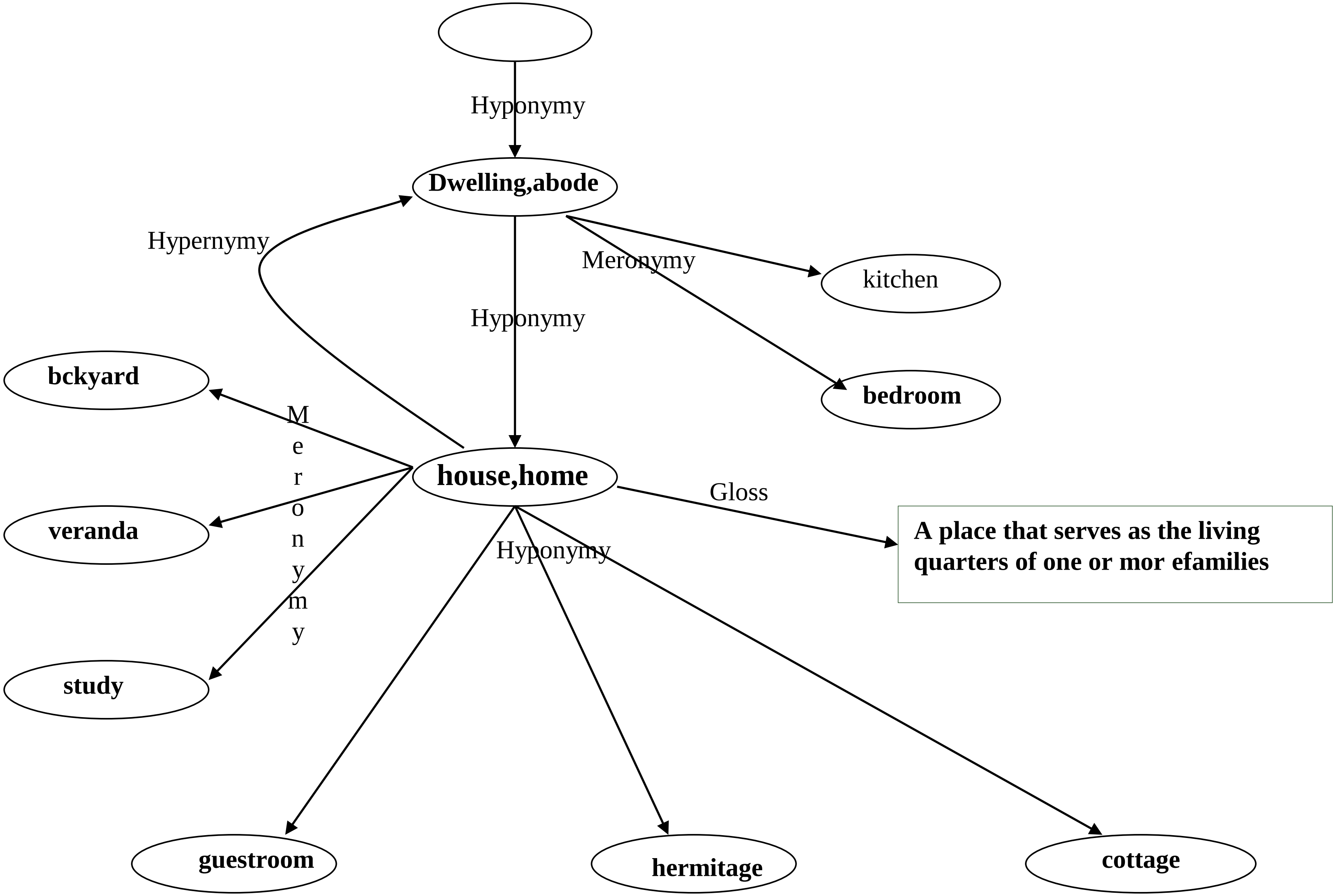
- Words in natural languages are polysemous.
- However, when synonymous words are put together, a unique meaning often emerges.
- Use is made of *Relational Semantics.*

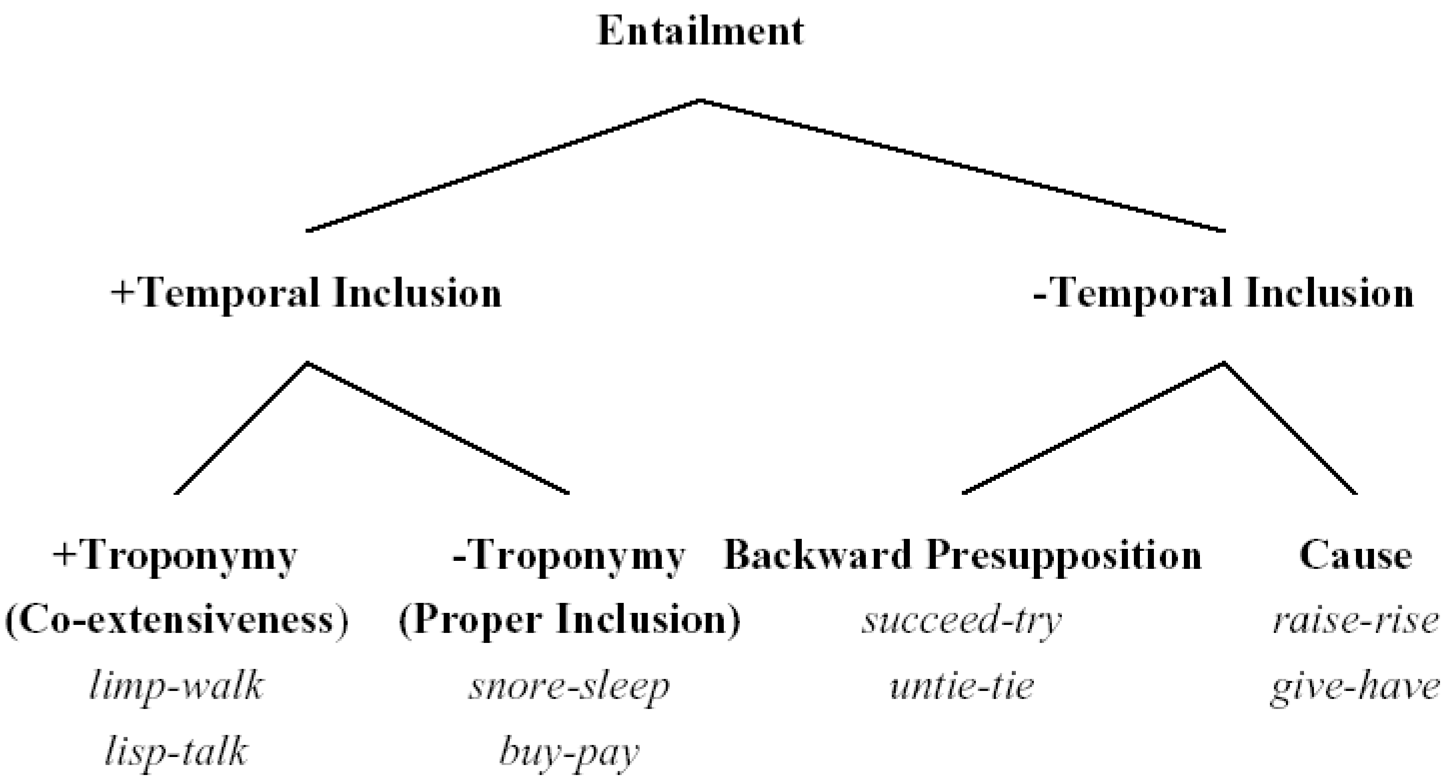
Lexical and Semantic relations in wordnet

1. Synonymy
2. Hypernymy / Hyponymy
3. Antonymy
4. Meronymy / Holonymy
5. Gradation
6. Entailment
7. Troponymy

1, 3 and 5 are lexical (*word to word)*, rest are semantic (*synset to synset).*

WordNet Sub-Graph




Entailment
+Temporal Inclusion
-Temporal Inclusion
+Troponymy
(Co-extensiveness)
limp-walk
lisp-talk
-Troponymy
(Proper Inclusion)
snore-sleep
buy-pay
Backward Presupposition
succeed-try
untie-tie
Cause
raise-rise
give-have

Principles behind creation of Synsets

Three principles:

- Minimality
- Coverage
- Replacability

Synset creation: from first principles

From first principles

- Pick all the senses from good standard dictionaries.
- Obtain synonyms for each sense.
- Needs hard and long hours of work.

Synset creation: Expansion approach

From the wordnet of another language preferably in the **same family**

- Pick the synset and obtain the sense from the gloss.
- Get the words of the target language.
- Often same words can be used- especially for words with the same etymology borrowed from the parent language in the typology.
- Translation, Insertion and deletion.

Illustration of expansion approach with noun[1]

English

- bank (sloping land (especially the slope beside a body of water)) "they pulled the canoe up on the bank"; "he sat on the bank of the river and watched the currents"

French (*wrong!*)

- ~~banque~~ (les terrains en pente (en particulier la pente à côté d'un plan d'eau)) "ils ont tiré le canoë sur la rive», «il était assis sur le bord de la rivière et j'ai vu les courants"

Illustration of expansion approach with noun[2]

No hypernymy
in the synset

English

- bank (sloping land (especially the slope beside a body of water)) "they pulled the canoe up on the bank"; "he sat on the bank of the river and watched the '

French

- {rive, rivage, ~~bord~~} (les terrains en pente (en particulier la pente à côté d'un plan d'eau)) "ils ont tiré le canot sur la rive», «il était assis sur le bord de la rivière et j'ai vu les courants"

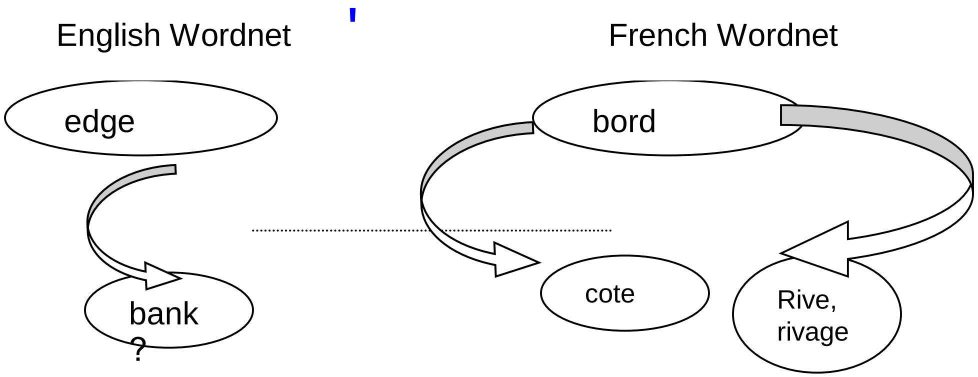


Illustration of expansion approach with verb[3]

English

- trust, swear, rely, bank (have confidence or faith in) "We can trust in God"; "Rely on your friends"; "bank on your good education"

French

- compter_sur, avoir_confiance_en, se_fier_a ’, faire_confiance_a’ (avoir confiance ou foi en) "Nous pouvons faire confiance en Dieu»,«Fiez-vous à vos amis",

Ordered by frequency

Lexical Relation

- Antonymy
  - Oppositeness in meaning
  - Relation between word forms
  - Often determined by phonetics, word length etc. ({rise, ascend} *vs.* {fall, descend})

Kinds of Antonymy

Size	**Small - Big**
Quality	**Good – Bad**
State	**Warm – Cool**
Personality	**Dr. Jekyl- Mr. Hyde**
Direction	**East- West**
Action	**Buy – Sell**
Amount	**Little – A lot**
Place	**Far – Near**
Time	**Day - Night**
Gender	**Boy - Girl**

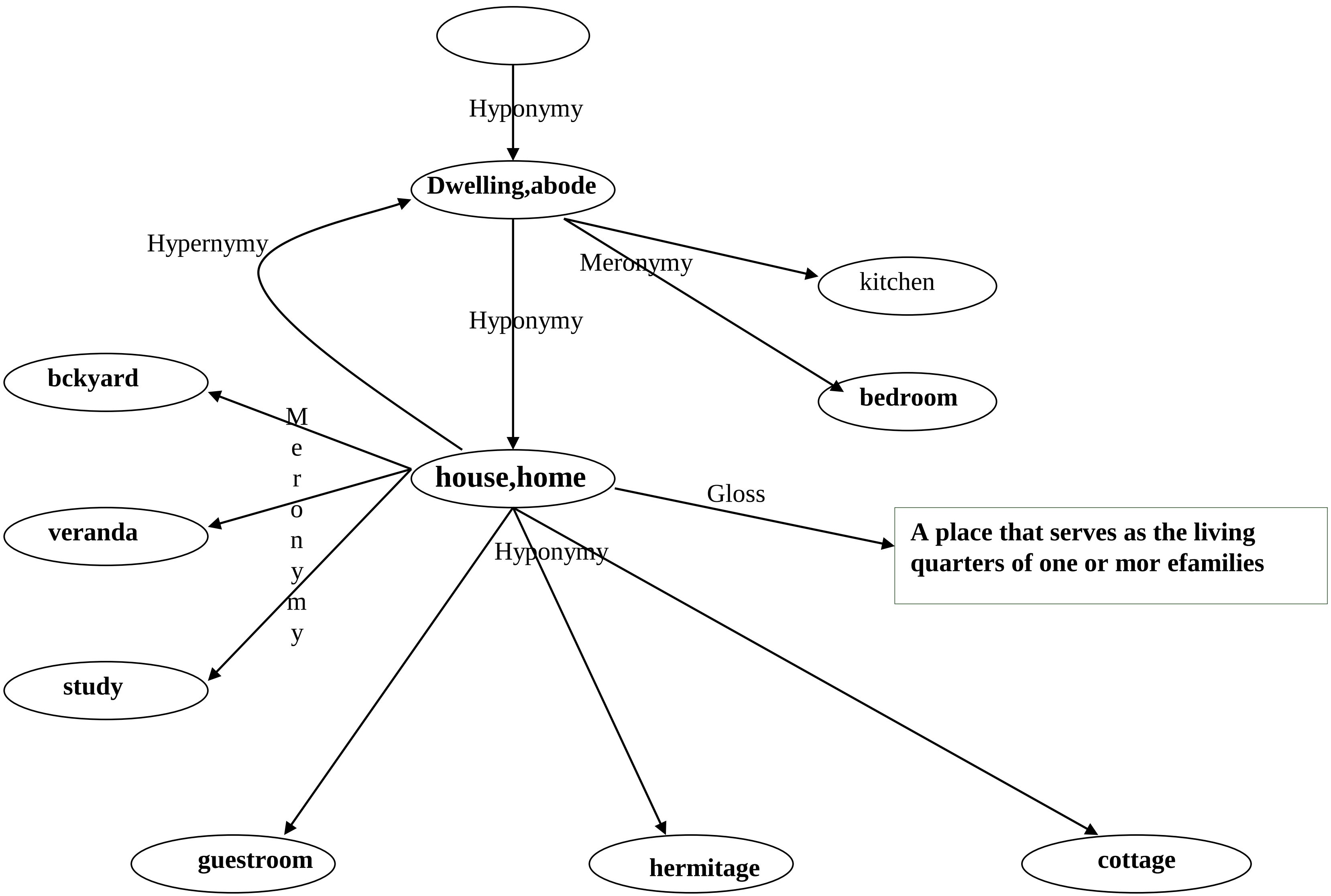
Kinds of Meronymy

Component-object	**Head - Body**
Staff-object	**Wood - Table**
Member-collection	**Tree - Forest**
Feature-Activity	**Speech - Conference**
Place-Area	**Palo Alto - California**
Phase-State	**Youth - Life**
Resource-process	**Pen - Writing**
Actor-Act	**Physician - Treatment**

Gradation

State	Childhood, Youth, Old age
Temperature	Hot, Warm, Cold
Action	Sleep, Doze, Wake

WordNet Sub-Graph



Metonymy

- Associated with *Metaphors* which are epitomes of semantics
- Oxford Advanced Learners Dictionary definition: “The use of a word or phrase to mean something different from the literal meaning”
- *Does it mean Careless Usage?!*

WordNet: limitations

- Contains little syntactic information
- No explicit predicate argument structures
- No systematic extension of basic senses
- Sense distinctions are very fine-grained, **IAA 73%**
- No hierarchical entries

ConceptNet

ConceptNet

- From MIT (Liu and Singh, 2004)
- Capture common sense information
- Emphasis on everyday knowledge rather than rigorous linguistic lexical differentiations (unlike wordnet)

Projects to Collect Commonsense[1]

- Cyc
  - Started in 1984 by Dr. Doug Lenat
  - Developed by CyCorp, with 3.2 millions of assertions linking over 280000 concepts and using thousands of micro-theories.
  - Cyc-NL is still a “potential application”, knowledge representation in frames is quite complicated and thus difficult to use.

Projects to Collect Commonsense[2]

- Open Mind Common Sense Project
  - Started in 2000 at MIT by Push Singh
  - WWW collaboration with over 20,123 registered users, who contributed 812,769 items
  - Used to generate *ConceptNet*, very large semantic network.

- Other such projects
  - HowNet (Chinese Academy of Science)
  - FrameNet (Berkley)

“I borrowed ‘Treasure Island’ for a couple of weeks”

- 1. `Treasure Island' is the name of a book
- 2. People borrow books to read
- 3. The book was most likely borrowed from a library
- 4. The book has to be returned to the lender in 14 days time

Flavors of common sense knowledge

- Emotive (“I feel awful")
- Functional (“Cups hold liquids")
- Causal (“Extracting a tooth causes pain")
- spatial (“Horses are usually found in stables")

OMCS *(Singh et al, 2004)*

- Open Mind Common Sense Project
- Volunteers contribute assertions (crowdsourcing?)
- 30 different activities of everyday life
- Short semi structured sentences (as opposed to synsets)
- Volunteers follow “patterns” in the “help” menu

Example of patterns

- *'Treasure Island' is a book: is-a pattern*
- *Books are found in a library: is-located-in pattern*
- Patterns make it possible to create machine processable data

ConceptNet: structure

- Directed Acyclic Graph formed by linking over 1.5 million assertions into a semantic network of about 300000 nodes
- Each node: a fragment of text (unlike synset) *aka* “concept”
- Nodes
  - NP: “watermelon”
  - VP: “breath air”

ConceptNet: Relations[1]

- 20 relations grouped into 8 thematic types
  - 1. K-Lines: ConceptuallyRelatedTo, ThematicKLine, SuperThematicK-Line
  - 2. Things: IsA, PropertyOf, PartOf, MadeOf, DefnedAs
  - 3. Agents: CapableOf

(note the difference from wordnet lexicon semantic relations like antonymy, hypernymy etc.)

ConceptNet: Relations[2]

- Themes:
  - 4. Events: PrerequisiteEvent, FirstSubEventOf, LastSubEventOf, SubEventOf
  - 5. Spatial: LocationOf
  - 6. Causal: EffectOf, DesirousEffectOf
  - 7. Functional: UsedFor, CapableOfReceivingAction
  - 8. Affective: MotivationOf, DesireOf

Twenty Semantic Relation Types in ConceptNet (Liu and Singh, 2004)	
THINGS (52,000 assertions)	IsA: (IsA "apple" "fruit") Part of: (PartOf "CPU" "computer") PropertyOf: (PropertyOf "coffee" "wet") MadeOf: (MadeOf "bread" "flour") DefinedAs: (DefinedAs "meat" "flesh of animal")
EVENTS (38,000 assertions)	PrerequisiteeventOf: (PrerequisiteEventOf "read letter" "open envelope") SubeventOf: (SubeventOf "play sport" "score goal") FirstSubeventOF: (FirstSubeventOf "start fire" "light match") LastSubeventOf: (LastSubeventOf "attend classical concert" "applaud")
AGENTS (104,000 assertions)	CapableOf: (CapableOf "dentist" "pull tooth")
SPATIAL (36,000 assertions)	LocationOf: (LocationOf "army" "in war")
TEMPORAL time & sequence	
CAUSAL (17,000 assertions)	EffectOf: (EffectOf "view video" "entertainment") DesirousEffectOf: (DesirousEffectOf "sweat" "take shower")
AFFECTIONAL (mood, feeling, emotions) (34,000 assertions)	DesireOf (DesireOf "person" "not be depressed") MotivationOf (MotivationOf "play game" "compete")
FUNCTIONAL (115,000 assertions)	IsUsedFor: (UsedFor "fireplace" "burn wood") CapableOfReceivingAction: (CapableOfReceivingAction "drink" "serve")
ASSOCIATION **K-LINES** (1.25 million assertions)	SuperThematicKLine: (SuperThematicKLine "western civilization" "civilization") ThematicKLine: (ThematicKLine "wedding dress" "veil") ConceptuallyRelatedTo: (ConceptuallyRelatedTo "bad breath" "mint")

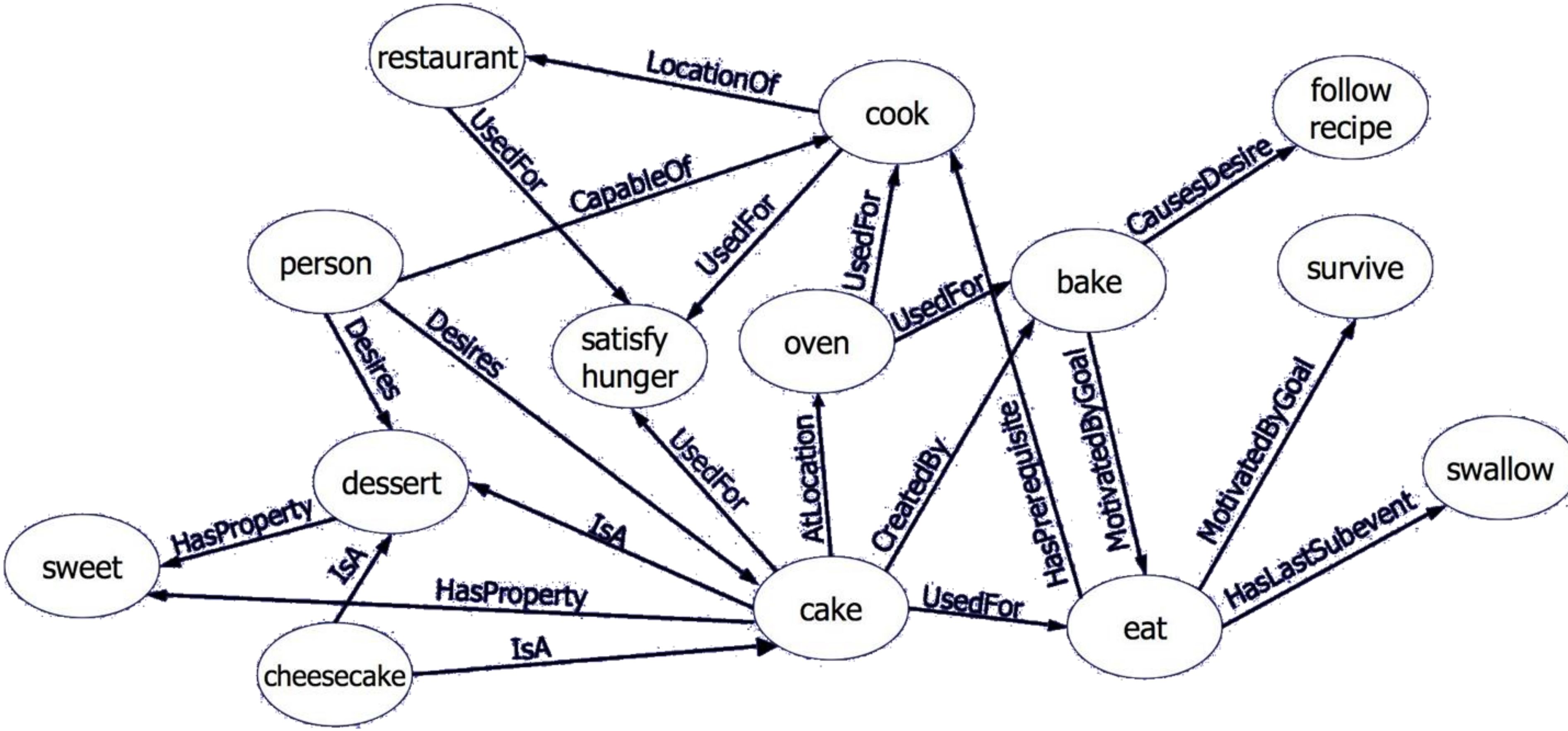
Table 1 ConceptNet's relational ontology of 20 link types.

ConceptuallyRelatedTo	IsA	FirstSubeventOf	DesirousEffectOf
ThematicKLine	MadeOf	SubeventOf	UsedFor
SuperThematicKLine	DefinedAs	LastSubeventOf	LocationOf
CapableOfReceivingAction	CapableOf	PrerequisiteEventOf	MotivationOf
PropertyOf	PartOf	EffectOf	DesireOf

Table 2 Ontology of concept types.

Events	Things	Places	Properties
Eat sandwich	Orange juice	At zoo	Furry
Sell car	Morning coffee	On table	Very expensive
Tell story	Policeman	Near school	Dark
Go to zoo	Leaf blower	Inside oven	Quickly
Type letter	Laptop computer	In closet	Dark

ConceptNet: Example1



ConceptNet: Example2

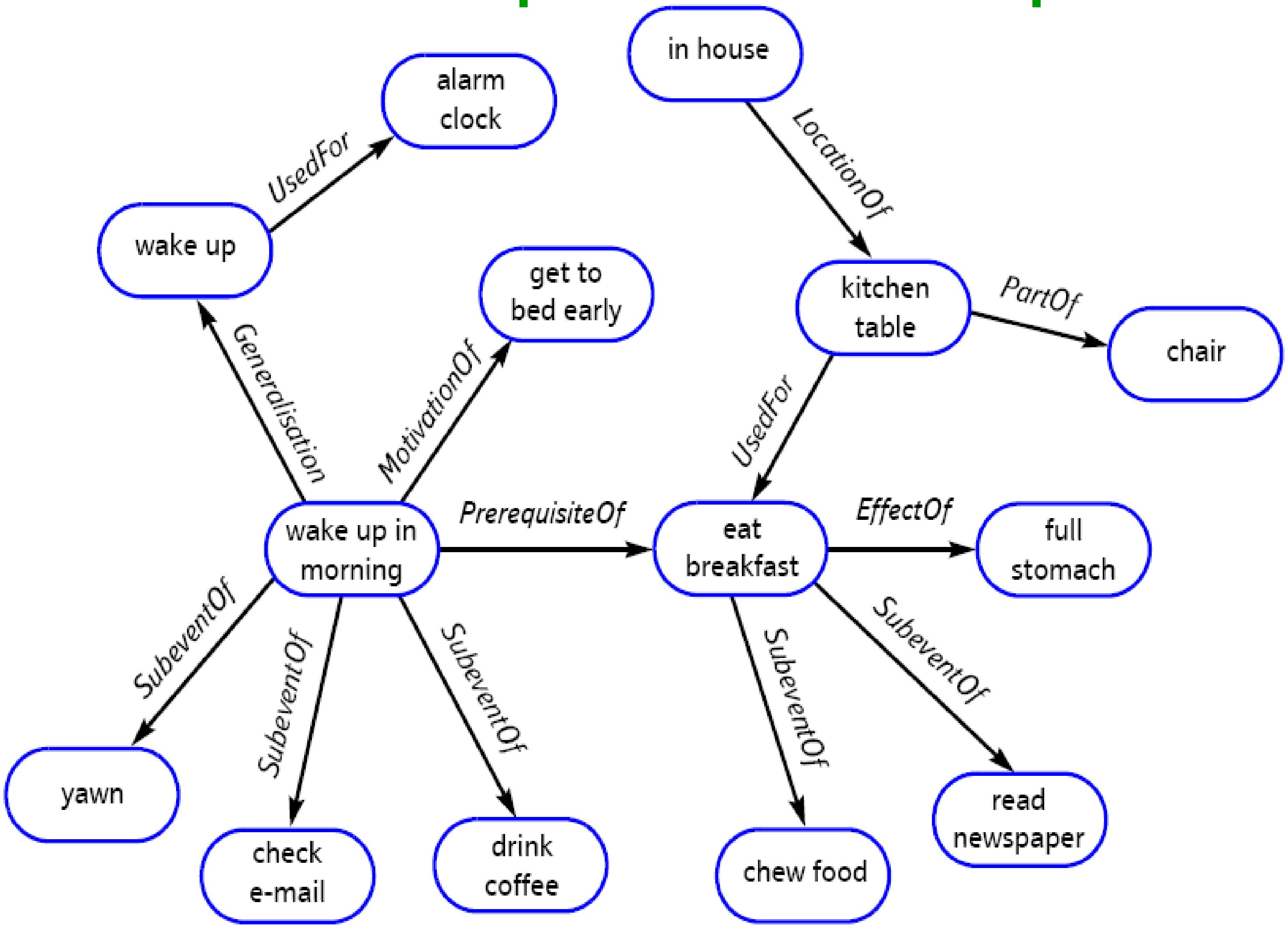


Fig 2 A subset of ConceptNet.

Sentence→ConceptNet

- Extraction
  - 50 regular expression rules run on OMCS sentences
  - Database creation
- Normalization
  - Spell corrrection
  - Stop word removal if needed
  - Lemmatization
- Relaxation

Database of ConceptNet

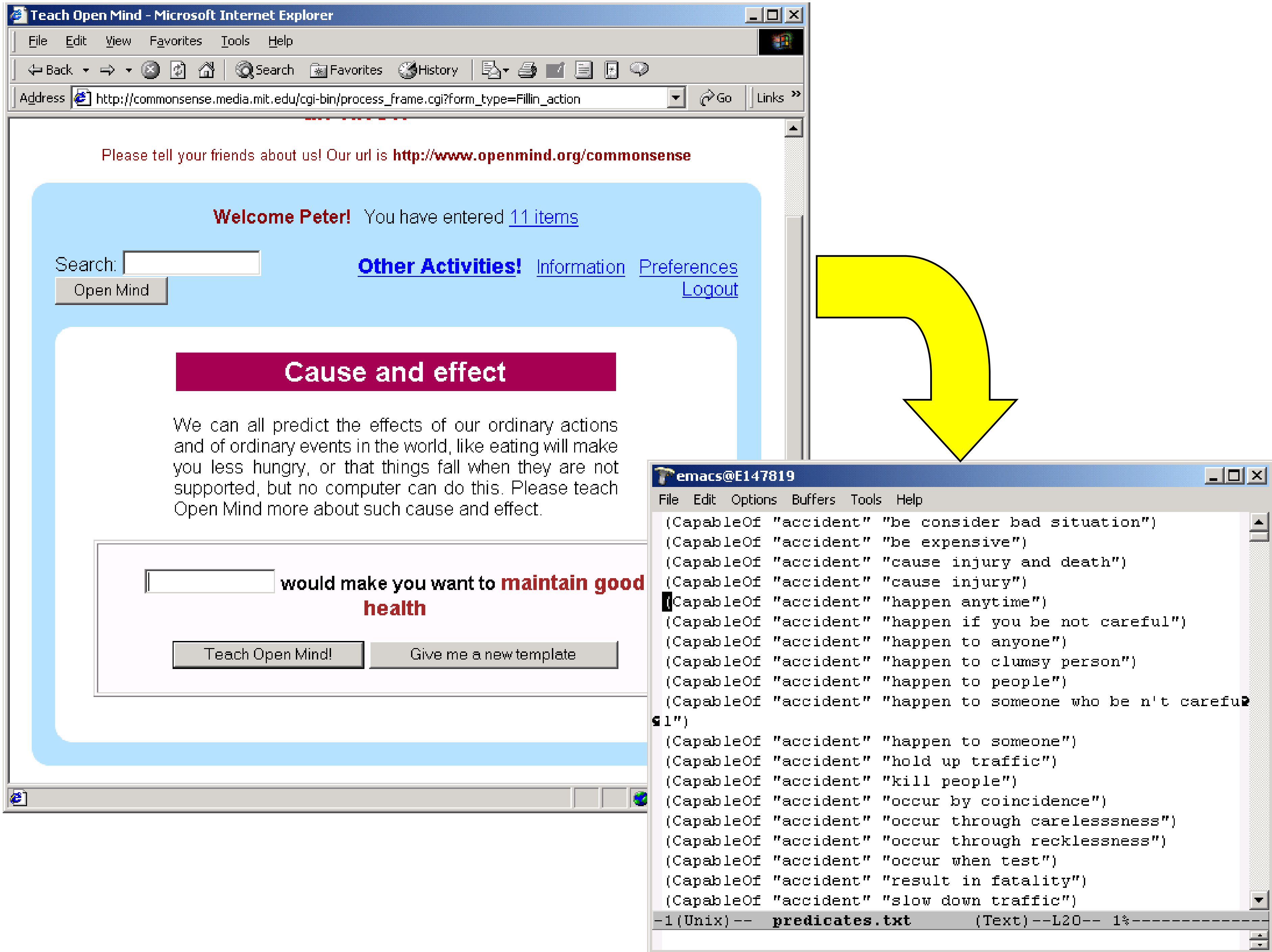
- Relations and facts are stored
  - (IsA “spider" “bug" “f=3; i=0;")
  - (LocationOf “waitress" “in restaurant" “f=2; i=0;")
- Frequency of seeing the assertion recorded as the number of times this relation was found through an inferencing procedure

Inference in ConceptNet

- Multiple assertions inferred from a single Open Mind sentence
- Example: ‘A lime is a sour fruit’
- Infer *IsA(lime, fruit)*
- Additionally infer *PropertyOf(lime, sour)*
- Infer Generalisations
  - if the majority of fruits have the property ‘sweet’,
  - then this property is lifted to the parent class, as: Property Of(fruit, sweet).

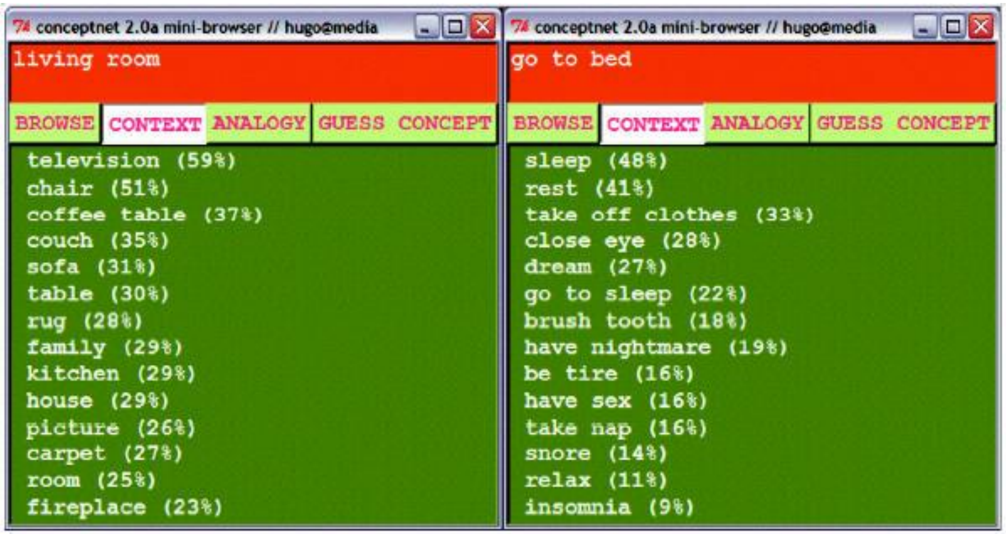
MontyLingua NLP engine

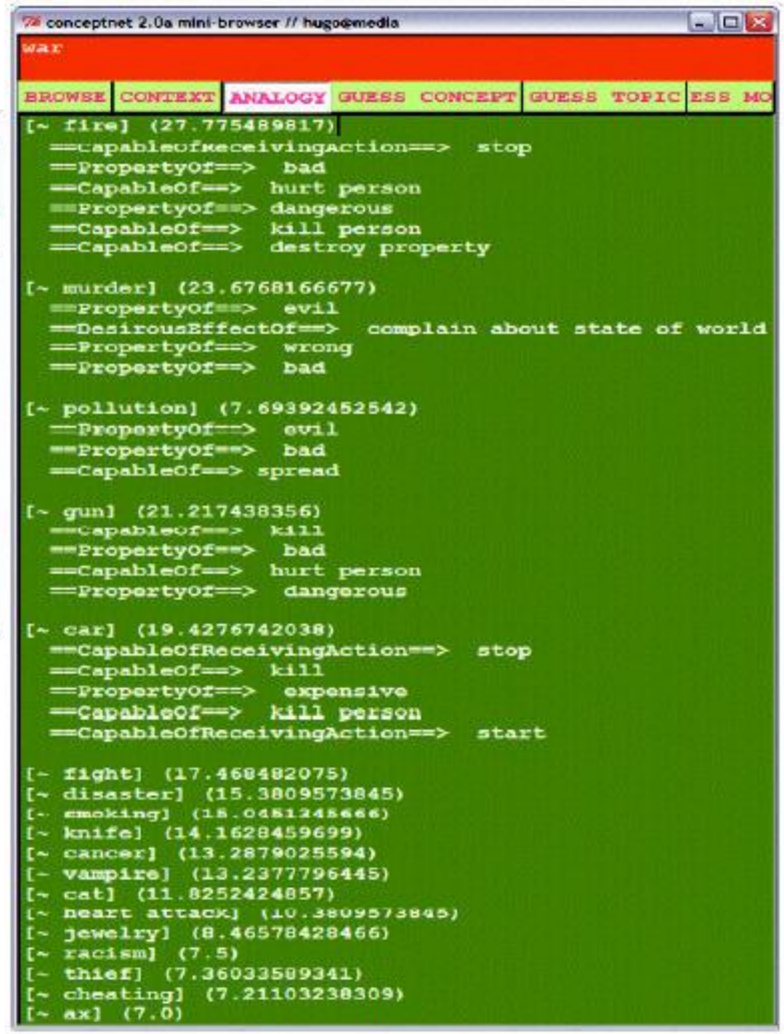
- Textual information management
- Written in Python, also available in Java
- Liberates ConceptNet from normalization of text
- Can take running paras and sentences
- API (2003) — http://web.media.mit.edu/~hugo/montylingua



Teach Open Mind - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites History
Address http://commonsense.media.mit.edu/cgi-bin/process_frame.cgi?form_type=Fillin_action
Go Links
Please tell your friends about us! Our url is http://www.openmind.org/commonsense
Welcome Peter! You have entered 11 items
Search:
Open Mind
Other Activities! Information Preferences Logout
Cause and effect
We can all predict the effects of our ordinary actions and of ordinary events in the world, like eating will make you less hungry, or that things fall when they are not supported, but no computer can do this. Please teach Open Mind more about such cause and effect.
would make you want to maintain good health
Teach Open Mind!
Give me a new template
emacs@E147819
File Edit Options Buffers Tools Help
(CapableOf "accident" "be consider bad situation")
(CapableOf "accident" "be expensive")
(CapableOf "accident" "cause injury and death")
(CapableOf "accident" "cause injury")
(CapableOf "accident" "happen anytime")
(CapableOf "accident" "happen if you be not careful")
(CapableOf "accident" "happen to anyone")
(CapableOf "accident" "happen to clumsy person")
(CapableOf "accident" "happen to people")
(CapableOf "accident" "happen to someone who be n't carefu
l")
(CapableOf "accident" "happen to someone")
(CapableOf "accident" "hold up traffic")
(CapableOf "accident" "kill people")
(CapableOf "accident" "occur by coincidence")
(CapableOf "accident" "occur through carelesssness")
(CapableOf "accident" "occur through recklessness")
(CapableOf "accident" "occur when test")
(CapableOf "accident" "result in fatality")
(CapableOf "accident" "slow down traffic")
-1(Unix)-- predicates.txt (Text)--L20-- 1%

Snapshot of ConceptNet





ConceptNet Application[1]

- Commonsense ARIA
  - Observes a user writing an e-mail and proactively suggests photos relevant to the user's story
  - Bridges semantic gaps between annotations and the user's story
- GOOSE
  - A goal-oriented search engine for novice users
  - Generate the search query

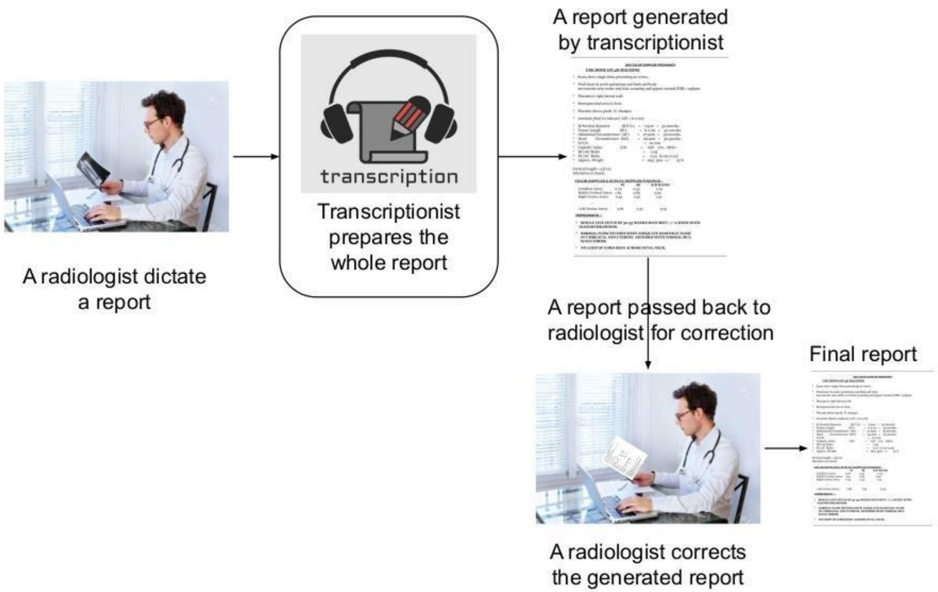
ConceptNet Application[2]

- Makebelieve
  - Story-generator that allows a person to interactively invent a story with the system
  - Generate causal projection chains to create storylines
- GloBuddy: A dynamic foreign language phrasebook
- AAA: Rrecommends products from Amazon.com by using ConceptNet to reason about a person's goals and desires,creating a profile of their predicted tastes.

Knowledge Graphs in Automatic Radiology Report Generation

Kaveri Kale, Pushpak Bhattacharyya and Kshitij Jadhav, *Replace and Report: NLP Assisted Radiology Report Generation*, **ACL 2023** Findings, Toronto, July 9-14, 2023.

Problem: Scarcity of Radiologists



- **Radiologist to Patient Ratio**
  - India: 1:100,000
  - US: 1:10,000
  - China: 1:14,772
- **High Patient Inflows**
  - Radiologists are extremely busy
  - Increased stress levels
- **Delays in Report Generation**
  - Significant delays in report turnaround time

Radiologist’s Dictation and Pathological Description

Radiologist’s dictation: *Chronic pancreatitis.*

Pathological description: *Pancreas is slightly small, reveals thin inhomogenous parenchyma. The pancreatic duct is dilated.*

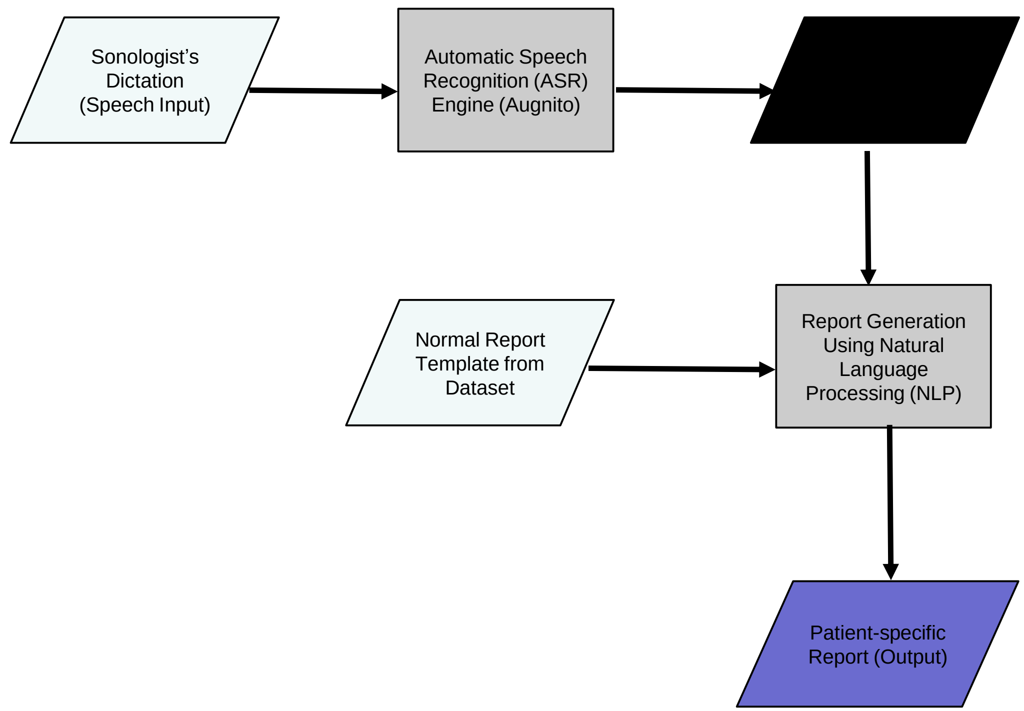
Radiologist’s dictation: *Cholecystitis with 3 mm gallbladder calculus in lumen.*

Pathological description: *Gallbladder is distended reveals wall thickening. Feature of note is presence of a calculus measuring 3 mm noted in lumen of gallbladder.*

Radiologist’s dictation: *Grade ii fatty liver.*

Pathological description: *Liver shows moderate increase in echogenicity.*

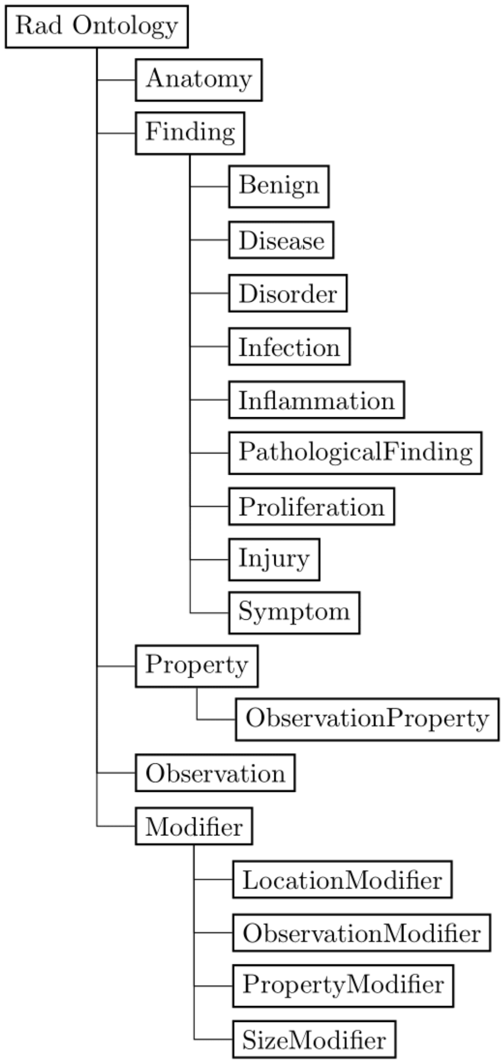
Automatic Radiology Report Generation Workflow



Dataset and Knowledge Graphs

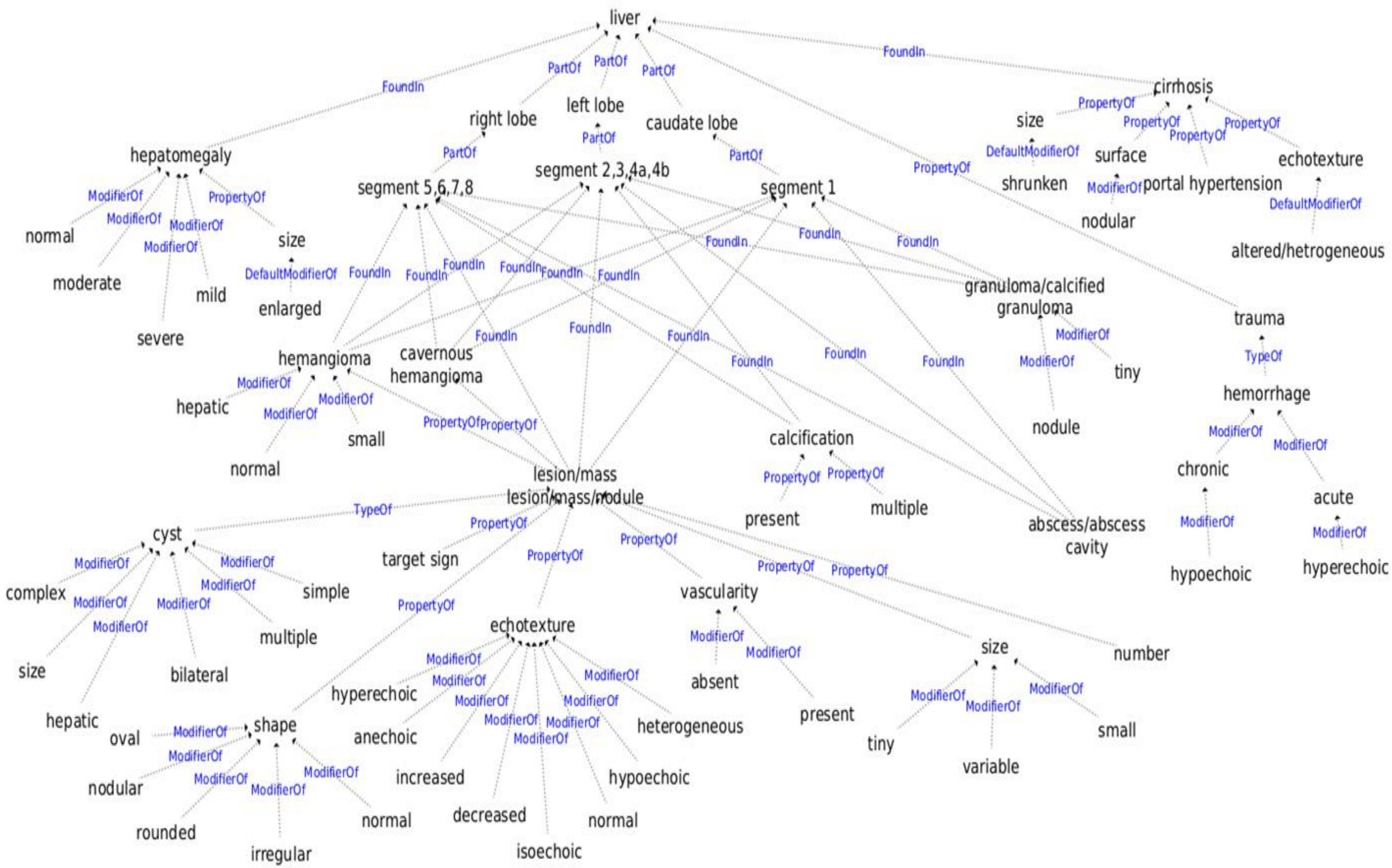
- **Dataset**: radiology reports collected from 5 different hospitals
- Approximately 100,000 reports
- Of these, 15,000 Ultrasound
- All USG reports anonymised
- **Knowledge Graphs**- *Gallbladder, Urinary bladder, Pancreas, Kidney, Liver, Prostate, Ovary, Uterus, Ureter (from ultrasound reports)*

Radiology Ontology



- Logical relations: **PartOf, TypeOf, PropertyOf, ModifierOf, and FoundIn**
- Attributes: prefered_name, synonyms, word_forms, E.g., lesion is instance of class Observation
- Class: Observation
- Attributes: prefered_name(lesion), synonyms (mass lesion, mass, nodule), word_forms(lesion, mass, nodule)

Liver Knowlegde Graph (Semin-automatic)



Radiologist's Dictation and Pathological Description

Radiologist's dictation: *Chronic pancreatitis.*

Pathological description: *Pancreas is slightly small, reveals thin inhomogenous parenchyma. The pancreatic duct is dilated.*

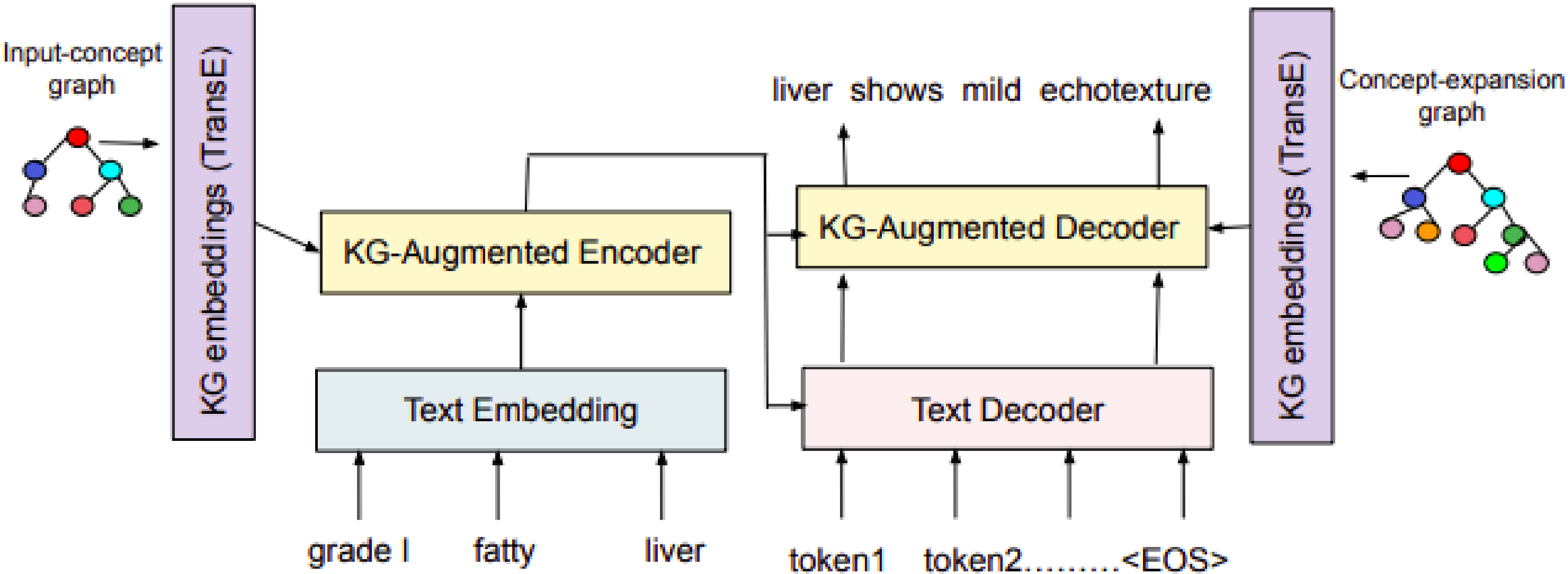
Radiologist's dictation: *Cholecystitis with 3 mm gallbladder calculus in lumen.*

Pathological description: *Gallbladder is distended reveals wall thickening. Feature of note is presence of a calculus measuring 3 mm noted in lumen of gallbladder.*

Radiologist's dictation: *Grade ii fatty liver.*

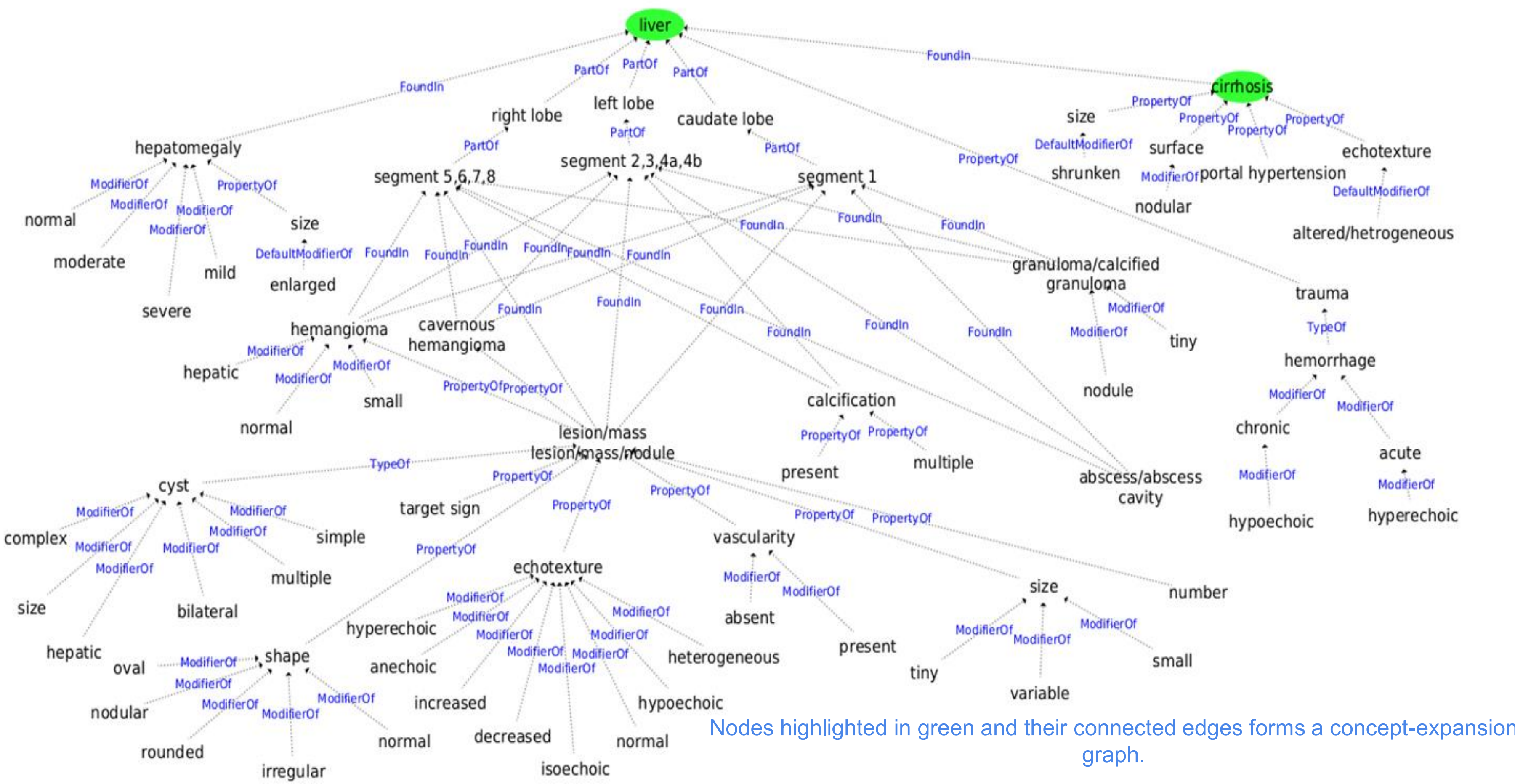
Pathological description: *Liver shows moderate increase in echogenicity.*

Deep Learning Model: KG-BART



Grounded KG: Input-concept Graph

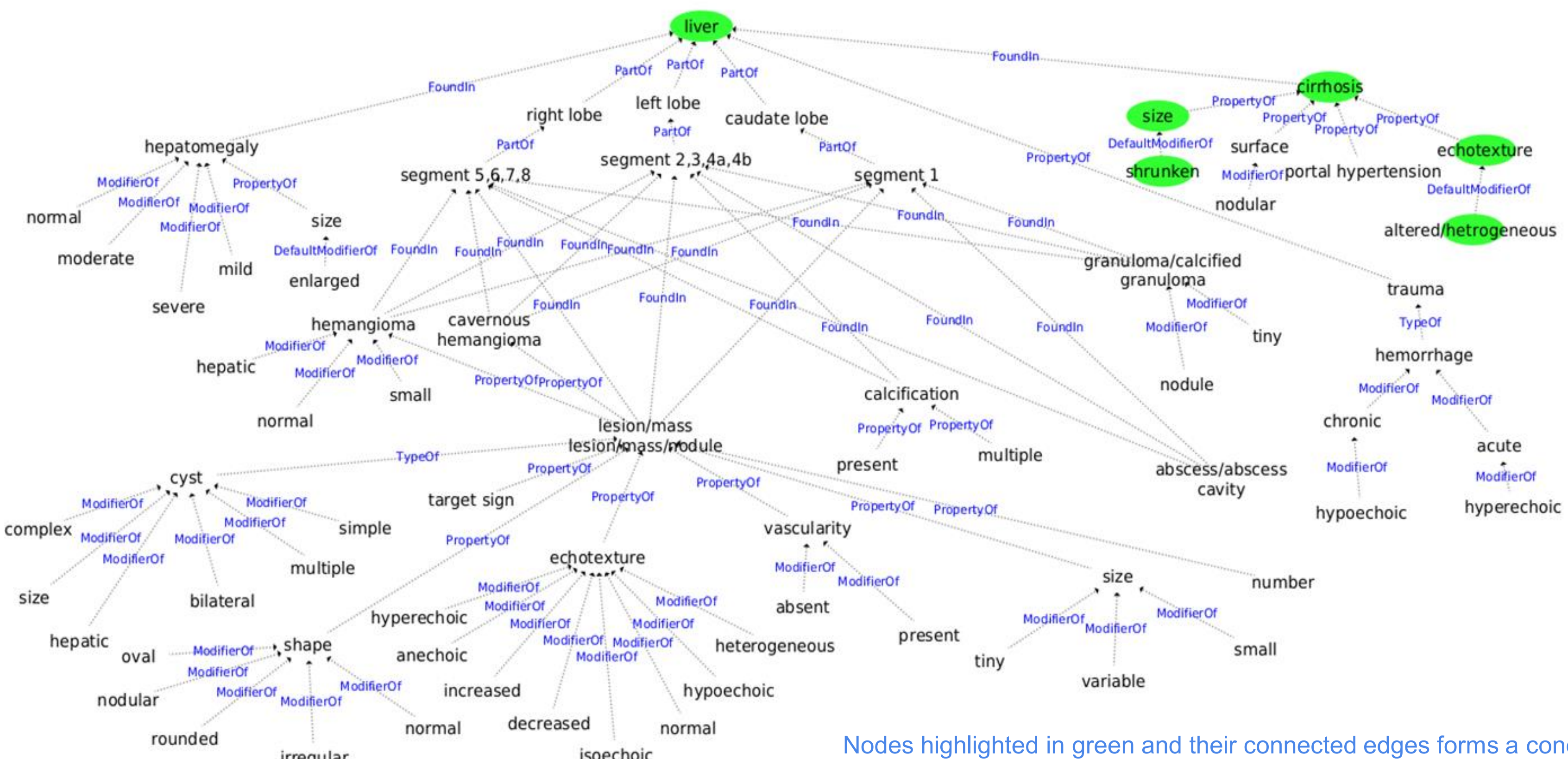
- Input Dictation: “liver cirrhosis”   Expected Output (PD): “Liver is small in size with altered echotexture.”
- Extracted Entities: liver, cirrhosis   Triplets from input-concept graph: (cirrhosis, FoundIn, liver)



Nodes highlighted in green and their connected edges forms a concept-expansion graph.

Grounded KG: Concept-expansion Graph (produces PD)

Triplets from concept-expansion graph: {(shrunken, DefaultModifierOf, size), (size, PropertyOf, cirrhosis), (altered/heterogeneous, DefaultModifierOf, echotexture), (echotexture, ProperyOf, cirrhosis), (cirrhosis, FoundIn, liver))}



Nodes highlighted in green and their connected edges forms a concept-expansion graph.

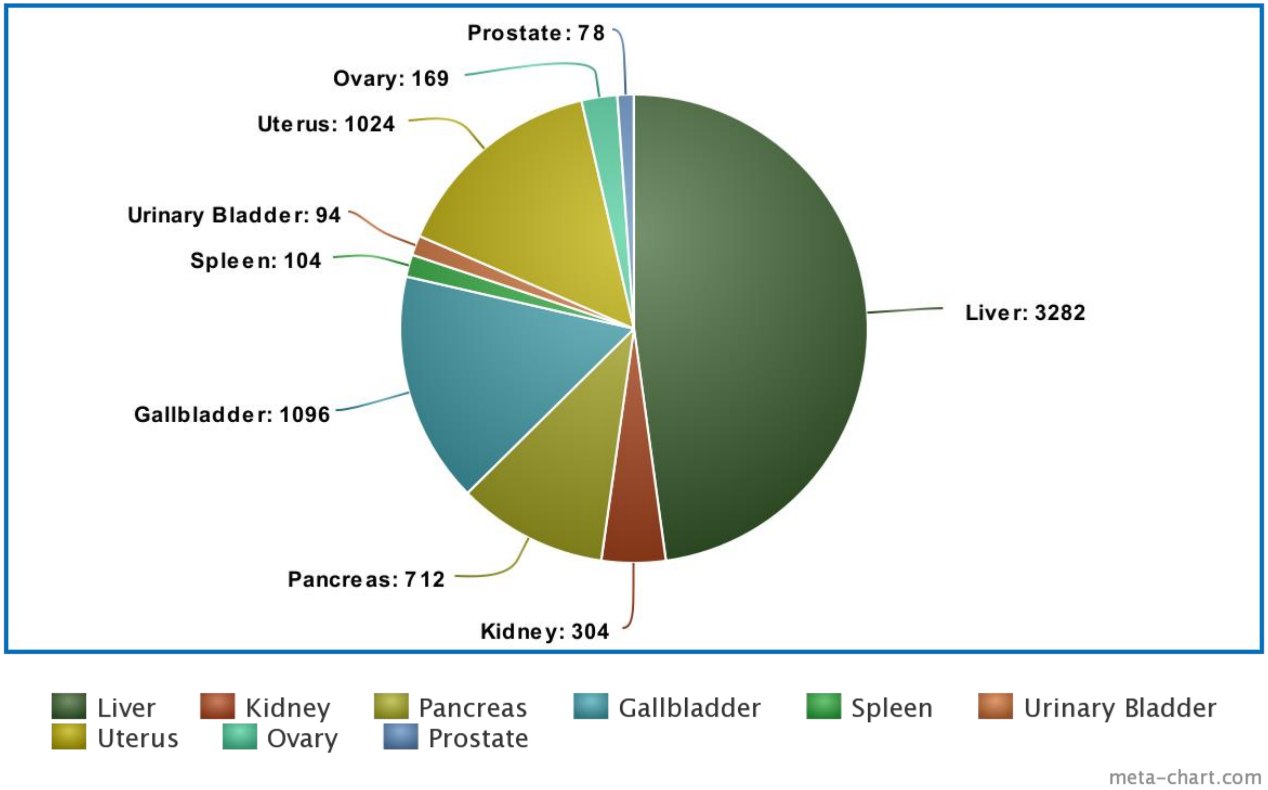
Parallel Dataset

Impression	Radiologists' Notes	Concept Set	Pathological Description
Bulky retroverted uterus with fundal fibroid.	Bulky retroverted uterus with fundal fibroid 2.3 x 5.6 mm.	uterus, fibroid, bulky, fundal, retroverted, 2.3 x 5.6 mm	Uterus is retroverted and bulky in size. Myometrial reflectivity is inhomogeneous with an illdefined fundal fibroid measuring 2.3 x 5.6 mm noted.
Calculus cholecystitis with multiple large calculi.	Calculus cholecystitis with multiple large calculi within lumen of gallbladder, largest measuring 2.4 mm.	multiple, calculi, calculus, lumen, cholecystitis, enlarged, measuring, 2.4 mm	Gallbladder is distended reveals thick wall. Feature of note is presence of multiple large calculi seen within lumen of gallbladder; largest calculus measures 2.4 mm.
Acute pancreatitis.	Acute pancreatitis.	acute pancreatitis	Pancreas is bulky, reveals reduced reflectivity with increased reflectivity of peripancreatic fat.

Data Statistics

Total Samples	Train Samples	Test Samples	Validation Samples
6860	6000	430	430

Data Statistics



Quantitative Results

Models	BLUE Score				ROUGE Score		
	1-gram	2-gram	3-gram	4-gram	1-gram	3-gram	L-gram
T5-base	0.81	0.74	0.68	0.63	0.88	0.77	0.85
T5-large	0.85	0.80	0.75	0.72	0.88	0.80	0.87
BART-base	0.86	0.82	0.78	0.75	0.90	0.82	0.89
BART-large	<u>0.89</u>	<u>0.85</u>	<u>0.84</u>	<u>0.81</u>	<u>0.92</u>	<u>0.86</u>	<u>0.92</u>
KG-BART	**0.93**	**0.89**	**0.86**	**0.83**	**0.96**	**0.89**	**0.94**

Error Analysis: Pathological Description Generation

Input	bulky and retroverted uterus 1 x 5 x 3.4 mm with endometrial thickness 7 mm.
T5-large	uterus is retroverted and bulky in size 1 x 5 x 3.4 mm. myometrial reflectivity is inhomogeneous with an echogenic lesion measuring 7 mm in the mid anterior wall.
BART-large	Uterus is retroverted and bulky in size 1 x 5 & 3.4 mm. Myometrial reflectivity is inhomogeneous and shows normal echotexture. It has uniform epigastrophic echopattern.
KG-BART	uterus is retroverted and bulky in size 1 x 5 x 3.4 mm. myometrial reflectivity is inhomogeneous. Endometrial thickness is 7 mm in size.
Target	Uterus is retroverted and bulky in size measuring 1 x 5 x 3.4 mm. myometrial reflectivity is inhomogeneous and shows bulky echotexture. Endometrial thickness is 7 mm in size.
Input	gross splenomegaly, maximum span of spleen is 8.2 mm with focal calcifications noted in spleen.
T5-large	e/o spleen is grossly enlarged and normal in echotexture, maximum span of stiple is 8.2 mm. multiple calcified granulomas noted in adnexa.
BART-large	spleen is grossly enlarged and normal in echotexture, maximum span of spleen is 8.2 mm. multiple calcified granulomas noted in gb.
KG-BART	Spleen is grossly enlarged and normal in echotexture, maximum span of spleen is 8.2 mm. multiple calcified granulomas noted in spleen.
Target	Spleen is grossly enlarged and normal in echotexture, maximum span of spleen is 8.2 mm. Multiple calcified granulomas noted in spleen.

Span Identification and Replacement

ULTRASONOGRAPHY OF THE ABDOMEN AND PELVIS

Liver:

Liver is normal in size and echotexture. (liver1) No focal areas of altered echotexture or mass lesion. (liver2) No intrahepatic biliary radicles dilatation seen. (biliary_radicals1) Portal vein appears normal. (portal_vein1) Portal vein measures _. (portal_vein2) Common duct at porta measures _ . (common_duct1)

Gall Bladder:

Gall bladder is physiologically distended reveals normal wall thickness. (gallbladder1) No evidence of calculi/calculus or sludge or polyp. (gallbladder2)

Spleen:

Spleen is normal in size with normal echotexture. (spleen1) The contours are smooth. (spleen2) The splenic vein and portal vein are normal in caliber. (spleen3)

Pancreas:

Pancreas appears normal in size and echotexture. (pancreas1)

Kidney:

Right Kidney measures _ x _. (kidney1) Left Kidney measures _ x _. (kidney2) Both the kidneys are normal in position, size, shape and contour. (kidney3) Cortical echogenicity is normal, corticomedullary differentiation is well maintained. (kidney4) No obvious calculus or mass is seen. (kidney5) No hydronephrosis noted. (kidney6)

Ureter:

Ureters are not dilated.

Urinary Bladder:

Urinary bladder appears normal. (urinary_bladder1) Wall thickness is normal. (urinary_bladder2) No evidence of calculus or mass is seen. (urinary_bladder3) Pre void is _ cc. Post void is _ cc. (urinary_bladder4)

Prostate:

The prostate is normal in size and echotexture measuring _. (prostate1)

Impression: No abnormality found.

Dataset Samples (Pathological Description:Class)

Pathological Description (PD)	**Class**
1 mm calculus noted in lower pole of right kidney.	kidney5
there is a 2 calcification seen in the right lobe of prostate.	prostate1
left ovary shows a cyst measuring 4 x 6.3 cm with thick septation.	ovary6
liver is enlarged in size with normal echopattern a tiny anechoic thin walled cyst measuring 3 x 5 x 4 mm in segment vi and segment vii of right lobe of liver.	liver1, liver2
Liver is enlarged in size and spleen is normal in size.	liver1, spleen1
uterus is anteverted showing enlarged in size with a fibroid of size 1 x 5 x 3.4 mm in posterior wall.	uterus3
spleen is normal in size and pancreas shows a well defined smooth walled hypoechoic area with multiple low level echoes seen in relation to tail of pancreas.	spleen1, pancreas1

The first row means that the PD should replace the 5th sentence in the “kidney block” of the normal report

Identification of the sentence in the normal report to be replaced with PD: Examples

Single sentence to replace:

Pathological Description: Liver is severely enlarged in size 6 cm and echotexture.	**Identified Normal Sentence:** Liver is normal in size and echotexture. (liver1)

Multiple sentences to replace:

Pathological Description: Right kidney measures 9.4 x 8 mm and left kidney measures 9.4 x 8 mm.	**Identified Normal Sentences:** Right Kidney measures _x_. (kidney1). Left Kidney measures _ x _. (kidney2)

Radiology Report Generation

chronic pancreatitis
cholecystitis with 3 mm gall bladder calculus in lumen
grade ii fatty liver

Select Gender: Male | Generate Report | Download Report | New Report

Toggle Report

Male Abdomen Pelvis Normal Report

Liver is normal in size and echotexture. No focal areas of altered echotexture or mass lesion. No intrahepatic biliary radicles dilatation seen. Portal vein appears normal. Portal vein measures _. common duct at porta measures _ .

Gall bladder is physiologically distended reveals normal wall thickness. No evidence of calculi/calculus or sludge or polyp.

Spleen is normal in size with normal echotexture. The contours are smooth. The splenic vein and portal vein are normal in caliber.

Pancreas appears normal in size and echotexture.

Right Kidney measures _ x _. Left Kidney measures _ x _. Both the kidneys are normal in position, size, shape and contour. Cortical echogenicity is normal, corticomedullary differentiation is well maintained. No obvious calculus or mass is seen. No hydronephrosis noted.

Ureters are not dilated.

Urinary bladder appears normal. Wall thickness is normal. No evidence of calculus or mass is seen. Pre void is _ cc. Post void is _ cc.

The prostate is normal in size and echotexture measuring _.

Generated Output

pancreas is slightly small, reveals thin inhomogenous paranchyma. The pancreatic duct is dilated.

gallbladder is distended reveals wall thickening. feature of note is presence of a calculus measuring 3 mm noted in lumen of gallbladder.

liver shows moderate increase in echogenicity.

Generated Report

Liver shows moderate increase in echogenicity. No focal areas of altered echotexture or mass lesion. No intrahepatic biliary radicles dilatation seen. Portal vein appears normal. Portal vein measures _. common duct at porta measures _ .

Gallbladder is distended reveals wall thickening. feature of note is presence of a calculus measuring 3 mm noted in lumen of gallbladder.

Spleen is normal in size with normal echotexture. The contours are smooth. The splenic vein and portal vein are normal in caliber.

Pancreas is slightly small, reveals thin inhomogenous paranchyma. the pancreatic duct is dilated.

Right Kidney measures _ x _. Left Kidney measures _ x _. Both the kidneys are normal in position, size, shape and contour. Cortical echogenicity is normal, corticomedullary differentiation is well maintained. No obvious calculus or mass is seen. No hydronephrosis noted.

Ureters are not dilated.

Urinary bladder appears normal. Wall thickness is normal. No evidence of calculus or mass is seen. Pre void is _ cc. Post void is _ cc.

The prostate is normal in size and echotexture measuring _.
Impression:
i) chronic pancreatitis, ii) cholecystitis and iii) grade ii fatty liver

Link

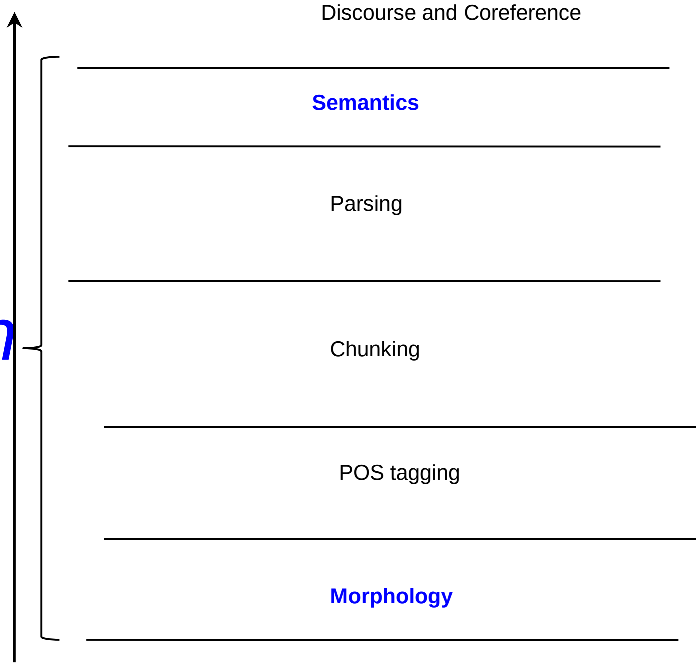
Recap

1-slide recap of 1st week

- Turing Test and Chinese Room Experiment
- Definition of NLP: art science and tech of NLU and NLG
- LINGUISTICS+PROBABILITY= NLP
- LLMs: what do they predict- language object and properties
- Comparisons with humans: size, energy requirement, carbon footprint
- Well known LLMs, MLMs, SLMs
- NL Stack

1-slide recap of week of 5th Aug

- Covered NLP stack with chatGPT's performance at each layer; chatGPT is an LLM based CAI
  - *To bank, I bank on the bank on the river bank*
- POS tag definition and argmax based formulation
- Should we apply Bayes theorem or not- discriminative (LHS of Argmax) vs. generative (RHS)
- HMM as the apt technique for POS



1-slide recap of week of 12th Aug

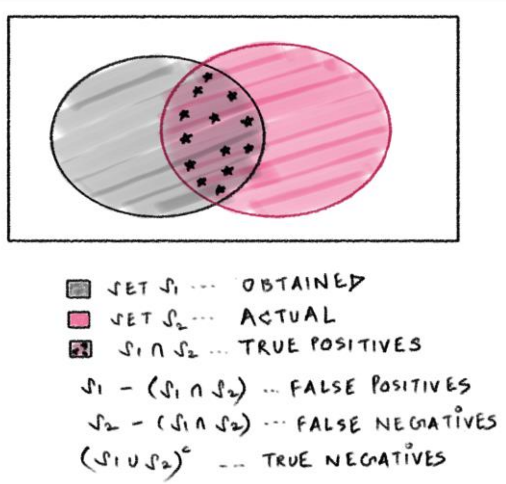
- Should I apply Bayes Rule: ***Cancer detection vs. Visa Application***
- Illustration of Viterbi with ***People laugh→ People_NN laugh_VB***
- Why is Viterbi linear time: ***Pruning of paths*** due to Markov Independence assumption
- What is "large" about large language models: ***probability of large (i.e. long sequences)***, plus the model having large number of parameters
- 3 tasks solved by HMM- ***Viterbi, Forward-Backward, Baum Welch***

1-slide recap of week of 19th Aug

- Discriminative POS tagging with Beam Search
- CRF and CRF based POS tagging

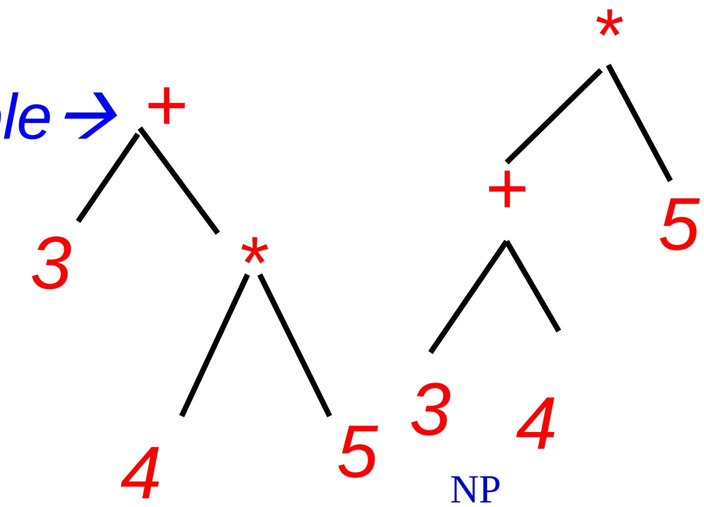
$$P(Y \mid X) = \frac{1}{Z(X)} \exp\left(\sum_{c} \varphi_c (Y_c, X_c) \right)$$

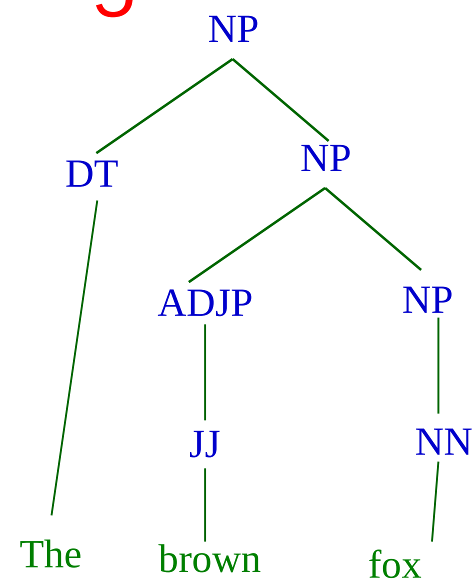
- Penn Tagset
- Evaluation metrics- *P, R, F*
- Inevitability of probabilistic POS tagging
- Assignment on POS- HMM, CRF, Benchmarking against ChatGPT



1-slide recap of week of 26th Aug

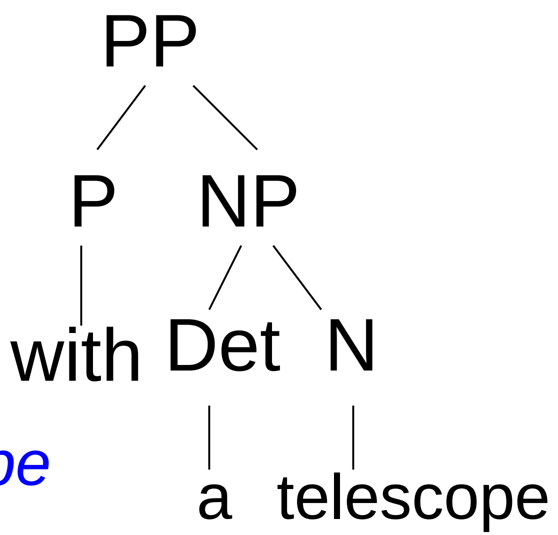
- Evidence of deep structure: *unlockable* → *un+lockable* or *unlock+able*
- **Structural ambiguity**
- Two kinds of parsing: CP and DP
- POS tagging facilitates chunking and parsing: short phrases and deep trees
- Parsing is important: e.g., aspect based SA
- Generative grammar, CFG
  - S→NP VP; NP→ NP PP
- Algorithmics of parsing: top down (TD), bottom up (BU), TDBU, CYK
- BI notation- vimp for NLP

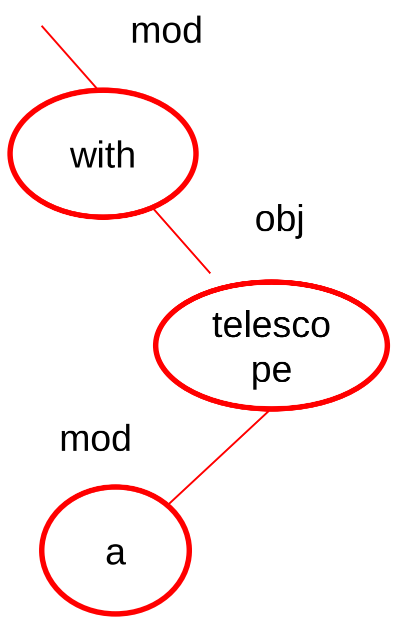




1-slide recap of week of 2nd Sep

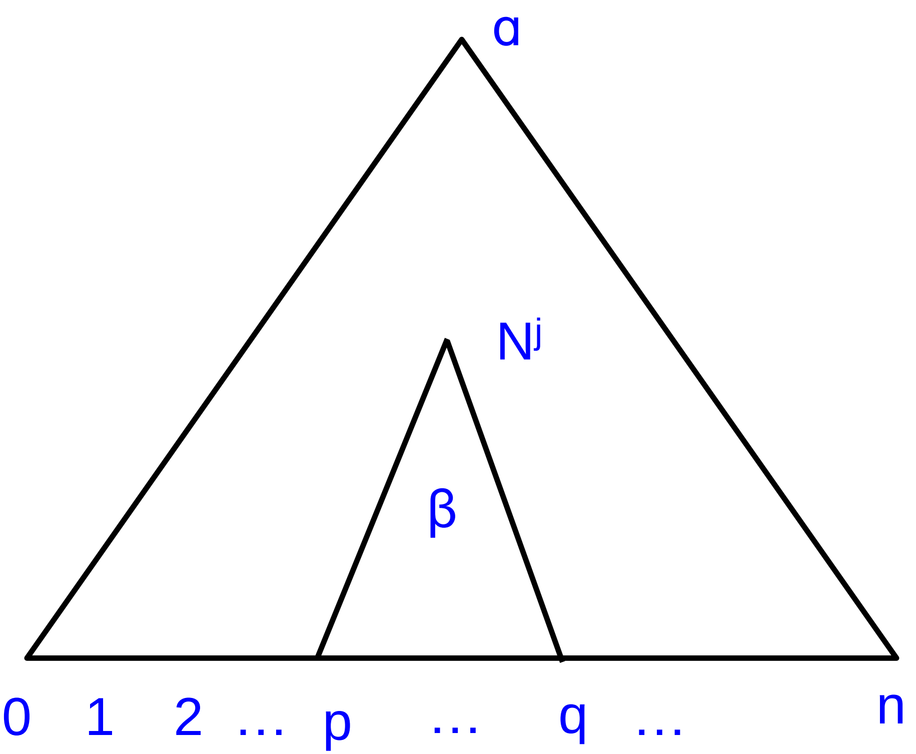
- CP: meaning of parent-child and sibling relationships- constituent; head-modifier
- Pre-modifier and post-modifier
- DP: Head-modifier expressed directly
- Ambiguity resolution by proximity: *telescope* example
- TD, BU, TDBY, CYK Parsing
- Notion of Domination: *X[p:q], the phrase/POS X dominates (generates) text segment from position p to position q*
- Probability of a CFG rule P(A→ B C) means P(B C|A)
- **Independence** of position, context and ancestry





1-slide recap of week of 9th Sep

- Domination
- Probabilistic Parsing:
  - *T*= argmax [P(T|S)]*
- Probability of a sentence
  - = *P(w0,l)=Σt P(t)*



$$\beta_j(p,q) = P(W_{p-q} \mid N^j_{pq})$$
$$= \sum_{k,r,l} P(N^j \to N^k\ N^l).P(W_{p-r} \mid N^k_{pr}).P(W_{r-q} \mid N^l_{rq})$$
$$= \sum_{k,r,l} P(N^j \to N^k\ N^l).\beta_k(p,r).\beta_l(r,q)$$

$$\delta_i(p,q) = \max_{j,r,k} P(N^i \to N^j\ N^k).\delta_j(p,r).\delta_k(r,q)$$

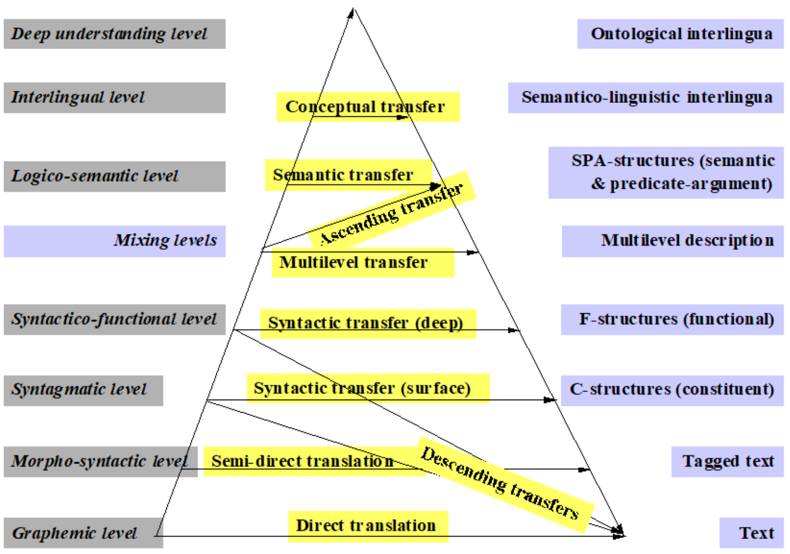
Stress Test for Parsing:
A very difficult parsing situation!-
the buffalo sentence

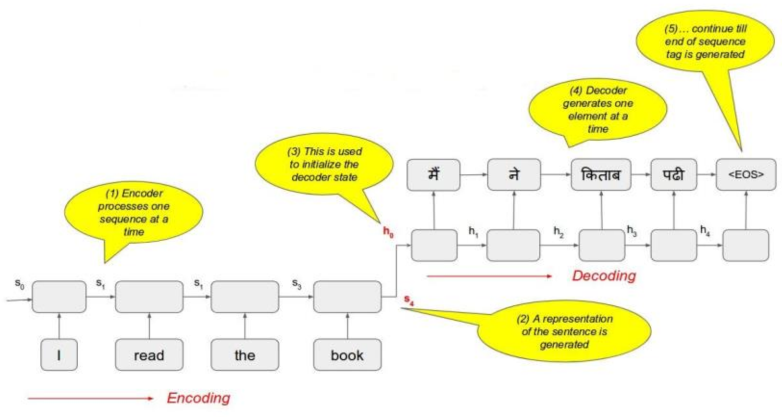
$$C_n = \frac{1}{n+1}\binom{2n}{n} = \prod_{k=2}^{n} \frac{k}{n+k}, \quad n \geq 0$$

1-slide recap of week of 23rd Sep

- Machine Translation: Definition, Paradigms
- Main Challenge: Language Divergence
- Vauquois Triangle as an abstraction of paradigms of MT
- A-T-G framework: Analysis Transfer Generation
- Encode Decoder Framework: basis of neural MT
- Data Driven MT- noisy channel model-

$$\bar{e} = \arg\max_{e} P(e|f)$$





1-slide recap of week of 30 Sep

- Language Divergence- Structural and Lexico-semantic
- Development of BLEU Score

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

$$p_n = \frac{\sum_{C \in \{Candidates\}} \sum_{n\text{-}gram \in C} Count_{clip}(n\text{-}gram)}{\sum_{C' \in \{Candidates\}} \sum_{n\text{-}gram' \in C'} Count(n\text{-}gram')}$$

- Another competing metric: Recall Oriented-Rouge score

$$\text{ROUGE-N} = \frac{\sum_{S \in \{ReferemceSummaries\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{ReferenceSummaries\}} \sum_{gram_n \in S} Count(gram_n)}$$

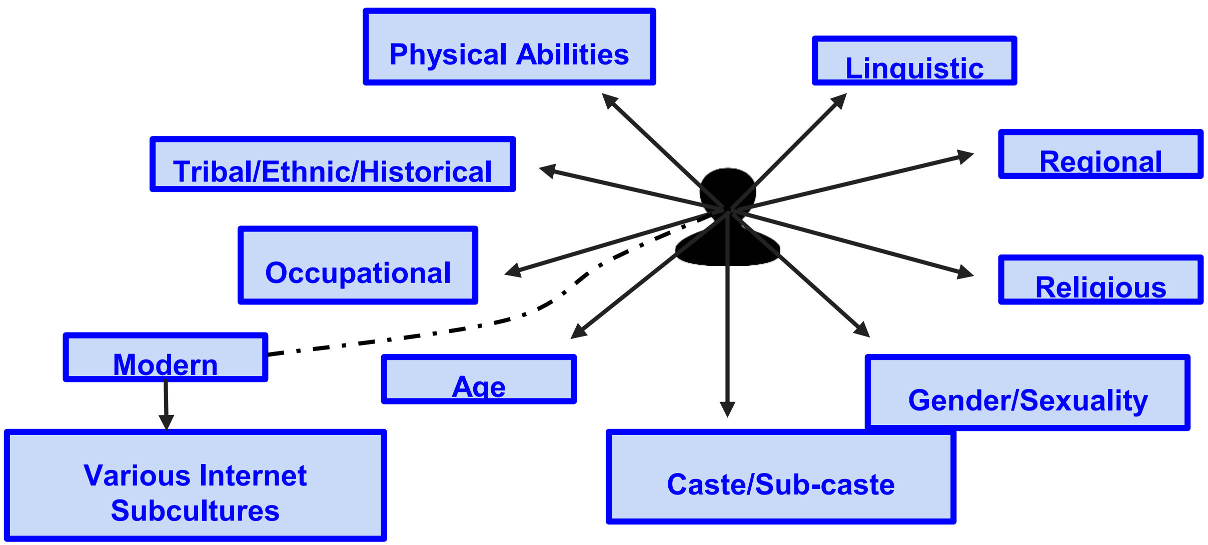
- The need for probability: Bridge Problem

1-slide recap of week of 7th Oct

- Bias: < S, L, T, C, R >
- Stereotyping

Fact $\xrightarrow{\textit{Overgeneralized}}$ Stereotype $\xrightarrow{\textit{Opinionated}}$ Bias

- Bias types



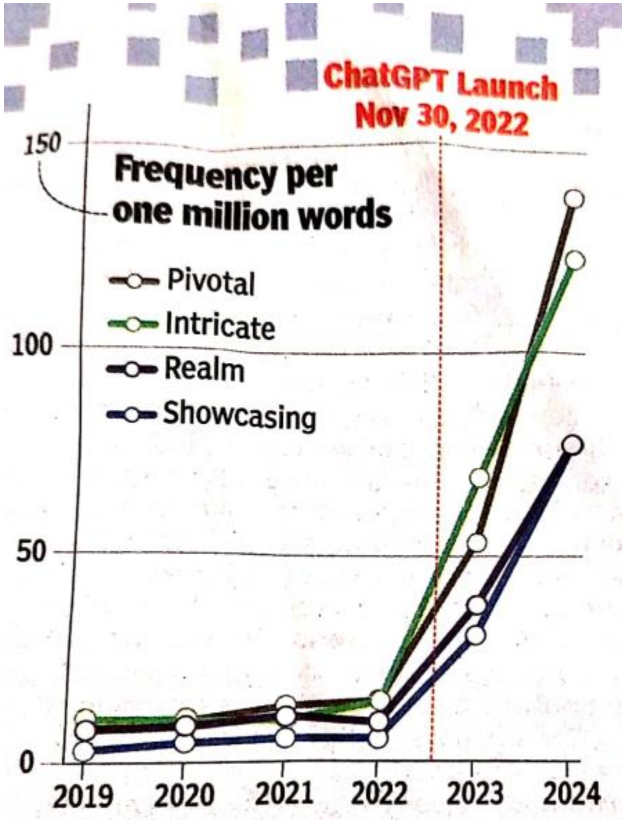
- Role of probability: the bridge problem- deterministic (conservative, liberal); probabilistic

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

1-slide recap of week of 21st Oct

- ChatGPT give-aways: ‘delve’, ‘additionally’, ‘additionally’, ‘nevertheless’, ‘a testament to…’
- **Pramana**- means of acquiring knowledge: **Pratyaksha** (perception), **Anumana** (inference), **Upamana** (comparison



- **Sabda** (verbal testimony**), Arthapatti** (postulation), **Anupalabdhi** (non-perception)
- CLT, LoLN

1-slide recap of week of 28th Oct (last week)

- NER contd.- ENAMEX, NUMEX, TIMEX
- Computation of NER- Supervised (classification), unsupervised (clustering), semi-supervised (label automatically-correct manually cycle)
- Discrimiasntive models very successful for NER (CRF, MEMM, NN)

- Knowledge Networks- wordnet, conceptnet etc.
- Knowledge triples augment power of DNNs
- Application of KG in health AI

Thank you and all the best

https://www.cse.iitb.ac.in/~pb/

https://www.cfilt.iitb.ac.in/