

# CS772: Deep Learning for Natural Language Processing (DL-NLP)

**Conversational AI, Closure**

Pushpak Bhattacharyya

Computer Science and Engineering  
Department

IIT Bombay

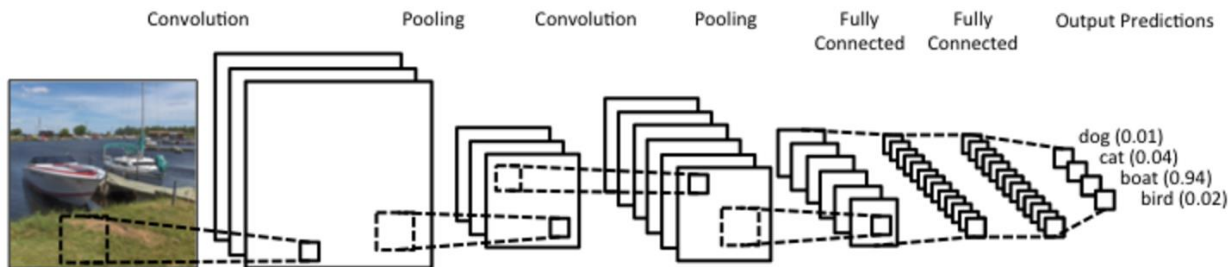
*Week 15 of 15apr24*

# 1-slide recap

## Three generations on LM:

1. Grammar Rules: Context Free Grammars, 2a. N-grams, 2b. Probabilistic Context Free Grammars, 3. Neural (solves the CURSE OF DIMENSIONALITY problem)

**CNNs:** stages of conv+RELU+Pooling layers terminating in FFNN and Softmax; very successful in Image and Vision; Partially in Language



**Stable Diffusion:** competitor to LLM. **LLMs:** high quality generation of text;

**Diffusion:** high-resolution image synthesis



# Conversational AI

# Ubiquity and impact of Chatbots

Credit: Economic Times, 18apr24

# The onward March of CAI

- The global AI Chatbot market will be worth \$27.2 billion by 2030 (Report from Grand View Research Inc)
- Chatbot personalization will significantly impact business outcome (Gartner report)
- Chinese company Baidu's chatbot gathered 200 million users
- After the success of chatGPT, enterprise chatbots are mushrooming
  - Shopify- helps build online store; Instacart- delivery company
  - Mercari- resale platform; Carrefour- retailer
  - Kering- owns Gucci and Balenciaga (high end clothing)

# CAI: ever increasing expectation

- CAI (*aka chatbot*) used for checking flight schedule, calling cabs, book hotel, bank statement.... (the list is endless)
- SoA- “well trained chatbots can handle 60-80% queries w/o human help” (Gupship CPO)
- Indigo’s Whatsapp Bot Dottie- contained 86% of chats and transferred 14% to experts
- 6Eskai- handles 91% of chats; number of experts needed down to 69 from 107 before

# Pifalls

- Data Residency
- Data Privacy
- Complexities of real time information
- Complexities of novel data streams
- Hallucination
- Lack of emotional intelligence
- Lack of empathy (serious in case of complaints)

## An AI Chatbot Was Caught Telling Businesses To Break The Law

It still gets to keep its job

**AN ARTIFICIAL INTELLIGENCE**-powered chatbot created by New York City to help small business owners dispensed bizarre advice recently, misstating local policies and advising companies to violate the law. But the city left the tool on its official website. Mayor Eric Adams defended the decision even as he acknowledged the chatbot's answers were "wrong in some areas".

In responses to questions, the chatbot falsely suggested it is legal for an employer to fire a worker who complains about sexual harassment, doesn't disclose a pregnancy or refuses to cut their dreadlocks.

At times, the bot's answers veered into the absurd. Asked if a restaurant could serve cheese nibbled on by a rodent, it responded: "Yes, you can still serve the cheese to customers if it has rat bites," before adding that it was important to assess "the extent of the damage caused by the rat" and to "inform customers about the situation".

The pitfalls of New York's chatbot should serve as a cautionary tale for other cities, said Suresh Venkatasubramanian, director of the Center for Technological Responsibility, Reimagination, and Redesign at Brown University, US.

"It should make cities think about why they want to use chatbots, and what problem they are trying to solve," he wrote in an email. "If the chatbots are used to replace a person, then you lose accountability while not getting anything in return."

Use with caution



# Solution to pitfalls

- Layer of human supervision
- **Make use of multiple models like BARD, ChatGPT etc.**
- Hallucination
- Human Bot synergy: when to pass control
- Zomato's chatbot- handle queries to a point and then pass control to human
- ***EMERGING WORLD ORDER: HUMAN MACHINE SYNERGY***

# AI chatbots compared: Bard vs. Bing vs. ChatGPT

<https://www.theverge.com/2023/3/24/23653377/ai-chatbots-comparison-bard-bing-chatgpt-gpt-4>

# Comparison: Chatbots

Google's Bard (<https://bard.google.com/>),

Microsoft's Bing  
(<https://www.theverge.com/2023/3/24/23653377/ai-chatbots-comparison-bard-bing-chatgpt-gpt-4>),

OpenAI's ChatGPT (<https://chat.openai.com/chat#>)

# Comparison: what is compared

Range of questions spanning  
common requests

from holiday tips to  
gaming advice to  
mortgage calculations.

Summary from the blog: “*ChatGPT is the most verbally dexterous, Bing is best for getting information from the web, and Bard is... doing its best*”

## Underlying systems

- OpenAI's ChatGPT: uses GPT-4
- Bing: GPT4; but Bing has other abilities
  - Can generate images
  - Can access the web
  - Offers sources for its responses
- However, “OpenAI announced it’s launching plug-ins for ChatGPT that will allow the chatbot to also access real-time data from the internet.”

“Give me a recipe for chocolate cake”

- Cake recipes offer room for creativity
- “Shift around the ratio of flour to water to oil to butter to sugar to eggs, and you’ll get a slightly different version of your cake: maybe drier, or moister, or fluffier”
- Chatbots combine different recipes to achieve a desired effect

## Recipe for chocolate cake: ChatGPT, the best

- “chose a chocolate cake recipe from one site, a buttercream recipe from another, shared the link for one of the two, and reproduced both of their ingredients correctly. It even added some helpful instructions, like suggesting the use of parchment paper and offering some (slightly rough) tips on how to assemble the cake’s layers, neither of which were found in the original sources. This is a recipe bot I can trust!”

## Recipe for chocolate cake: Bing

- “Bing gets in the ballpark but misses in some strange ways. It cites a specific recipe but then changes some of the quantities for important ingredients like flour, although only by a small margin. For the buttercream, it fully halves the instructed amount of sugar to include. Having made buttercream recently, I think this is probably a good edit! But it’s not what the author called for.”



## Recipe for chocolate cake: Bard

- “Bard, ... makes some changes that meaningfully affect flavor: it swaps buttermilk for milk and coffee for water. Later on, it fails to include milk or heavy cream in its buttercream recipe, so the frosting is going to end up far too thick. The buttercream recipe also seems to have come from an entirely different source than the one it cited. If you follow ChatGPT or Bing, I think you’d end up with a decent cake. But right now, it’s a bad idea to ask Bard for a hand in the kitchen”

# Linguistic Foundations of CAI (*aka* ChatBot)

# Gricean Maxims: Cooperative Principle in Conversation (Wikipedia)

- **Quantity, Quality, Relation, and Manner**
- Paul Grice, philosopher of language
- *“Make your contribution such as is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged”.*
- Captures the LINK between utterances

# Examples of Gricean Maxims (1/3)

- **Quantity:**
  - Question: *Do you know where the IITB library is*
  - Answer: *Yes I do, it is near the maths building*
  - (violation): *Yes, I do*
- **Quality:**
  - Question: *Do you know where the IITB library is*
  - Answer: *Yes I do, it is near the maths building*
  - (violation, if the answerer purposefully misleads):  
*Yes, it is near the main gate*

# Examples of Gricean Maxims (2/3)

- **Relation:**

- Question: *Do you know where the IITB library is*
- Answer: *Yes I do, it is near the maths building*
- (violation): *Mumbai is a big city*
- (violation of a different kind and degree): *IITB campus is big! (so I do not know)*
- (flouting, but ok): *I saw Ram enter the library from the maths building*

- **Manner:**

- Question: *Do you know where the IITB library is*
- Answer: *Yes I do, it is near the maths building*
- (violation, if the answerer purposefully misleads): *Yes, it is near the main gate*

## Examples of Gricean Maxims (3/3)

- ***Manner:***

- Question: *Do you know where the IITB library is*
- Answer: *Yes I do, it is near the maths building*
- (violation): *Up yonder in the citadel of learning where polynominals are the bread and operators are the butter and where Hardy and Ramanujam permeate the atmosphere, thither will thee find the storehouse of what bibliophiles love (maybe used for humorous effect)*

# What does ChatGP “think”? (1/2)

If for the question, "where is the library?" someone answers "Up yonder in the citadel of learning where polynominals are the bread and operators are the butter and where Hardy and Ramanujam permeate the atmosphere, thither will thee find the storehouse of what bibliophiles love", what would you say?

# What does ChatGP “think”? (2/2)

Based on the answer given, it seems that the person is using figurative and poetic language to describe the location of the library. They are painting a vivid picture of the library being located in a place of high learning where mathematics and academic excellence are highly valued.

However, their answer may not be helpful if the person asking the question simply needs a straightforward answer to their question. In such a case, it might be more helpful to ask the person to provide a more direct and concrete answer to the question of where the library is located.



# 3 stages of LLM based CAI

- Generative Pretraining (GP)
- Supervised Fine Tuning (SFT)
- Reinforcement Learning based on Human Feedback (RLHF)

Enter Pragmatics

# Modeling

**$P(e)$ : “language”  
model**

$$\begin{aligned} e^* &= \arg \max_e P(e \mid f) \\ &= \arg \max_e [P(e)P(f \mid e)] \end{aligned}$$

- Dialogue Act Classification (DAC):  $f \rightarrow$  *Dialogue Sequence*,  $e \rightarrow$  *Dialogue turn labels*
- Dialogue Intent:  $f \rightarrow$  *dialogue sequence*,  $e \rightarrow$  *dialogue turns with Intent like ‘question’, ‘elaboration’, ‘affirmation’, ‘command/request’ etc.*

# Example of sequence of dialogue and dialogue acts

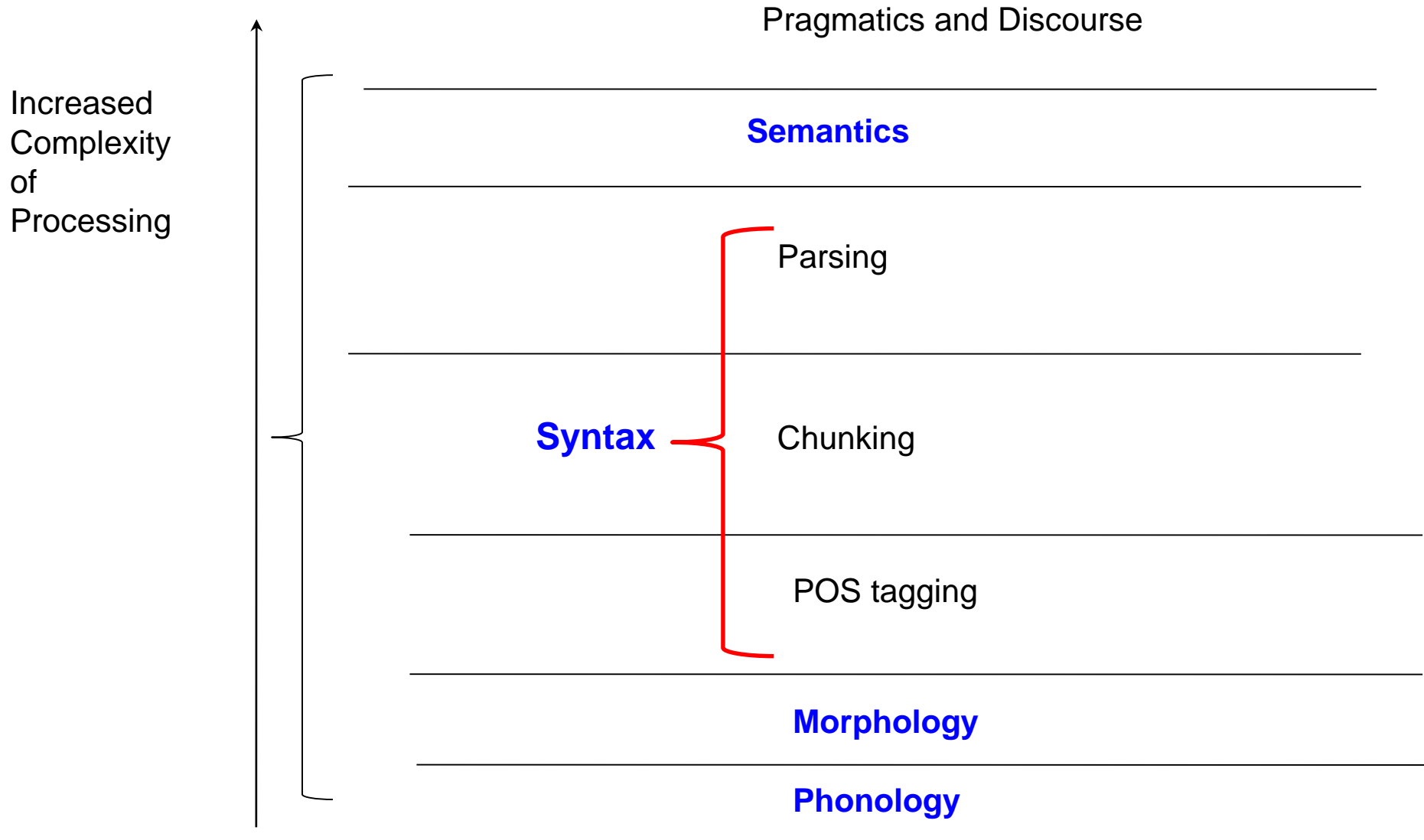
Fragment of a labeled conversation (from the Switchboard corpus).

Speaker	Dialogue Act	Utterance
A	YES-NO-QUESTION	So do you go to college right now?
A	ABANDONED	Are yo-,
B	YES-ANSWER	<i>Yeah,</i>
B	STATEMENT	<i>it's my last year [laughter].</i>
A	DECLARATIVE-QUESTION	You're a, so you're a senior now.
B	YES-ANSWER	<i>Yeah,</i>
B	STATEMENT	<i>I'm working on my projects trying to graduate [laughter].</i>
A	APPRECIATION	Oh, good for you.
B	BACKCHANNEL	<i>Yeah.</i>
A	APPRECIATION	That's great,
A	YES-NO-QUESTION	um, is, is N C University is that, uh, State,
B	STATEMENT	<i>N C State</i>
A	SIGNAL-NON-UNDERSTANDING	What did you say?
B	STATEMENT	<i>N C State</i>

# Digression: What if there are many “abandoned”s

- Raise a flag
- Linguistic Limitation
- Or, Mental Health Problem
- Mental health doctors would like help of chatbots that can give preliminary help to mental health patients by engaging in a dialogue
- Or, they can do a preliminary screening based on disfluencies, abandoned statements and so on

# NLP (and linguistics) Layers

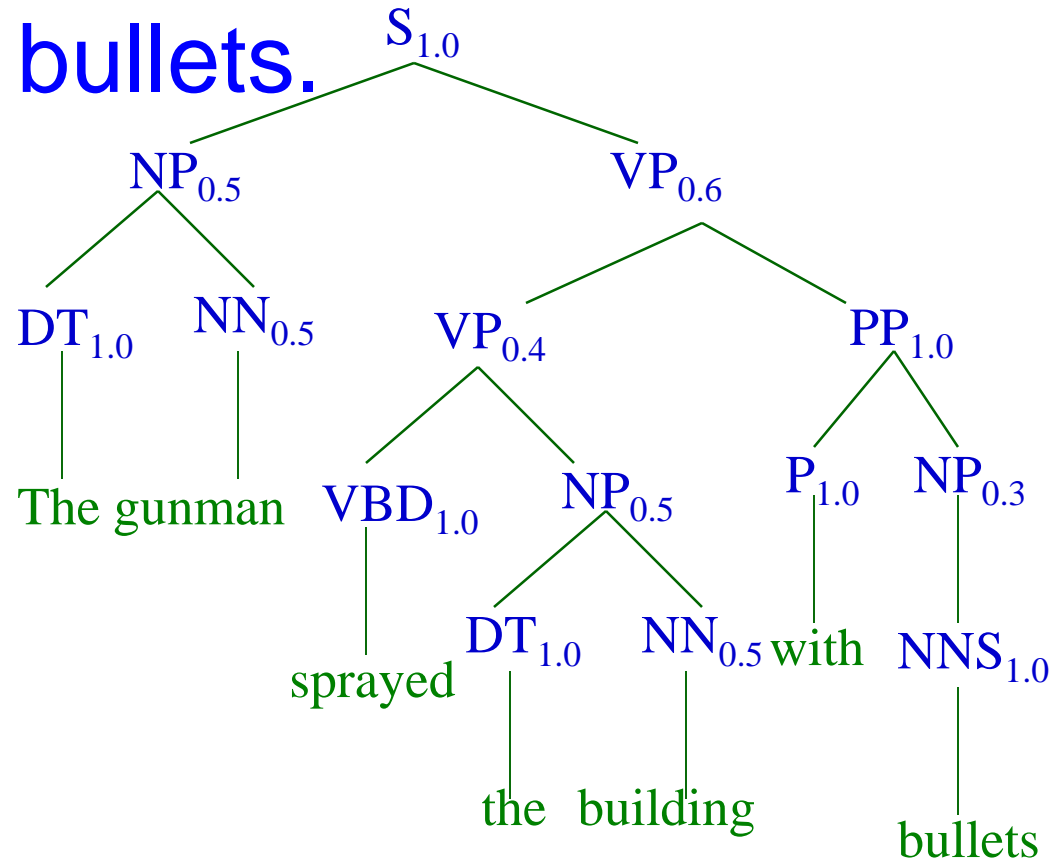


Which meaning of “*the gunman sprayed the building with bullets*” is more likely?:  
Semantics-Pragmatics interface

- The gunman has the bullets
- The building has the bullets
- The former is more likely
- Corroborated by data

# Parse $t_1$

- The gunman sprayed the building with bullets.

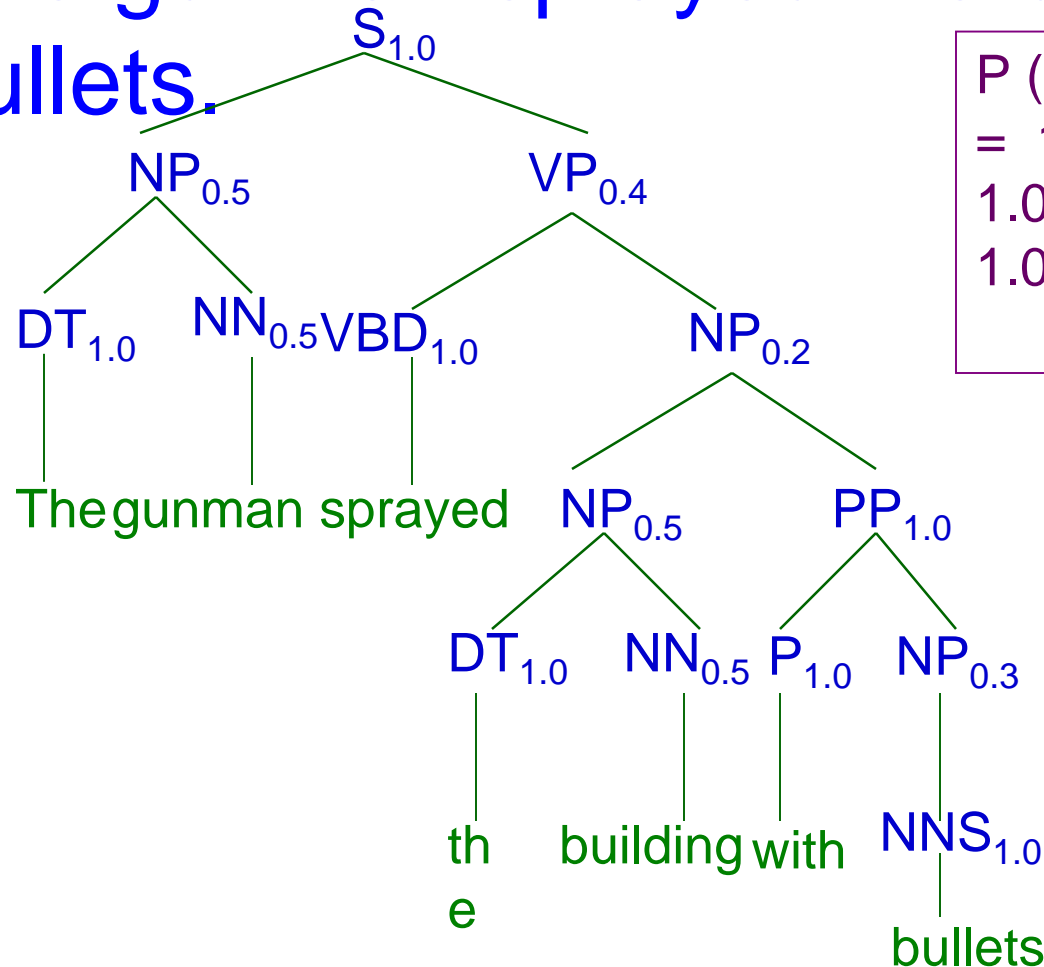


$$\begin{aligned}
 P(t_1) &= 1.0 * \\
 &0.5 * 1.0 * 0.5 * 0.6 * 0.4 * 1.0 \\
 &* 0.5 * 1.0 * 0.5 * 1.0 * 1.0 * \\
 &0.3 * 1.0 \\
 &= 0.00225
 \end{aligned}$$



# Parse $t_2$

- The gunman sprayed the building with bullets.



$$\begin{aligned}
 P(t_2) &= 1.0 * 0.5 * 1.0 * 0.5 * 0.4 * \\
 &1.0 * 0.2 * 0.5 * 1.0 * 0.5 * \\
 &1.0 * 1.0 * 0.3 * 1.0 \\
 &= 0.0015
 \end{aligned}$$

# Meaning of “Pragmatics”

- Meaning in a Context
- To be contrasted with **Lexical Semantics**- word meanings
- **Sentential semantics**- truth value of a sentence and entailment (*a la* Montague)
  - *Today, the sky is blue*  $\models$  *Today, there is no rain*
- Pragmatics is extra-sentential
- Arises due to the limitations of lexical and formal semantics

# Example of Pragmatics

- Dialogue or Conversation Setting
  - Speaker A: *shall we go for a walk?*
  - Speaker B: *It is raining outside*
- Implication: answer is NO

# More examples

- (Person entering a room where there is an AC running): *Isn't it cold here?*
- Implication- stop or reduce the AC.

## Another dimension of pragmatics

- Pragmatics is language in use
- Priest conducting a wedding ceremony in church: *I pronounce you man and wife*
- Leads to solemnisation of marriage

# Elements of Pragmatics (1/2)

- Deixis (literally, 'pointing with words': temporal- *now, then*; spatial- *here, there*; personal- *I, you, he, they*; definite-indefinite- *this, that, those*)
- Presupposition: (*untie the shoe* → presupposes *the shoe was tied before*)

# Elements of Pragmatics (1/2)

- Speech Acts: (*I pronounce you man and wife*)- **locutionary, illocutionary, and perlocutionary**
- Implicatures: (*A: shall we go for a walk? B: It is raining outside*)
- Politeness: (*close the door* → *please close the door* → *can you close the door* → *would you mind closing the door*)
- Information Structure: ordering of information (?? *The table is under the flower pot*- odd: smaller object first mention)

# Sanskrit Tradition

- **Shabdshakti (power inherent in word)**
  - Abhidha, Lakshana, Vyanjana
- **Meaning of Hall:**
  - *The hall is packed* (avidha)
  - *The hall burst into laughing* (lakshana)
  - *The Hall is full* (unsaid: and so we cannot enter) (vyanjana)



# Abhidha, Lakshana, Vyanjana giving rise to

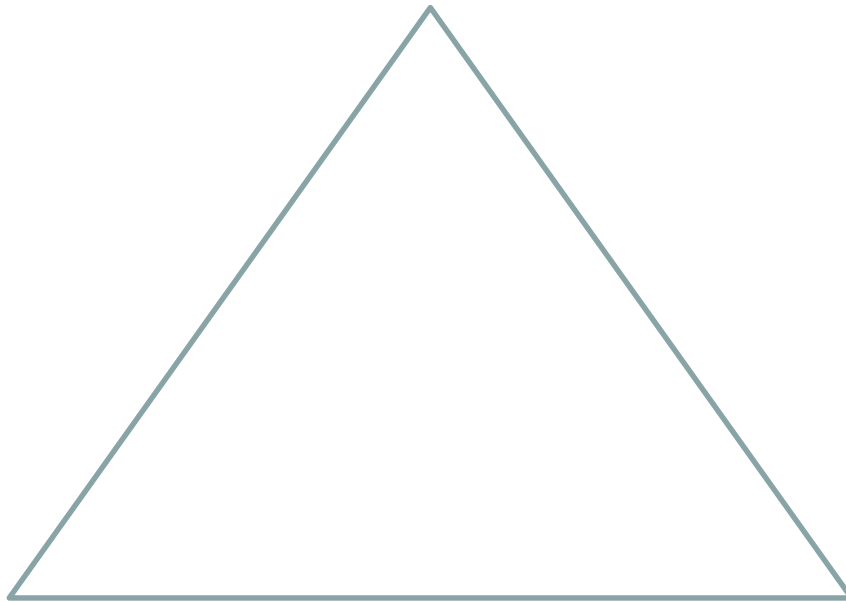
- Vachyarthā, Lakshyarthā, Vyangaarthā
- **“Gangaa”:**
  - *vaachyaarthā*: The river Gangaa (due to abhidhaa)
  - *lakshyaarthā*: *gangaayaaM ghoshaH*: the house on river gangaa, meaning “on the bank of” (due to lakshanaa)
  - *Vyangaarthā*: the house will have nice view, breeze etc. (unsaid) (due to vyanjana)

# Crucial to Pragmatics

- Sentence vs. Utterance
- Semantics + Intent → Pragmatics

# The Trinity of Pragmatics

**Linguistic Expression**



**Speaker**

**Hearer**

# Computational Perspective: Conversational AI

# A dialogue

S:	How may I help you?
U:	<i>can you tell me how much it is to tokyo?</i>
S:	You want to know the cost of a call?
U:	<i>yes that's right</i>
S:	Please hold on for rate information.
S:	How may I help you?
U:	<i>what's the area code for chicago?</i>
S:	You want an area code.
U:	<i>(silence)</i>
S:	The area code for Chicago is 312.

Figure 22.2 The call routing domain: two conversations between a user (U) and the AT&T How May I Help You (HMIHY) system (S) (Gorin et al., 1997).

# Modeling

$$\begin{aligned} e^* &= \arg \max_e P(e | f) \\ &= \arg \max_e [P(e)P(f | e)] \end{aligned}$$

**$P(e)$ : “language” model**

- Dialogue Act Classification (DAC):  $f \rightarrow$  *Dialogue Sequence*,  $e \rightarrow$  *Dialogue turn labels*
- Dialogue Intent:  $f \rightarrow$  *dialogue sequence*,  $e \rightarrow$  *dialogue turns with Intent like ‘question’, ‘elaboration’, ‘affirmation’, ‘command/request’ etc.*

# Dialogue Based Computation

Zihao He, Leili Tavabi, Kristina Lerman, and Mohammad Soleymani.  
2021. Speaker Turn Modeling for Dialogue Act Classification. In Findings of the Association for Computational Linguistics: EMNLP 2021, pages 2150–2157, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tulika Saha, Aditya Patra, Sriparna Saha and Pushpak  
Bhattacharyya, Towards Emotion-aided Multi-modal Dialogue Act Classification, Association of Computational Linguistics Conference (**ACL 2020**), Seattle USA, 5-10 July, 2020.

# Dialogue Act Classification and Sentiment

Tulika Saha, Aditya Patra, Sriparna Saha and Pushpak  
Bhattacharyya, *Towards Emotion-aided Multi-modal Dialogue Act  
Classification*, Association of Computational Linguistics Conference  
(**ACL 2020**), Seattle USA, 5-10 July, 2020.



# Contributions (1/2)

- Investigated the role of **emotion and multi-modality** in determining **DAs** of an utterance
- Created of a novel dataset, **EMOTyDA**, containing emotion-rich videos of dialogues collected from various open-source datasets manually annotated with DAs
- Given an **attention** based (self, inter-modal, inter-task) multi-modal, multi-task framework for joint optimization of DAs and emotions

## Contributions (2/2)

- Multi-modality and multi-tasking **boosted** the performance of DA identification compared to its unimodal and single task DAC variants
- Plan in future to incorporate conversation history, speaker information, fine-grained modality encoding to predict DA with more accuracy and precision

# Dialogue Act Classification (DAC) and Multimodality

- DAC → Intent
- Each turn primarily *a question, a statement, or a request for action*
- Prior work: Jurafsky et al. (1997), Stolcke et al (2000), Verbree et al (2006), Kalchbrenner and Blunsom (2013), Liu et al. (2017), Ortega et al (2019), Saha et al (2019) etc.

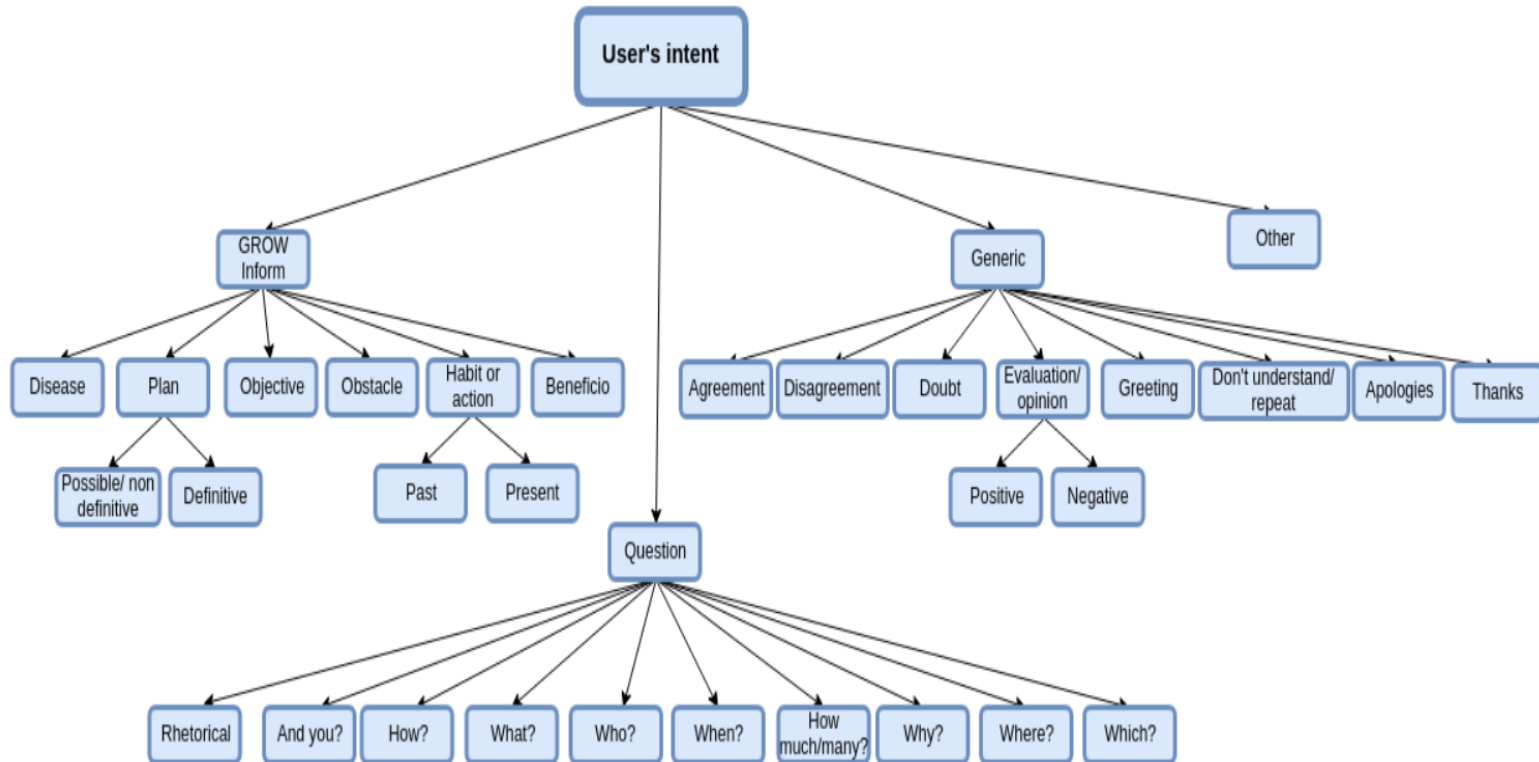
# Emotion and Dialogue

- Non-verbal features
  - change of tone, facial expressions
  - provide beneficial cues to identify DAs
  - Emotion aided multi-modal DAC
  - “ha ji ha” in Hindi can denote agreement (statement) or disagreement (*sarcasm*)
- Contributions
  - Emotion-aware DA dataset (EMOTyDA)
  - Multi-modal, multi-task DNN for DAs and emotions identification
  - Showed: Multi-modality and multi-tasking DAC better than uni-modal and single task DAC

# Prior Work

- Dating back to late 1990's (Reithinger and Klesen, 1997), (Stolcke et al., 1998) and early 2000's (Stolcke et al., 2000), (Grau et al., 2004)
- DA specification and taxonomy got established
- ***DAC = sequence labelling***
- DL Based:
  - Khanpour et al. (2016)- stacked LSTM
  - Kumar et al (2018)- hierarchical bi-LSTM and CRF
  - Raheja and Tetreault (2019)- contextual self-attention framework fused with hierarchical recurrent units
  - Yu et al (2019)- CNN

# Dialogue Taxonomy



# Example of dialogue and dialogue acts

Fragment of a labeled conversation (from the Switchboard corpus).

Speaker	Dialogue Act	Utterance
A	YES-NO-QUESTION	So do you go to college right now?
A	ABANDONED	Are yo-,
B	YES-ANSWER	<i>Yeah,</i>
B	STATEMENT	<i>it's my last year [laughter].</i>
A	DECLARATIVE-QUESTION	You're a, so you're a senior now.
B	YES-ANSWER	<i>Yeah,</i>
B	STATEMENT	<i>I'm working on my projects trying to graduate [laughter].</i>
A	APPRECIATION	Oh, good for you.
B	BACKCHANNEL	<i>Yeah.</i>
A	APPRECIATION	That's great,
A	YES-NO-QUESTION	um, is, is N C University is that, uh, State,
B	STATEMENT	<i>N C State</i>
A	SIGNAL-NON-UNDERSTANDING	What did you say?
B	STATEMENT	<i>N C State</i>

# Dataset: EMOTyDA

- Short videos of dialogue conversations manually annotated with its DA along with its pre-annotated emotions
- Studied existing emotion recognition data sources
  - Youtube (Morency et al 2011), MOUD (P´erez-Rosas et al., 2013), **IEMOCAP** (Busso et al., 2008), ICT-MMMO (Wollmer et al., 2013), CMU-MOSI (Zadeh et al., 2016), CMU-MOSEI (Zadeh et al., 2018) and **MELD** (Poria et al., 2019)
- Zeroed down on IEMOCAP and MELD datasets
- Manually annotated for DAs



# Data Annotation (1/2)

- SWBD-DAMSL tag-set consisting of 42 DAs (Jurafsky, 1997) for task-independent dyadic conversation such as SWBD corpus used
- Out of the 42 DAs of the SWBD-DAMSL tag-set, 12 most commonly occurring tags selected
- 12 frequently occurring chosen tags are
  - Greeting (g), Apology (ap), Command (c),
  - Question (q), Answer (ans), Agreement (ag),
  - Disagreement (dag), Statement-Opinion (o), Statement-Non-Opinion (s),
  - Acknowledge (a), Backchannel (b) and Others (oth).

## Data Annotation (2/2)

- A subset of 1039 dialogues from MELD amounting to 9989 utterances and the entire IEMOCAP dataset of 302 dialogues amounting to 9376 utterances
- Three annotators graduate in English were assigned to annotate the utterances
- The inter-annotator score with more than 80% was considered as reliable agreement
- Mapped the *joy* tag of MELD to *happy* tag of the IEMOCAP

# Particulars of EMOTyDA

1341 dyadic and multi-party conversations resulting in a total of 19,365 utterances or annotated videos with the corresponding DA and emotion tags considering the dialogue history.

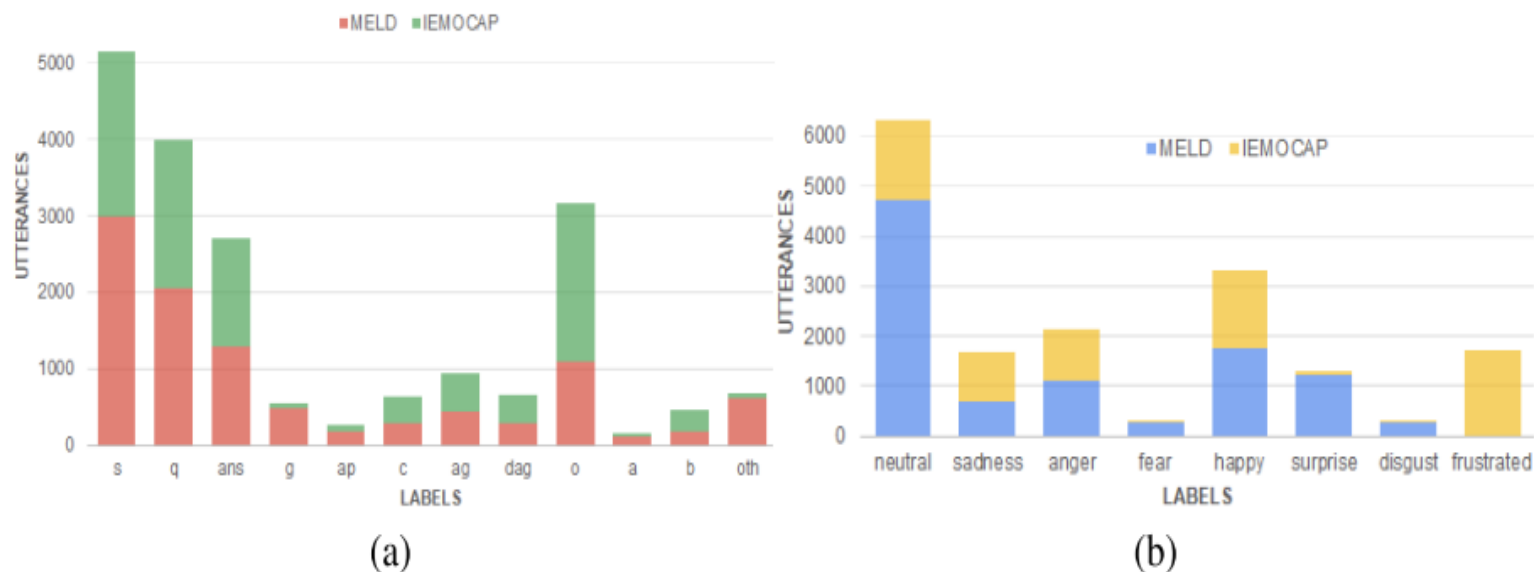


Figure 1: Statistics across the datasets : (a) Distribution of DA labels, (b) Distribution of emotion labels

# Case for modality

## Utterance

- 1) **Phoebe:** Fine! Then you tell Roger because he was really looking forward to this!

- **Text** : suggests agreement or opinion
- **Audio** : commanding tone
- **Video** : furious



- 2) **M\_1:** That's very amusing indeed.

- **Text** : agreement
- **Audio** : sarcastic tone
- **Video** : slight anger



(a)

## Utterance

- 1) **Monica:** I can't leave it! You gouged a hole in my dingy floor.

DA: disagreement

- 2) **M\_2:** Well, you know I appreciate you coming over and talking to me, I mean it definitely helps.

DA: acknowledge

## Emotion

anger

sad

(b)

Figure 2: (a) Incongruent modalities in DAC, (b) Importance of emotion in DAC.

# Technique: Feature Extraction (1/2)

- **Text**: transcripts of each video; concatenation of pretrained GloVe (Pennington et al., 2014)
- **Audio**: OpenSMILE (Eyben et al., 2010), an open source software used
  - 12 Mel-frequency coefficients, glottal source parameters (Drugman et al., 2011), maxima dispersion quotients (Kane and Gobl, 2013),
  - several low-level descriptors (LLD) such as voice intensity, MFCC, voiced/unvoiced segmented features (Drugman and Alwan, 2011), pitch and their statistics (for example, root quadratic mean, mean etc.), voice quality (for example, jitter and shimmer), etc.

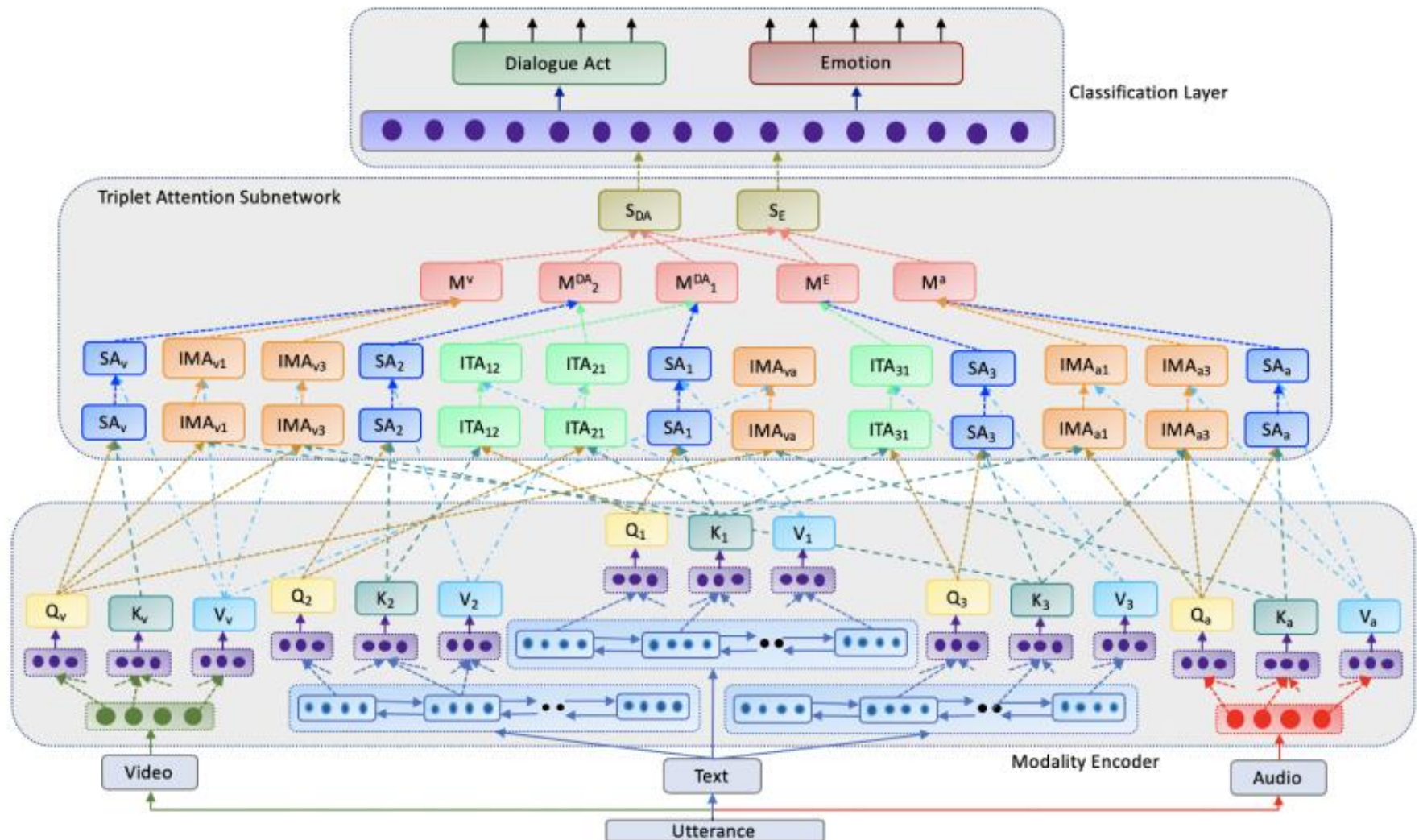
# Technique: Feature Extraction (2/2)

- **Audio (cntd):**
  - Extracted features concatenated together to form a  $dq = 256$  dimensional representation for each window. The final audio representation of
  - each utterance (A) is obtained by concatenating  $dq$  for every window
- **Video:**
  - ImageNet (Deng et al., 2009) pretrained ResNet-152 (He et al., 2016) used
  - Visual representation of each utterance (F ) is obtained by concatenating the obtained  $df = 4096$  dimensional feature vector for every frame

# Network Architecture

- Three main components:
  - (i) Modality Encoders (ME) which typically takes as input the uni-modal features and outputs the modality encodings,
  - (ii) Triplet Attention Subnetwork (TAS) that encompasses self, inter-modal and inter-task attention and
  - (iii) Classification layer that encompasses outputs of both the tasks (DAC and ER) to be learned jointly conditioned on the output of the TAS

# Network Diagram



SA, IMA, ITA represents self, inter-modal and inter-task attentions respectively.



# Results

Modality	Dataset											
	EMOTyDA:dyadic				EMOTyDA:multparty				EMOTyDA			
	DA		DA + ER		DA		DA + ER		DA		DA + ER	
	Acc.	F1-score	Acc.	F1-score	Acc.	F1-score	Acc.	F1-score	Acc.	F1-score	Acc.	F1-score
Text (T)	63.75	60.67	65.23	62.35	46.20	39.23	48.90	41.10	53.56	49.17	53.02	50.22
Audio (A)	32.06	24.95	35.42	38.92	25.76	19.45	26.58	21.01	27.13	23.09	28.65	24.87
Video (V)	35.94	29.71	36.88	30.34	27.23	20.26	28.12	21.03	30.16	26.85	32.09	27.73
T + A	65.43	60.67	<b>66.98</b>	<b>62.08</b>	47.17	40.30	<b>49.42</b>	<b>41.69</b>	54.12	50.00	<b>56.62</b>	<b>51.99</b>
A + V	38.59	34.98	40.07	36.00	27.91	22.76	28.95	23.89	32.09	28.86	33.76	29.13
T + V	67.12	64.14	<b>70.55</b>	<b>68.12</b>	49.80	41.90	<b>51.00</b>	<b>44.52</b>	57.31	53.20	<b>60.88</b>	<b>57.96</b>
T + A + V	66.35	62.30	<b>69.45</b>	<b>67.00</b>	49.02	41.00	<b>50.65</b>	<b>44.00</b>	56.77	52.09	<b>59.86</b>	<b>56.05</b>
T + V (emotional cue)	65.26	60.20	-	-	46.88	39.70	-	-	54.31	50.02	-	-

Table 1: Results of the various models. Higher the values of accuracy and F1-score, better the performance of the corresponding model. All the reported results are statistically significant

Dyadic- IEMOCAP; multparty- MELD;  
EMOyDA- combined IEMOCAP and MELD

# Case Study

Utterance	True Label	MT(T+V)	ST (T+V)
She is not Larry's girl	dag	dag	s
I know, it was amazing! I mean, we totally nailed it, it was beautiful.	ag	ag	o
Then why is she still single?,New York is full of men.,Why hasn't she married?	o	s	q
Probably a hundred people told her she's foolish, but she's waited.			
God, I feel so guilty about Ross.	ap	ap	s

Table 2: Sample utterances with its predicted labels for the best performing multi-task (MT) (T+V) model and its single task (ST) DAC variants; These examples show that ER as an auxiliary task helps DAC for better performance in MT.

Greeting (g), Apology (ap), Command (c),  
Question (q), Answer (ans), Agreement (ag),  
Disagreement (dag), Statement-Opinion (o), Statement-Non-Opinion (s),  
Acknowledge (a), Backchannel (b) and Others (oth).

Closure

# 1-slide recap of week 1

- Nature of language- displacement, recursion etc.
- Neurophysiology- Broca and Wernicke
- Nature of NLP: NLP stack;  
NLP=linguistics+probability; 3 gens of NLP
- Main Challenge: Ambiguity
- ChatGPT's (an LLM) amazing capability-  
“Buffalo” sentence
- Course info- evaluation, references
- Heart of ML-NLP:  $\operatorname{argmax}(P(B|A))$

# 1-slide recap of week 2

- Perceptrons: only  $O(2^N)$  out of  $O(2^{2^N})$  Boolean Functions are threshold; XOR not computable
- PTA: find  $W$  s.t.  $W \cdot X_i > 0$ , for all  $i$ ;  
 $W_{next} = W + X_{fail}$
- PTA guaranteed to converge if vectors are from linearly separable function  
Proof by contradiction:  $G(w_n) = (W_n \cdot W^*) / |W_n|$
- Sigmoid and Softmax functions and their derivatives: (a) sigmoid  $O(1-O)$ ; (b) Softmax-  
 $O_k(1-O_k)$  and  $-O_k O_{k'}$

# 1-slide recap of week 3

- Recurrent Perceptron- concept of state, feedback
- Responsible AI- Toxicity, Hallucination, Deepfake
- Sigmoid and Softmax
- Derivative of sigmoid and softmax
- Weight change with gradient descent- sigmoid neuron and softmax neural net (without hidden layer)

# 1-slide recap week 4

- Small and Large LMs
- BP weight change rules
- Recurrent Perceptron
- Application of BP- skin disease prediction
- Vanishing gradient
- Derivation of word embeddings

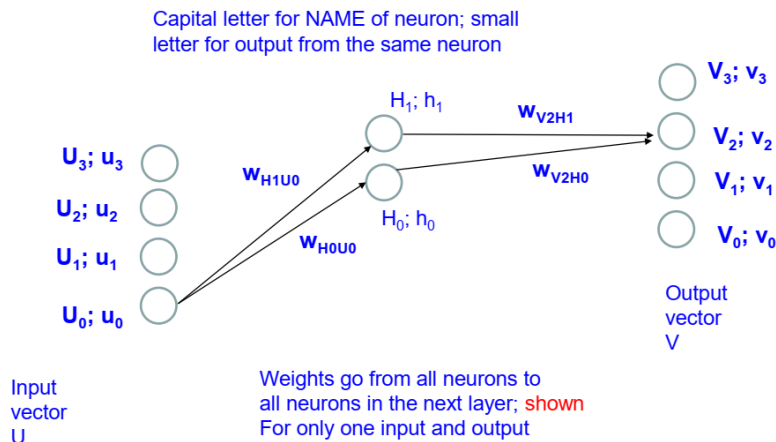
# 1-slide recap of week 5

- Linguistic foundation and derivation of WE
- Objective to be maximized- probability of context words (1-hot embeddings)
- Discriminative vs. Generative modeling- sentiment analysis example
  - Approach1:  $C^* = \operatorname{argmax}_C [P(C|S)]$ ; S is the sentence and C is the sentiment class
  - Approach2:  $C^* = \operatorname{argmax}_C [P(C)P(S|C)]$ ; apply Bayes theorem
  - Approach1 discriminative; features like adjectives give rise to the classification (e.g., 'wonderful' → positive sentiment)
  - Approach2 generative; class C 'generates' the features like adjectives (positive sentiment → 'wonderful')



# 1-slide recap of week 6

- Derivation of weight change rule for skip gram



$$net_{V_0} = W_{U_0} \cdot W_{V_0}^T$$

$$net_{V_1} = W_{U_0} \cdot W_{V_1}^T$$

$$net_{V_2} = W_{U_0} \cdot W_{V_2}^T$$

$$net_{V_3} = W_{U_0} \cdot W_{V_3}^T$$

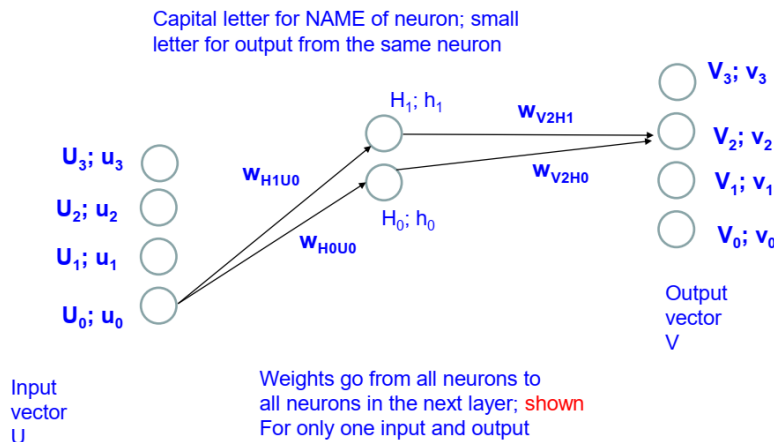
$$\Delta w_{V_2 H_0} = \eta(1 - v_2) \cdot w_{H_0 U_0} = \eta(1 - v_2) o_{H_0}$$

$$\Delta w_{H_0 U_0}$$

$$= \eta[(1 - v_2)w_{V_2 H_0} + (0 - v_0)w_{V_0 H_0} + (0 - v_1)w_{V_1 H_0} + (0 - v_3)w_{V_3 H_0}] \cdot u_0$$

# 1-slide recap of week 7

- Derivation of weight change rule for skip gram



$$net_{V_0} = W_{U_0} \cdot W_{V_0}^T$$

$$net_{V_1} = W_{U_0} \cdot W_{V_1}^T$$

$$net_{V_2} = W_{U_0} \cdot W_{V_2}^T$$

$$net_{V_3} = W_{U_0} \cdot W_{V_3}^T$$

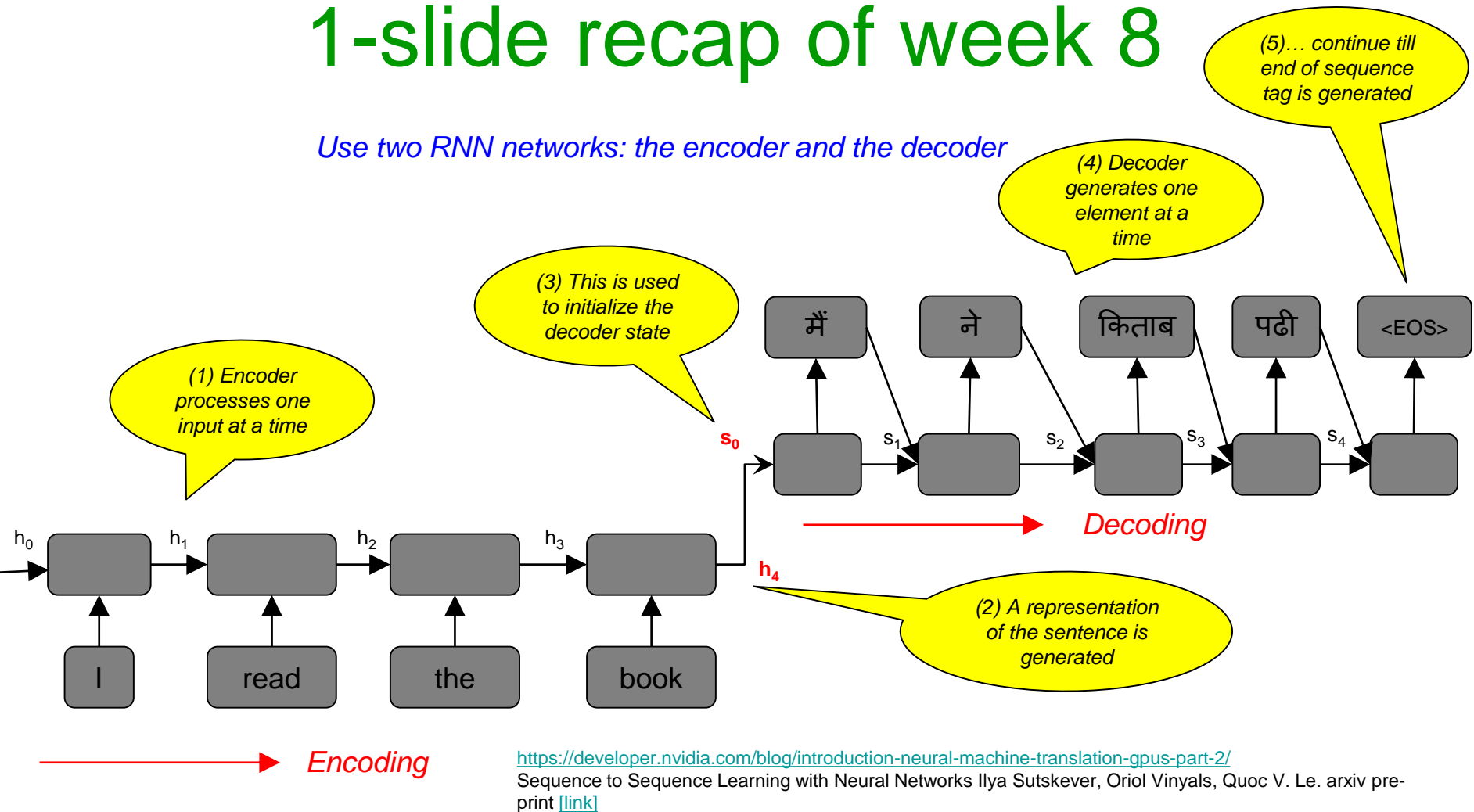
$$\Delta w_{V_2 H_0} = \eta(1 - v_2) \cdot w_{H_0 U_0} = \eta(1 - v_2) o_{H_0}$$

$$\Delta w_{H_0 U_0}$$

$$= \eta[(1 - v_2)w_{V_2 H_0} + (0 - v_0)w_{V_0 H_0} + (0 - v_1)w_{V_1 H_0} + (0 - v_3)w_{V_3 H_0}] \cdot u_0$$

# 1-slide recap of week 8

Use two RNN networks: the encoder and the decoder

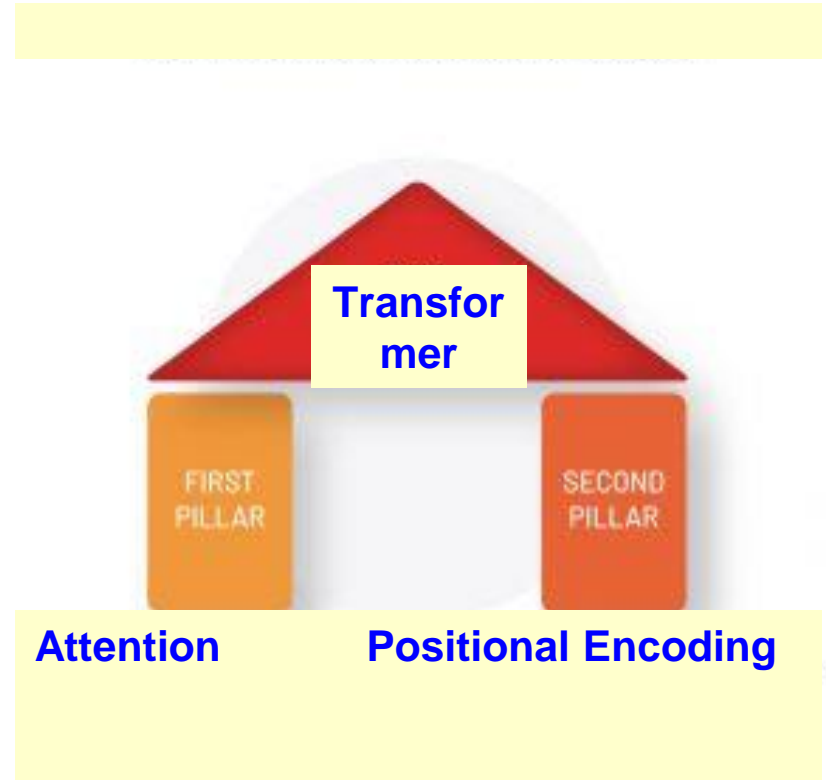


In the context of decoding, did A\*, Viterbi and Beam Search

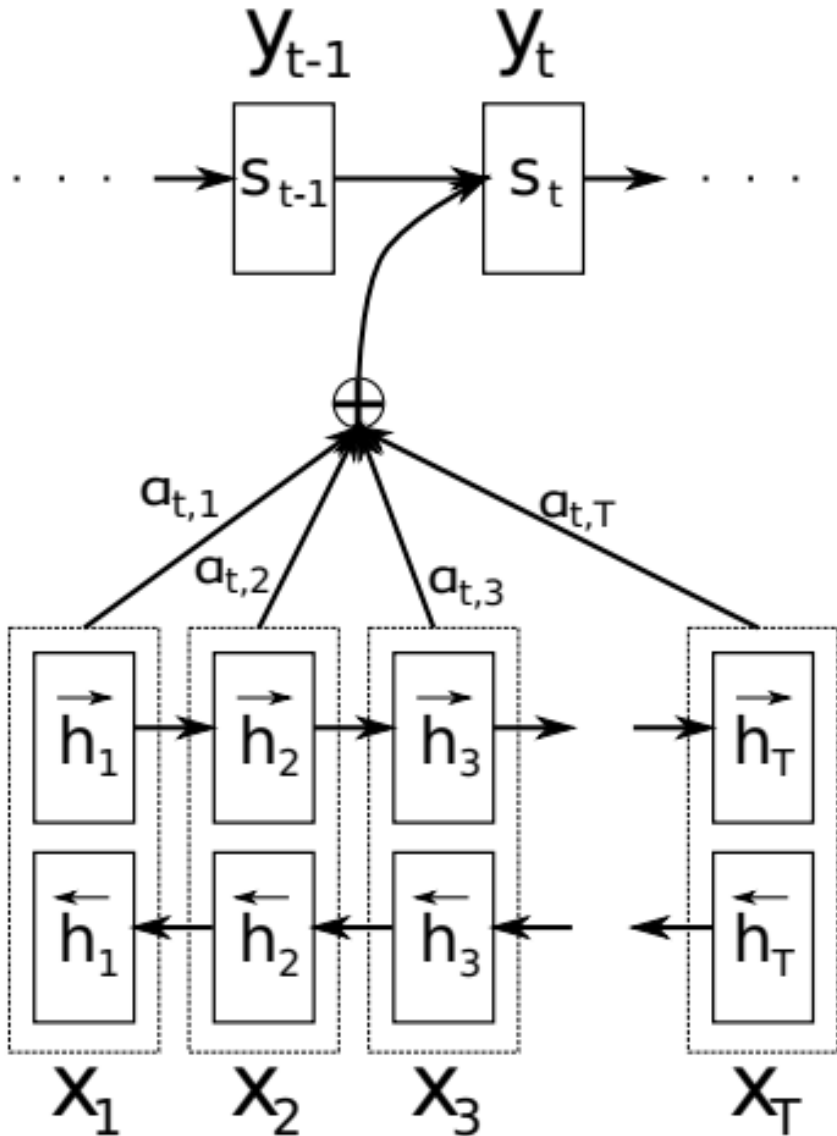
# 1-slide recap of week 9

[nesu]	“I carry”
[ponese]	“He will carry”
[nese]	“He carries”
[nesou]	“They carry”
[yedu]	“I drive”
[plavou]	“They swim”

$$pos(t, 2i) = \sin\left(\frac{1}{10000^{\frac{2i}{d}}} t\right)$$
$$pos(t, 2i + 1) = \cos\left(\frac{1}{10000^{\frac{2i}{d}}} t\right)$$



# 1-slide recap of week 10



$$h_t = f(x_t, h_{t-1})$$

$$c = g(\{h_1, h_2, h_3, \dots, h_{T_x}\})$$

$$P(\bar{y}) = \prod_{t=1}^T P(y_t | \{y_1, y_2, y_3, \dots, y_{t-1}\}, c)$$

$$P(y_t | \{y_1, y_2, y_3, \dots, y_{t-1}\}, c) = g(y_{t-1}, s_t, c)$$

$$p(y_i | y_1, y_2, \dots, y_{i-1}) = g(y_{i-1}, s_i, c_i)$$

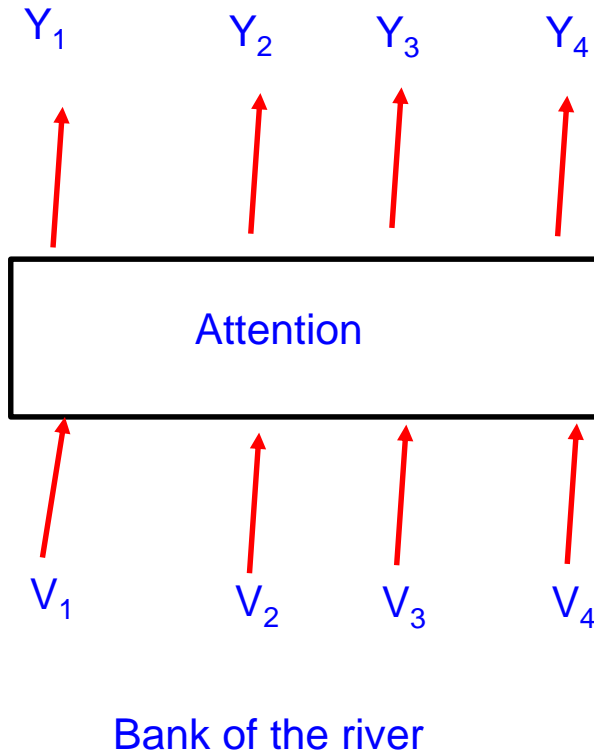
$$s_i = f(s_{i-1}, y_{i-1}, c_i)$$

$$c_i = \sum_{j=1}^{T_x} \alpha_{ij} h_j$$

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{ik})}$$

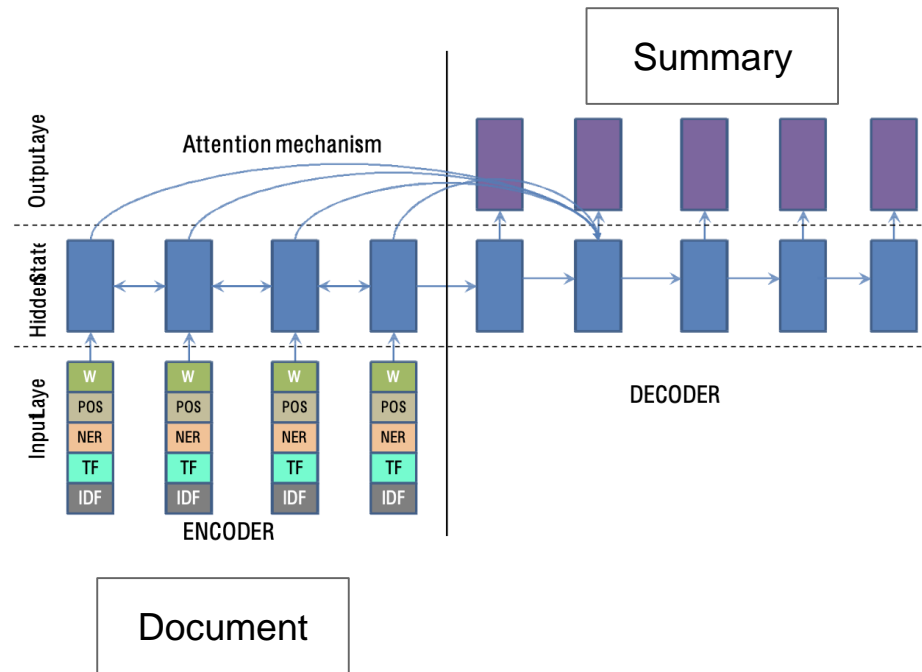
$$e_{ij} = a(s_{i-1}, h_j)$$

# 1-slide recap of week 11



$Y_i$  – vector

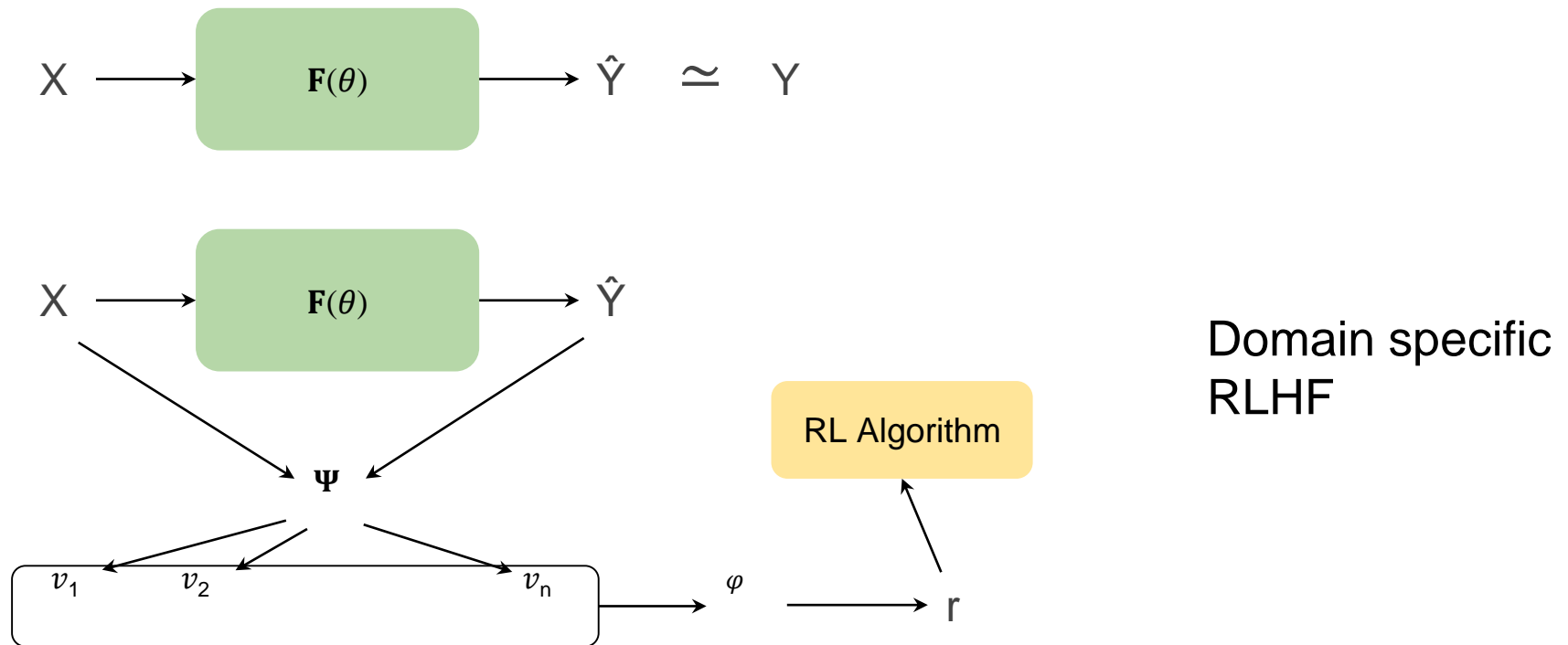
$$= w_{11} \cdot V_1 + w_{12} \cdot V_2 + w_{13} \cdot V_3 + w_{14} \cdot V_4$$



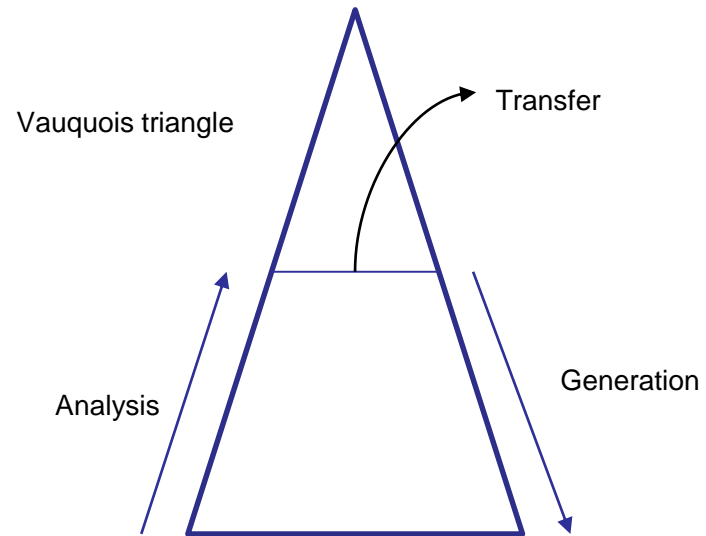
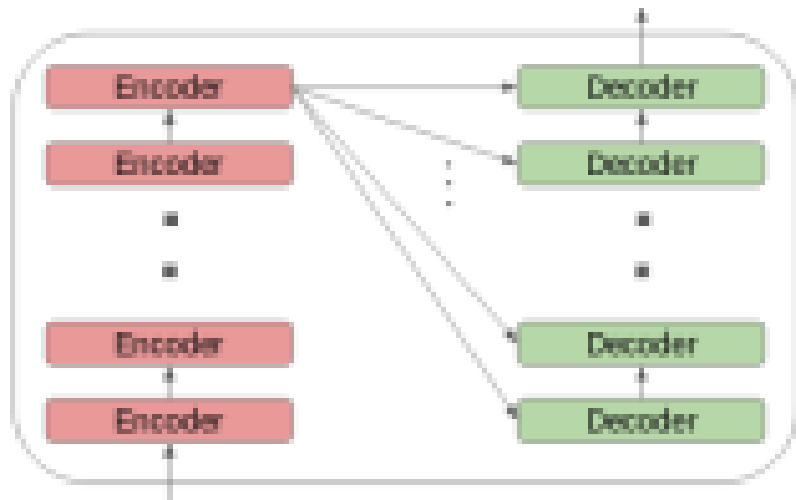
# 1-slide recap of week 12

$$\tilde{e} = \operatorname{argmax}_{e \in e^*} p(e|f) = \operatorname{argmax}_{e \in e^*} p(f|e)p(e)$$

SMT



# 1-slide recap of week 13



## Efficient and elastic Large Models (Dr. Prateek Jain, Google Research)

Key challenges in improving efficiency of LLM serving.

Matformers: train one model but read-off 100s of smaller models + speed up decoding in LLMs

Depends on Matryoshka Representation Learning, Neurips 24

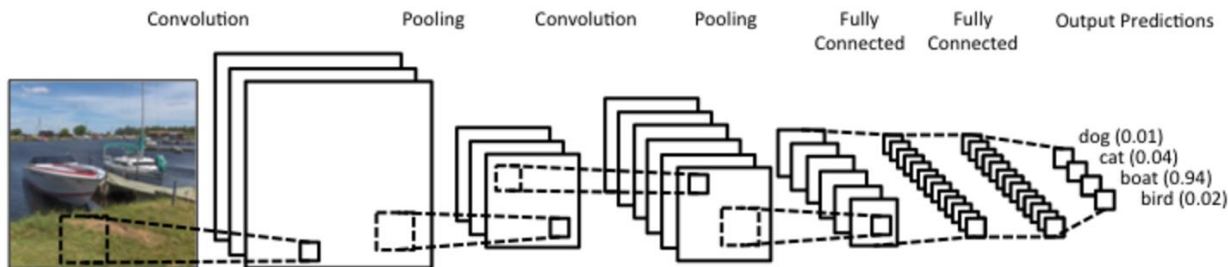


# 1-slide recap of week 14

## Three generations on LM:

1. Grammar Rules: Context Free Grammars, 2a. N-grams, 2b. Probabilistic Context Free Grammars, 3. Neural (solves the CURSE OF DIMENSIONALITY problem)

**CNNs:** stages of conv+RELU+Pooling layers terminating in FFNN and Softmax; very successful in Image and Vision; Partially in Language



**Stable Diffusion:** competitor to LLM. **LLMs:** high quality generation of text;

**Diffusion:** high-resolution image synthesis



# 1-slide recap of week 15

**Reasoning in LLMs (Sravanthi et al)**

**Conversational AI: Impact, Linguistics,  
Computation**

**Closure**

Thank you

All the best!!!