NLP, LLM and Applications

Pushpak Bhattacharyya
Computer Science and Engineering Department
IIT Bombay
AI Workshop, INS Balsura
11th August, 2023

Ode to *Scientists and Engineers*

Scientists ask WHY
Engineers ask WHY NOT
Scientists wonder at WHAT-IS
Engineers wonder WHAT-COULD-BE
World couldn’t do without either.

Scientists STUDY
Engineers MAKE
And ever the twain shall meet.

What is NLP?

Ans: Language+Computation

NLP

The branch of AI dealing with the study and implementation of endowing natural language **understanding and generation** ability onto computers.

A bit of history

- Born in 1940s: need for translating between English and Russian- cold war days
- 1957-70: Symbolic vs. Stochastic- rule and knowledge based vs. data and probability
- 1970: Logic based NLP- Prolog
- 1983-93: more of empiricism and reign of data, ML, probability- noisy channel, graphical model
- 2010 onwards: Neural models, Gradient Descent, Backpropagation

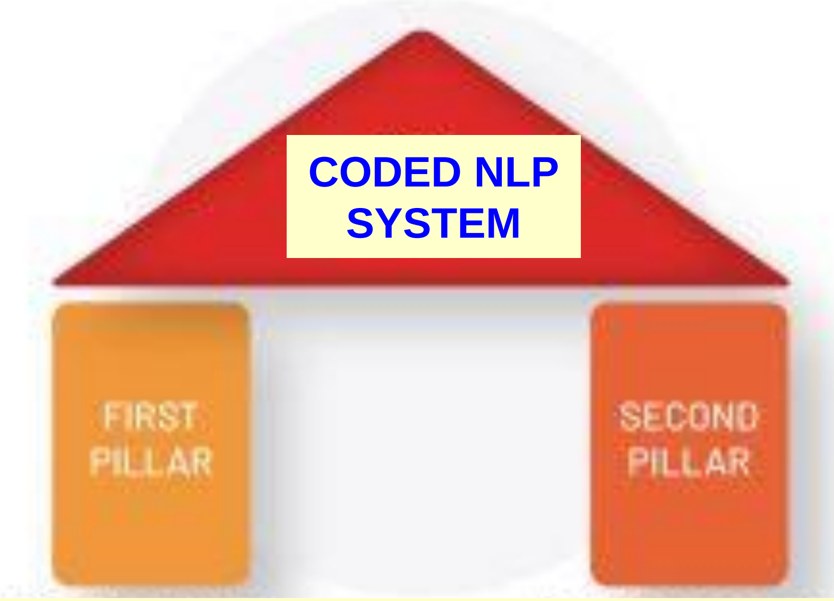
Volume of NLP Business

- “The global natural language processing (NLP) market was worth $15.7 billion in 2022. It is expected to grow to $49.4 billion by 2027, at a compound annual growth rate (CAGR) of 25.7%.” (https://www.marketsandmarkets.com/Market-Reports/natural-language-processing-nlp-825.html)
- Caused by data, ML, systems and architecture
- Lets look at two key developments
- WORD VECTORS and TRANSFORMERS

NLP: Task *vs.* Technique Matrix

Task (row) vs. Technique (col) Matrix	**Rules Based /Knowledge-Based**	**Classical ML**				**Deep Learning, LLM, Stable diffusion**		
		Perceptron	Logistic Regression	SVM	Graphical Models (HMM, MEMM, CRF)	Dense FF with BP and softmax	RNN-LSTM	CNN
Morphology								
POS								
Chunking								
Parsing								
NER, MWE								
Coref								
WSD								
Machine Translation								
Semantic Role Labeling								
Sentiment								
Question Answering								

Two Pillars of NLP



CODED NLP SYSTEM

LINGUISTICS + PROBABILITY
= NLP

Linguistics → Language Insight: 1st pillar of NLP

- **BEFORE** the model operates:
  - *Preparing the training data; Deciding on the annotation scheme (e.g., POS tag labels)*
  - *Guidelines for annotation (e.g., in ‘golf club’, ‘golf’ is an adjectival, i.e., noun functioning as adjective)*
- **AFTER** the model has operated
  - *Interpretation of model output (e.g., “I love being ignored” labeled as ‘sarcastic’, incongruity of ‘love’ and ‘ignore’)*
  - *Error Analysis (MT of Dravidian languages has low BLEU score due to morph complexity)*

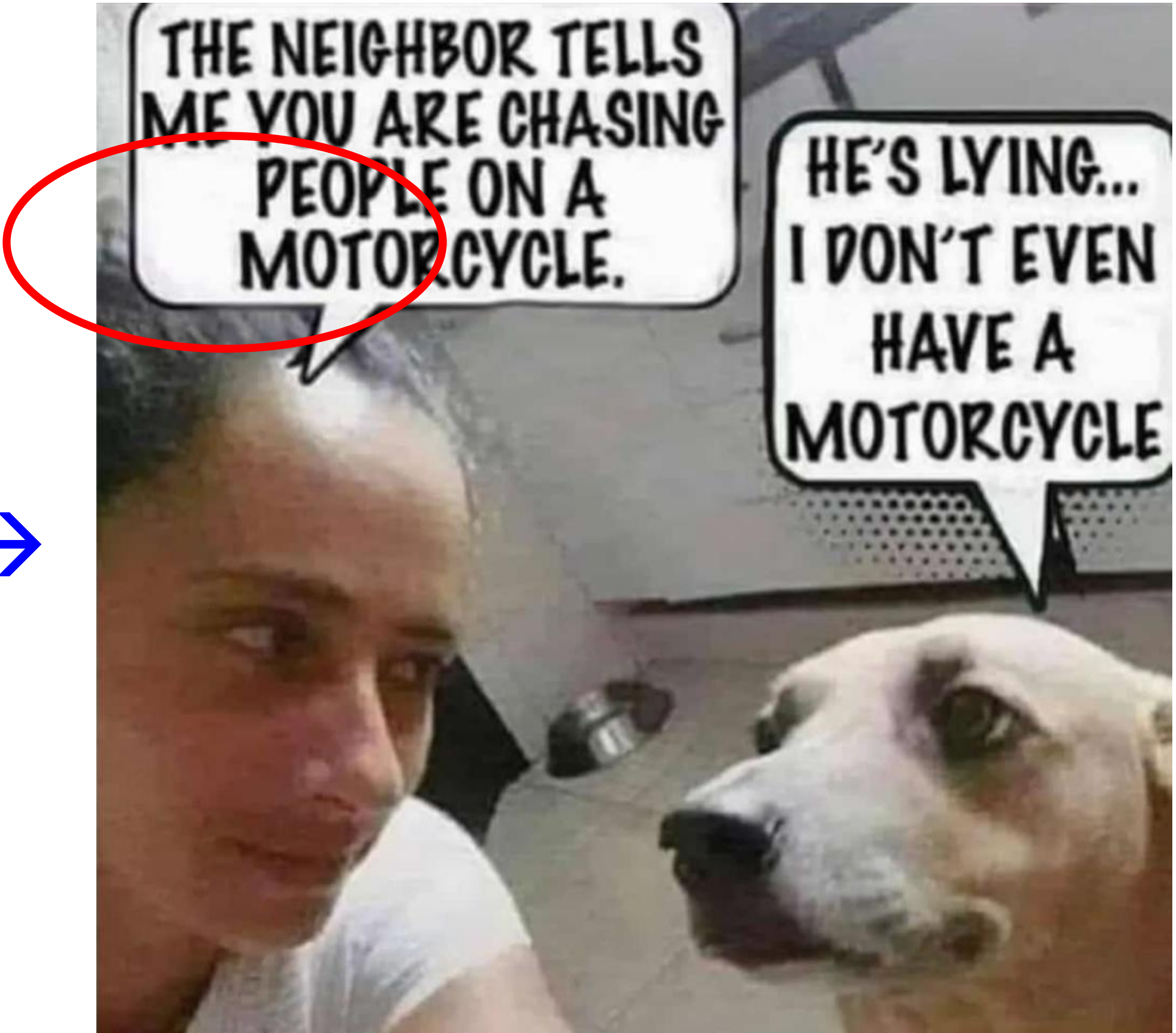
Probability → Data Insight: 2nd pillar of NLP

- **LAW OF LARGE NUMBERS** operates
  - *NN follows JJ 80% of the time: P(NN|JJ)=0.8*
  - *Credibility increases as data increases*
- **CENTRAL LIMIT THEOREM** operates
  - *Beyond a threshold (~30) on the amount of data, we can use normal distribution approximation, whatever the underlying distribution to estimate the parameters*
  - *Can use hypothesis testing to ascertain the robustness of generalization and also measure statistically the PROGRESS (aka, improvement over baseline)*

Main Challenge: **AMBIGUITY (1/3)**

- Elderly with young **face** increased covid 19 risk (ToI Oct 20)
  - *Old people with young children…*
  - *Old people whose faces are young…!!!*
- Maharashtra reports increased covid-19 cases
  - *it is reported by Maharashtra Govt. that covid-19 cases have increased*
  - *it is the Maharashtra reports that have increased covid-19 cases!!!*

Ambiguity (2/3)

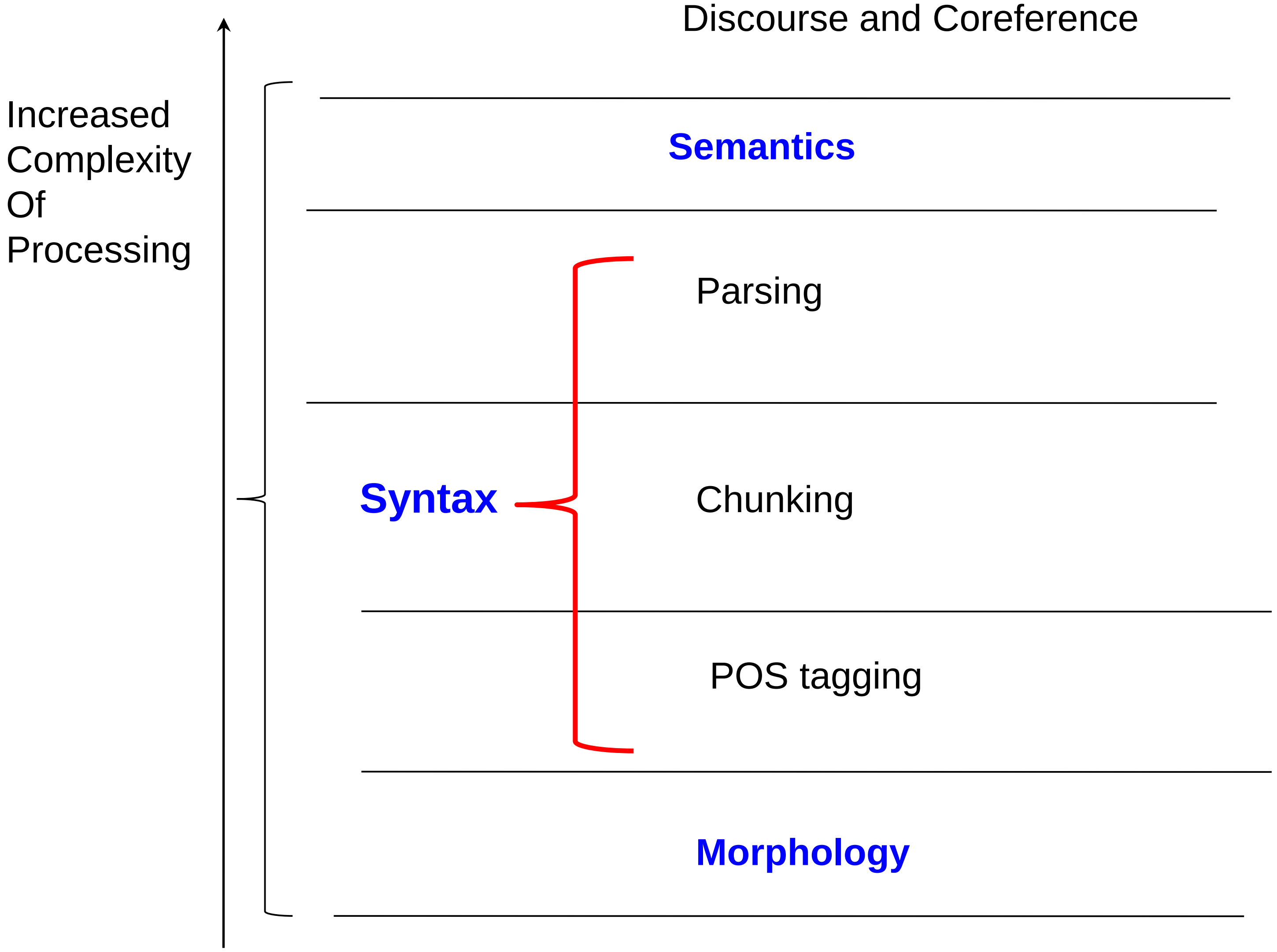

THE NEIGHBOR TELLS ME YOU ARE CHASING PEOPLE ON A MOTORCYCLE.
HE'S LYING... I DON'T EVEN HAVE A MOTORCYCLE

Incongruity→
Ambiguity→
Humour

Ambiguity (3/3)

- Vividh Bhartai sakhi saheli 29mar23:
  - "aaj kaaryaalay band rakhaa jaayegaa“ (today the office will remain closed)
  - "aaj kaaryaalay bandar khaa jaayegaa“ (today the office will be eaten by a monkey)
- ToI, 3dec22
  - “Korea leave Uruguay in Tears” (beat Uruguay in FIFA): who was in tears? Uruguay place or team? leave--> go or cause?

NLP Layers



Increased Complexity Of Processing

Discourse and Coreference

Semantics

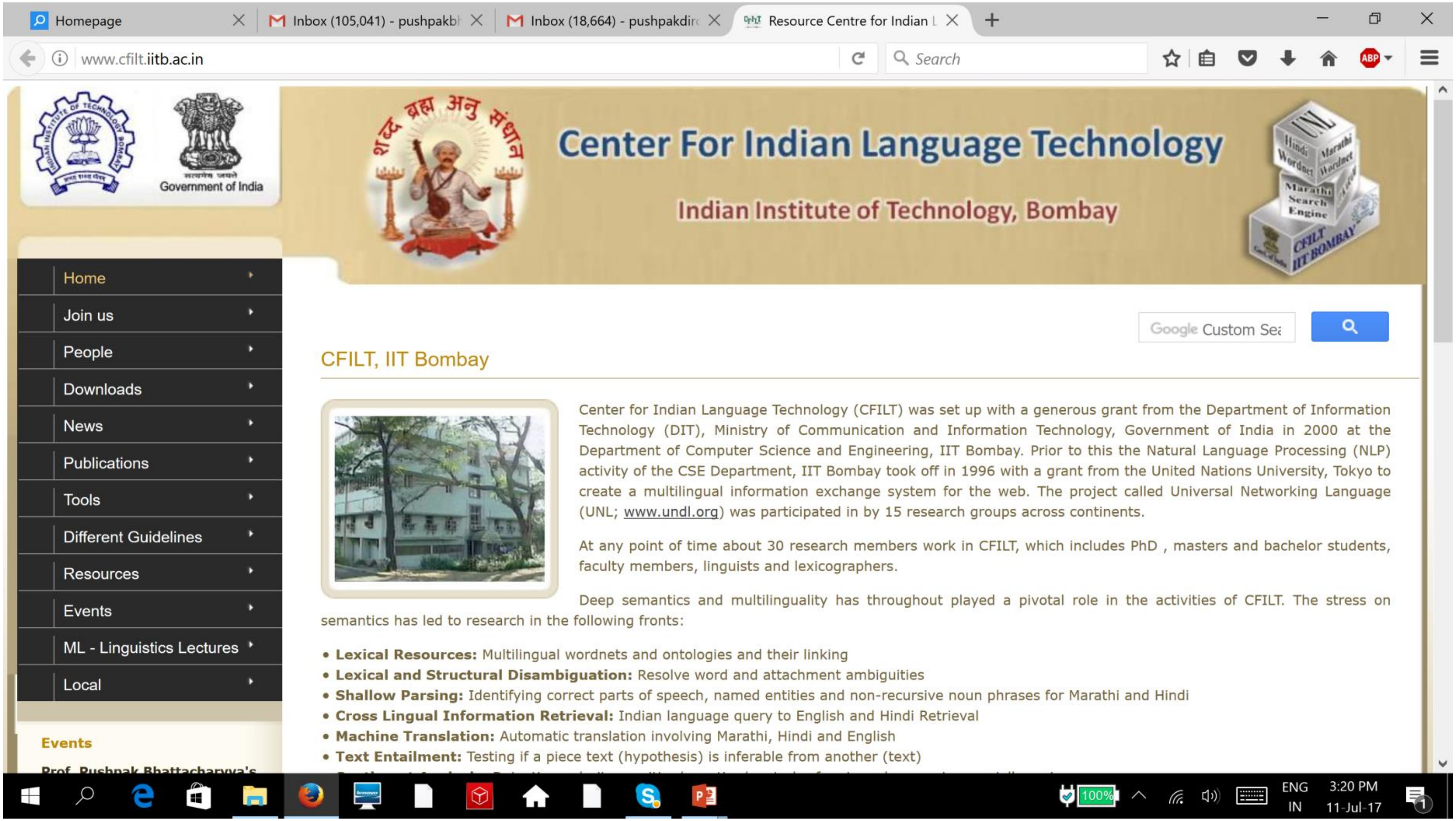
Syntax

Parsing

Chunking

POS tagging

Morphology



CFILT, IIT Bombay

Center for Indian Language Technology (CFILT) was set up with a generous grant from the Department of Information Technology (DIT), Ministry of Communication and Information Technology, Government of India in 2000 at the Department of Computer Science and Engineering, IIT Bombay. Prior to this the Natural Language Processing (NLP) activity of the CSE Department, IIT Bombay took off in 1996 with a grant from the United Nations University, Tokyo to create a multilingual information exchange system for the web. The project called Universal Networking Language (UNL; www.undl.org) was participated in by 15 research groups across continents.

At any point of time about 30 research members work in CFILT, which includes PhD , masters and bachelor students, faculty members, linguists and lexicographers.

Deep semantics and multilinguality has throughout played a pivotal role in the activities of CFILT. The stress on semantics has led to research in the following fronts:

- **Lexical Resources:** Multilingual wordnets and ontologies and their linking
- **Lexical and Structural Disambiguation:** Resolve word and attachment ambiguities
- **Shallow Parsing:** Identifying correct parts of speech, named entities and non-recursive noun phrases for Marathi and Hindi
- **Cross Lingual Information Retrieval:** Indian language query to English and Hindi Retrieval
- **Machine Translation:** Automatic translation involving Marathi, Hindi and English
- **Text Entailment:** Testing if a piece text (hypothesis) is inferable from another (text)

Our NLP Lab at IIT Bombay*: Since 2000;* Works in all areas of NLP- **translation, QA, Sentiment and Emotion, Natural Language Generation** and so on.

Key Research Areas

Machine Translation

Sentiment Analysis

Information Retrieval

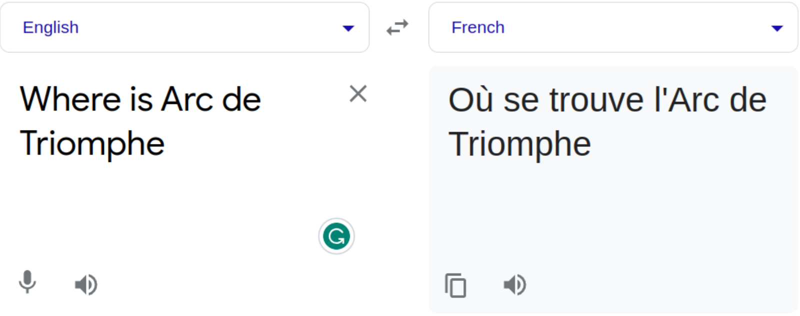
Lexical Semantics

Information Extraction

Cognitive NLP

LMs, LLMs, Comparisons

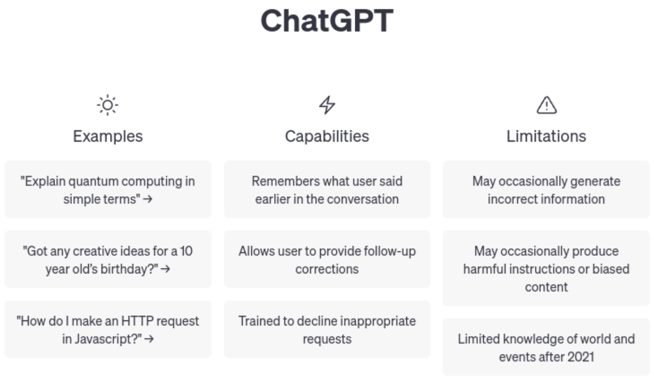
LMs everywhere



Where is Arc de Triomphe

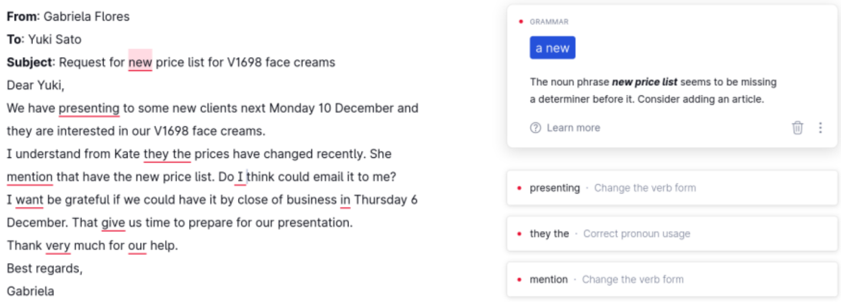
Où se trouve l'Arc de Triomphe

Machine Translation



Examples	Capabilities	Limitations
"Explain quantum computing in simple terms" →	Remembers what user said earlier in the conversation	May occasionally generate incorrect information
"Got any creative ideas for a 10 year old's birthday?" →	Allows user to provide follow-up corrections	May occasionally produce harmful instructions or biased content
"How do I make an HTTP request in Javascript?" →	Trained to decline inappropriate requests	Limited knowledge of world and events after 2021

QA

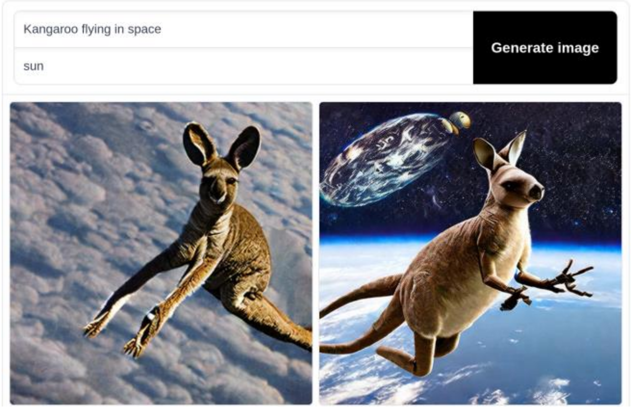


From: Gabriela Flores
To: Yuki Sato
Subject: Request for new price list for V1698 face creams
Dear Yuki,
We have presenting to some new clients next Monday 10 December and they are interested in our V1698 face creams.
I understand from Kate they the prices have changed recently. She mention that have the new price list. Do I think could email it to me?
I want be grateful if we could have it by close of business in Thursday 6 December. That give us time to prepare for our presentation.
Thank very much for our help.
Best regards,
Gabriela

The noun phrase ***new price list*** seems to be missing a determiner before it. Consider adding an article.

- presenting · Change the verb form
- they the · Correct pronoun usage
- mention · Change the verb form

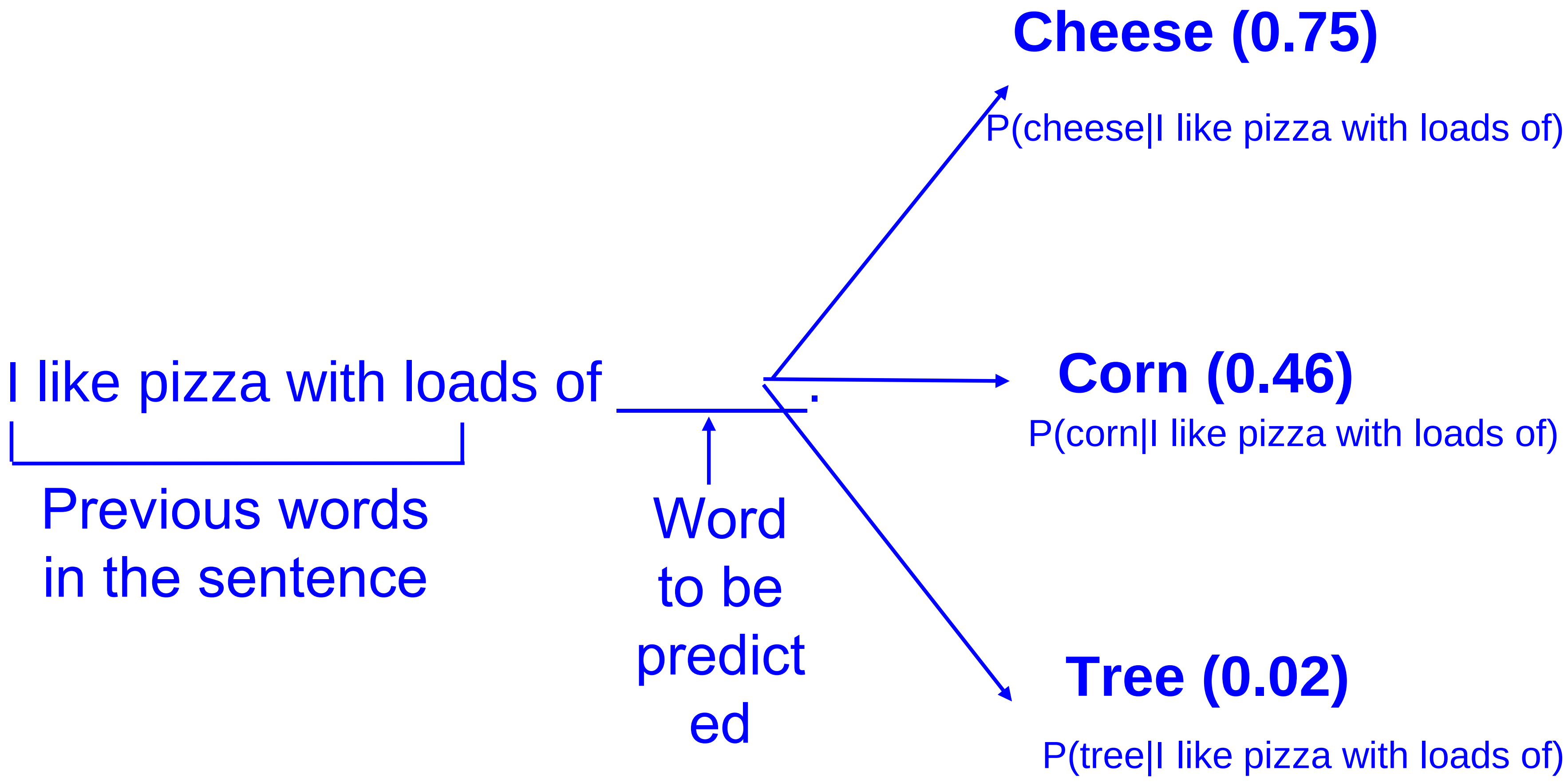
Writing aid (grammarly uses LLM)



Text Encoding
(Stable Diffusion)

Why Language Models?

Language Modelling



I like pizza with loads of ________.

Previous words in the sentence

Word to be predicted

Cheese (0.75)

P(cheese|I like pizza with loads of)

Corn (0.46)

P(corn|I like pizza with loads of)

Tree (0.02)

P(tree|I like pizza with loads of)

P(cheese| I like pizza with loads of) > P(corn| I like pizza with loads of) >> P(tree| I like pizza with loads of)

Two most famous LMs

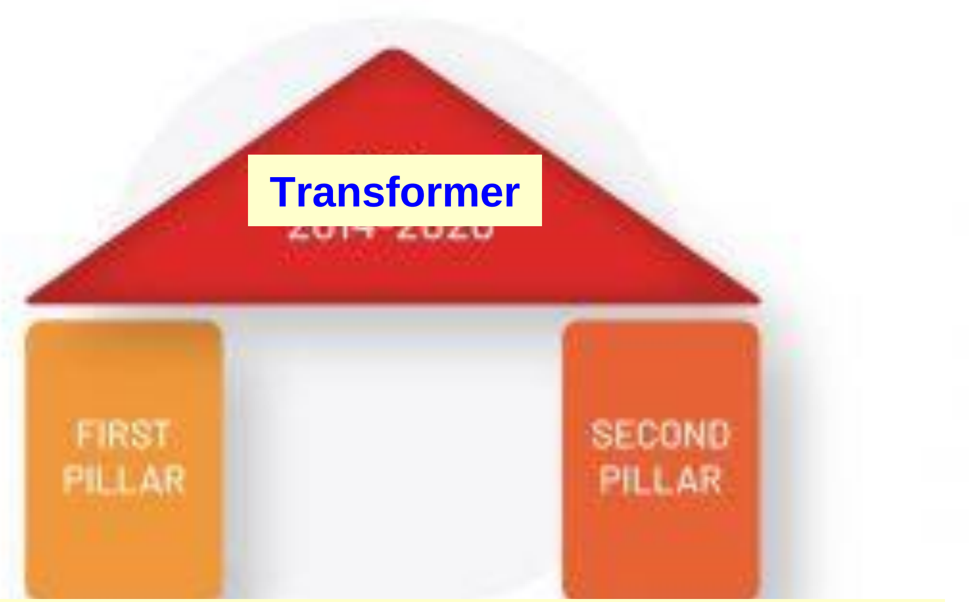
- InstructGPT:
  - ***Command/Request/Order → Response***
- ChatGPT:
  - Carry out a ***conversation***
  - Respect context (state), personalization, quality and quantity and respond
    - Input: *I have been promoted*
    - Appropriate response: *I am delighted/congratulations/great* ..
    - Inappropriate: *why did they promote you?*

Transformer based famous models

- BERT
- BART
- XLM
- MURIL
- INDICBERT
- BIOBERT

and so on

Two Pillars of Transformer

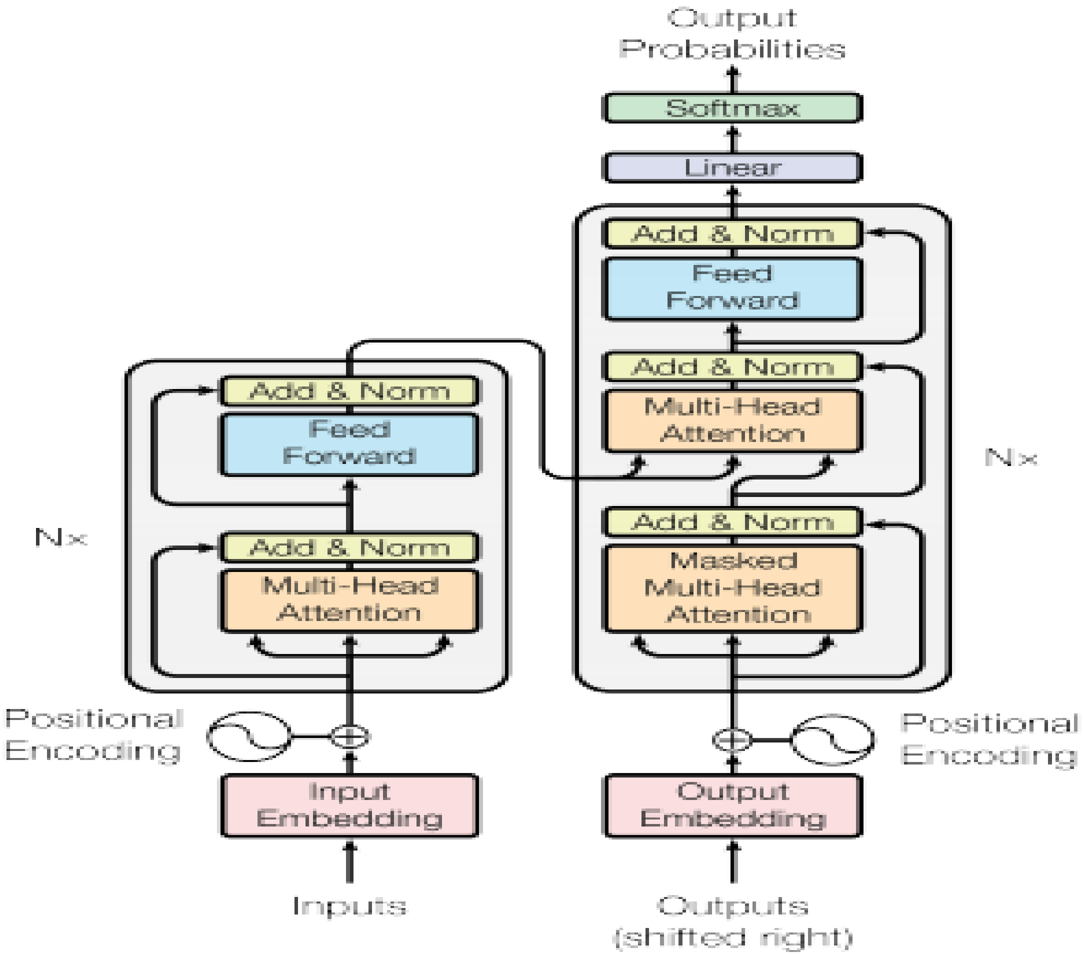


Transformer

Attention + Positional Encoding

= Transformer

A classic diagram and a classic paper



Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. "Attention is all you need." NeurIPS (2017).

http://nlp.seas.harvard.edu/2018/04/03/attention.html
http://jalammar.github.io/illustrated-transformer/

Chronology

- IBM Models of Alignment- Brown *et al.* 1990, 1993
- Phrase Based MT- Koehn 2003
- Encoder Decoder- Sutskever *et al.* 2014, Cho *et al.* 2014
- Attention- Bahadanu *et al.* 2015
- Transformer- Vaswani *et al.* 2017

3 stages of LLM based CAI

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

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#)

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!”

“How do I install RAM into my PC?”

“The instructions should definitely guide people to their motherboard manual to ensure RAM is being installed optimally. ChatGPT does pick up on a key part of the RAM install process — checking your system BIOS afterward — but it doesn’t go through another all-important BIOS step. If you’ve picked up some Intel XMP-compatible RAM, you’ll typically need to enable this in the BIOS settings afterward, and likewise for AMD’s equivalent. Otherwise, you’re not running your RAM at the most optimized timings to get the best performance. Overall, the advice is solid but still very basic. It’s better than some PC building guides, but I’d like to have seen the BIOS changes or dual-channel

“Write me a poem about a worm” (1/2)

- Anapestic tetrameter: an arcane meter
  - 1.“Twas the night before Christmas, when all through the house/Not a creature was stirring, not even a mouse;
    2. The stockings were hung by the chimney with care,/In hopes that St. Nicholas soon would be there;
    3. The children were nestled all snug in their beds;/While visions of sugar-plums danced in their heads;”

“What’s the average salary for a plumber in NYC? (And cite your sources)” (1/2)

- Bing’s cited sources include [Zippia](), [CareerExplorer](), and [Glassdoor]()

- ChatGPT: gave a ballpark figure, explained that there were caveats, and told about what sources one could check for more detailed numbers

“Design a training plan to run a marathon” (1/2)

- “In the race to make a marathon training plan, ChatGPT is the winner by many miles.
- Bing linked out to a *Runner’s World* article. This isn’t necessarily an irresponsible decision — I suspect that *Runner’s World* is an expert on marathon training plans! — but if I had just wanted a chatbot to tell me what to do, I would have been disappointed.

I gave this test (1/2)

Read this story and examine everything is right about this story: "once a ship was going from England to France. After a few days of journey, it was caught in a violent storm. The captain tried a lot to steer the ship clear from the eye of the storm, but the ship sped to a cliff and dashed against it. All those on board died and only the captain survived. With a binocular, he could spot a spot of jungle very far. With great difficulty, he took his articles near the jungle and after going there fell unconscious. He regained consciousness after 3 days. After becoming conscious, the captain felt very hungry. He fortunately found some potatoes amongst his articles. (cntd.)

I gave this test (2/2)

He managed to get a fire ignited by striking two pieces of stone. There were some dry pieces of wood which he lighted and boiled the potatoes on the fire and appeased his hunger. When he had slept for some time, he woke up and found some shining objects on the sand where the fire was burnt. He kept these objects with him. After several days, he spied a ship far away with the binocular. He managed to attract attention of the ship with his torn shirt waving it vigorously. The ship rescued him from the island. The captain took those shining objects with him to England. Later they found interesting properties of these shining objects. These objects were transparent, but could be used to make containers. The material was given the name "glass". This is how glass

ChatGPT response

There are several issues with this story, indicating that it is not entirely accurate.

Firstly, glass was not invented by a shipwrecked captain as described in the story. The history of glassmaking goes back thousands of years and involves the contributions of many different cultures.

Secondly, the story implies that glass is made from potatoes, which is completely false. Glass is made from a mixture of silica, soda ash, and limestone, among other materials.

Applications

Automatic Radiology/Sonology Reports Generation

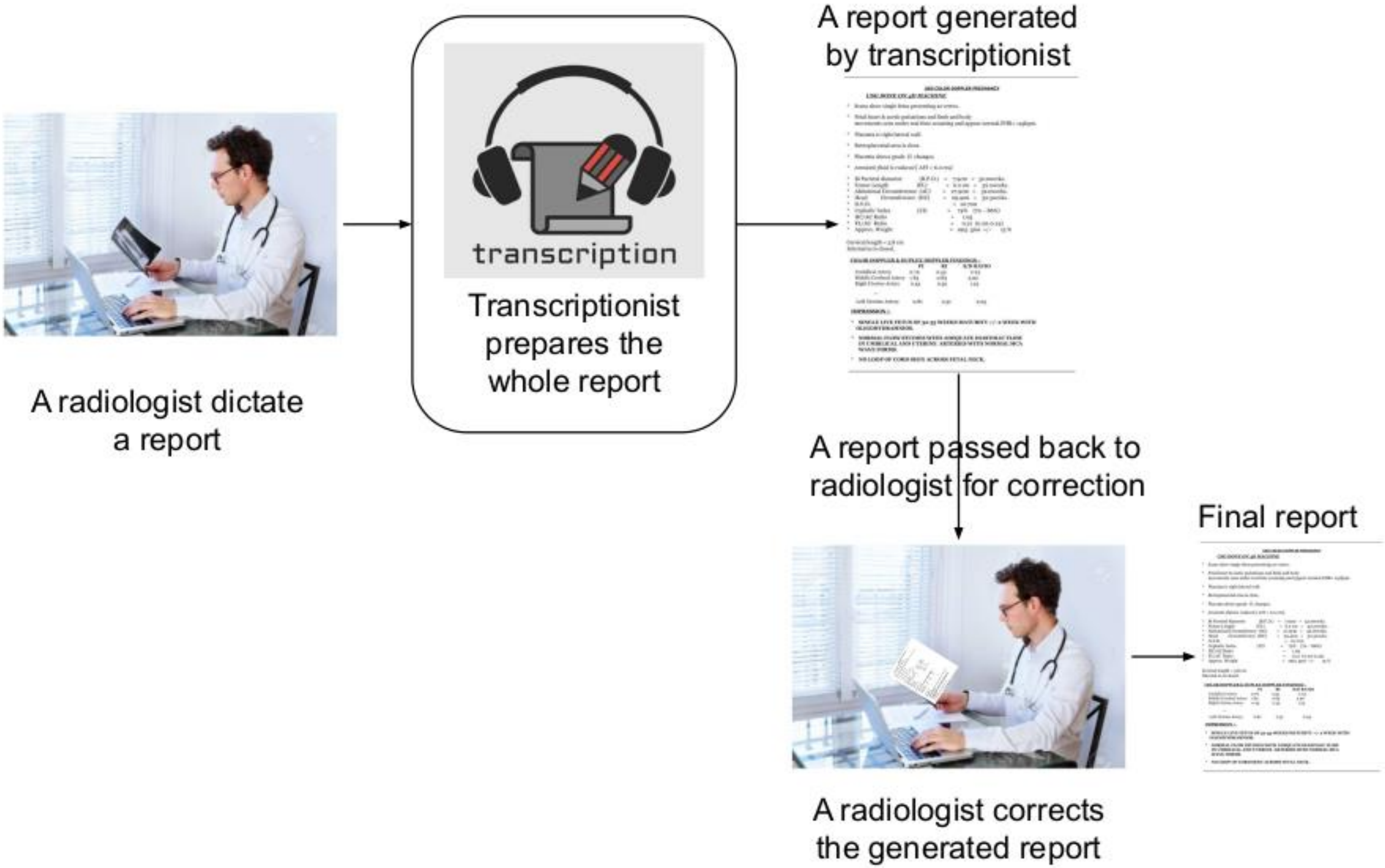
1: From Dictation

2. From Image

(EACL21, ACL23)

Scarcity of Radiologists

- Radiologist to Patient ratio in India is, 1:100,000, the corresponding ratio in the US is 1:10,000, and for China, it is 1:14,772.
- It results in very high patient inflows, making radiologists incredibly busy and stressed out.
- Current situation of radiology report generation causes
  - significant delays in report turnaround time,
  - errors in the reports, and
  - burnout.



A radiologist dictate a report

Transcriptionist prepares the whole report

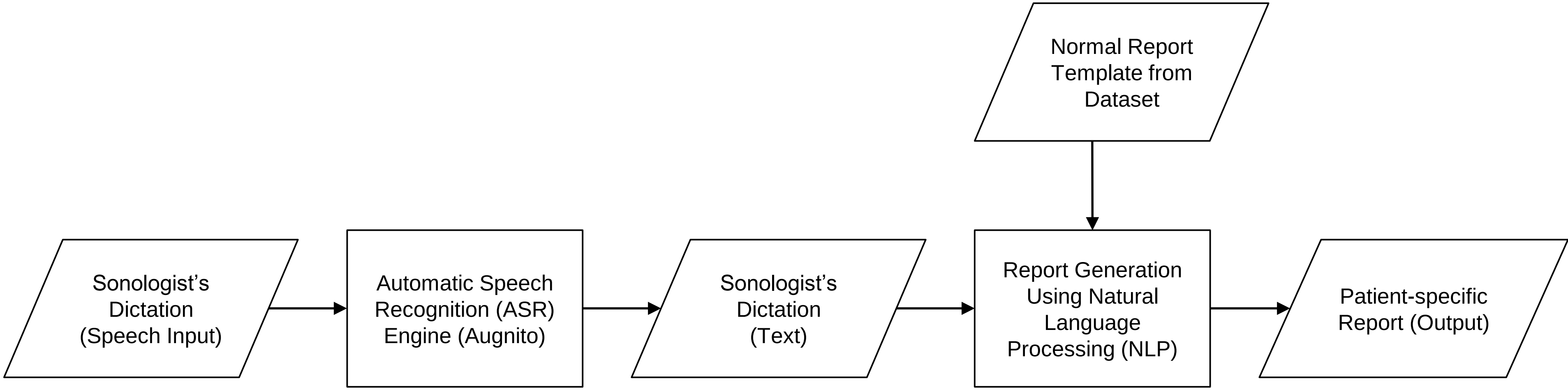
A report generated by transcriptionist

A report passed back to radiologist for correction

A radiologist corrects the generated report

Final report

Workflow



Normal Report Template from Dataset

Sonologist's Dictation (Speech Input)

Automatic Speech Recognition (ASR) Engine (Augnito)

Sonologist's Dictation (Text)

Report Generation Using Natural Language Processing (NLP)

Patient-specific Report (Output)

Problem Statement 1 (pertains to Ultrasound) (1/2)

Design a system that generates a structured patient-specific report from radiologist’s dictation and domain knowledge.

- Input
  1. Input text (radiologist’s dictation)
  2. Generic radiology report with all normal findings (normal report template)
- Output
  3. Radiology report with patient-specific findings

Domain knowledge comes from the Knowledge Graph (KG).

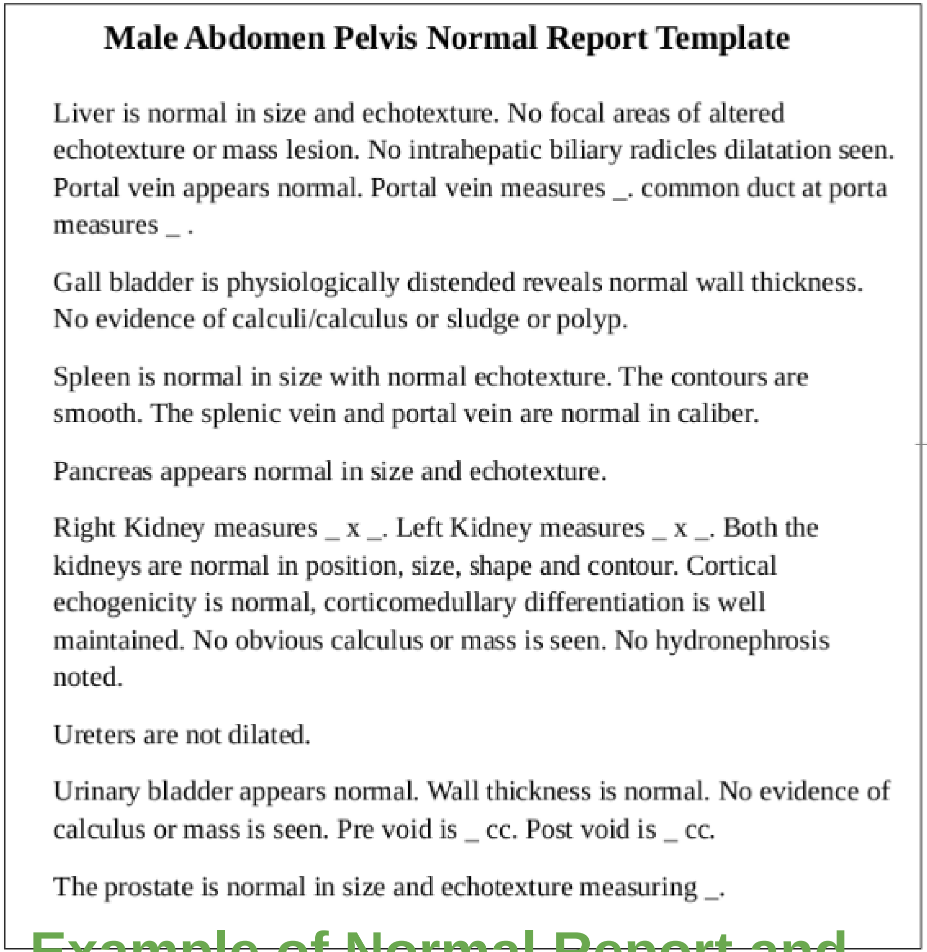
Problem Statement 1 (pertains to Ultrasound) (2/2)

Sub-Problem:

Develop a system that automatically constructs a KG of essential medical information from radiology free-text reports.

- Input: Radiology free-text report corpus
- Output: A formalized representation of the essential medical information contained within the free-text reports in the form of a hierarchical KG

Example of Normal Report and Patient Specific Report



Male Abdomen Pelvis Normal Report Template

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.

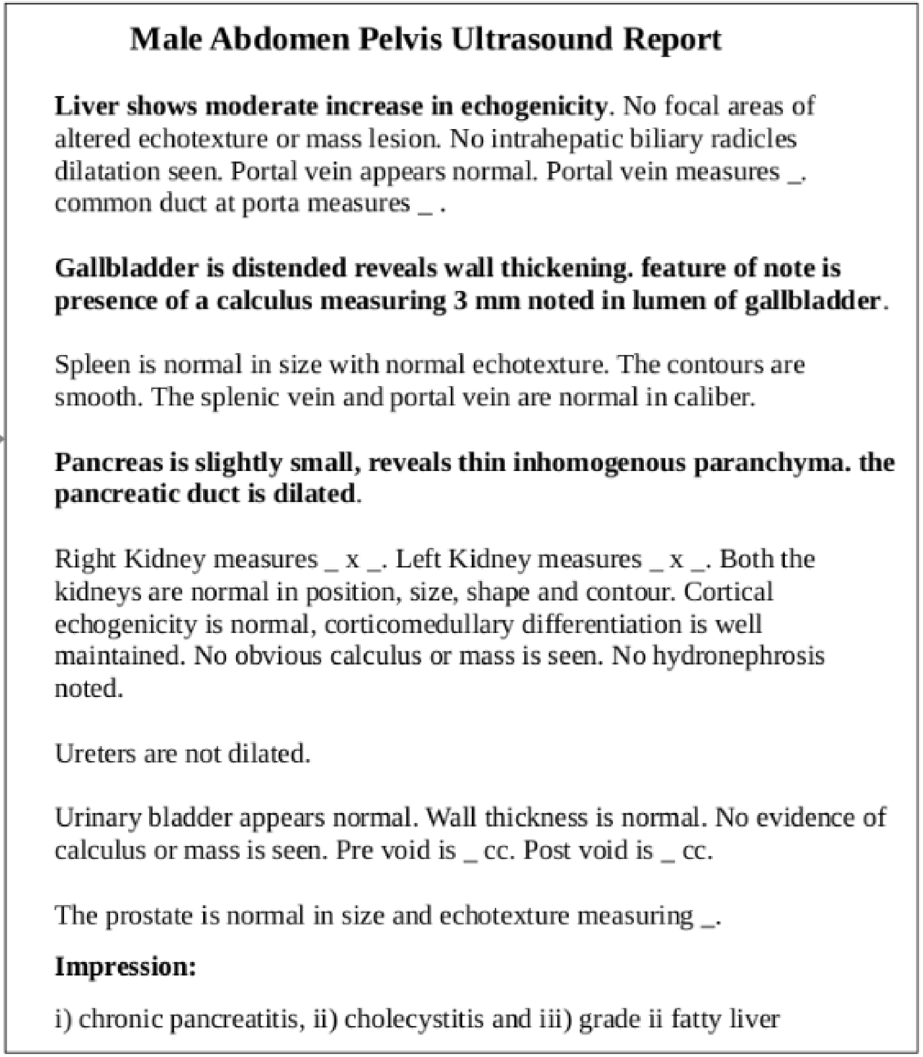
Pancreas appears normal in size and echotexture.

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 _.



Male Abdomen Pelvis Ultrasound 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

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.*

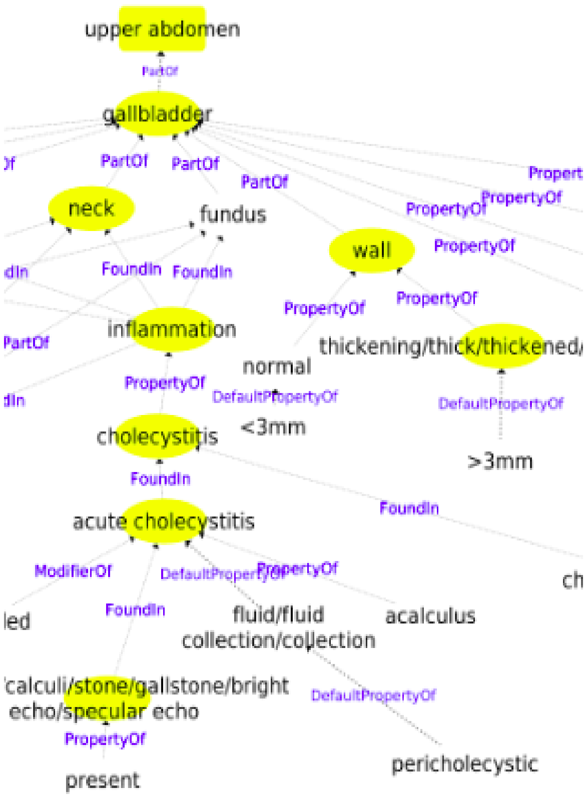
Pathological description: *Liver shows moderate increase in echogenicity.*

Example of Concept-expansion Graph

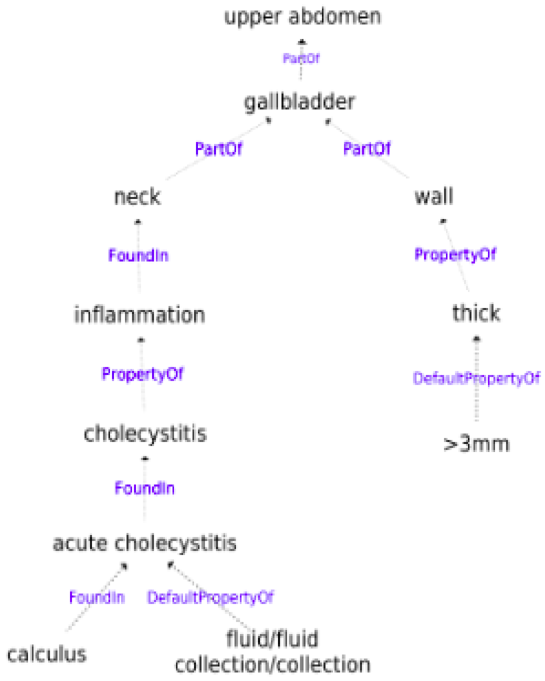
Radiologist's Notes: *Acute cholecystitis with few small calculi, largest measuring 2.4 mm with thickened wall impacted at neck of gallbladder*

Concept Set: *gallbladder,wall,thick,acute cholecystitis,calculus,neck,impacted,measuring,small,few,largest,2.4-mm*

Most Appropriate Path Found in KG



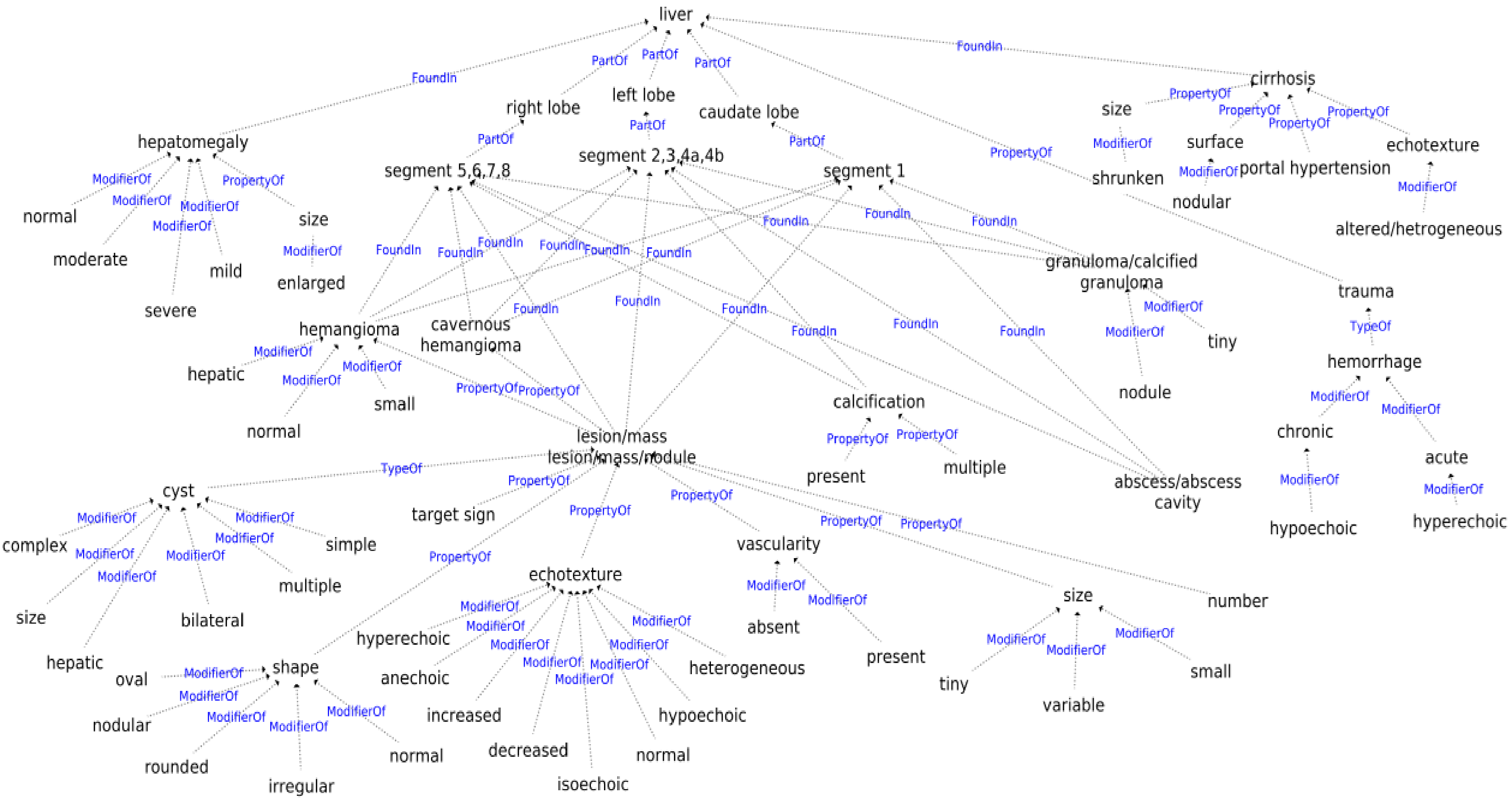
Concept-expansion Graph



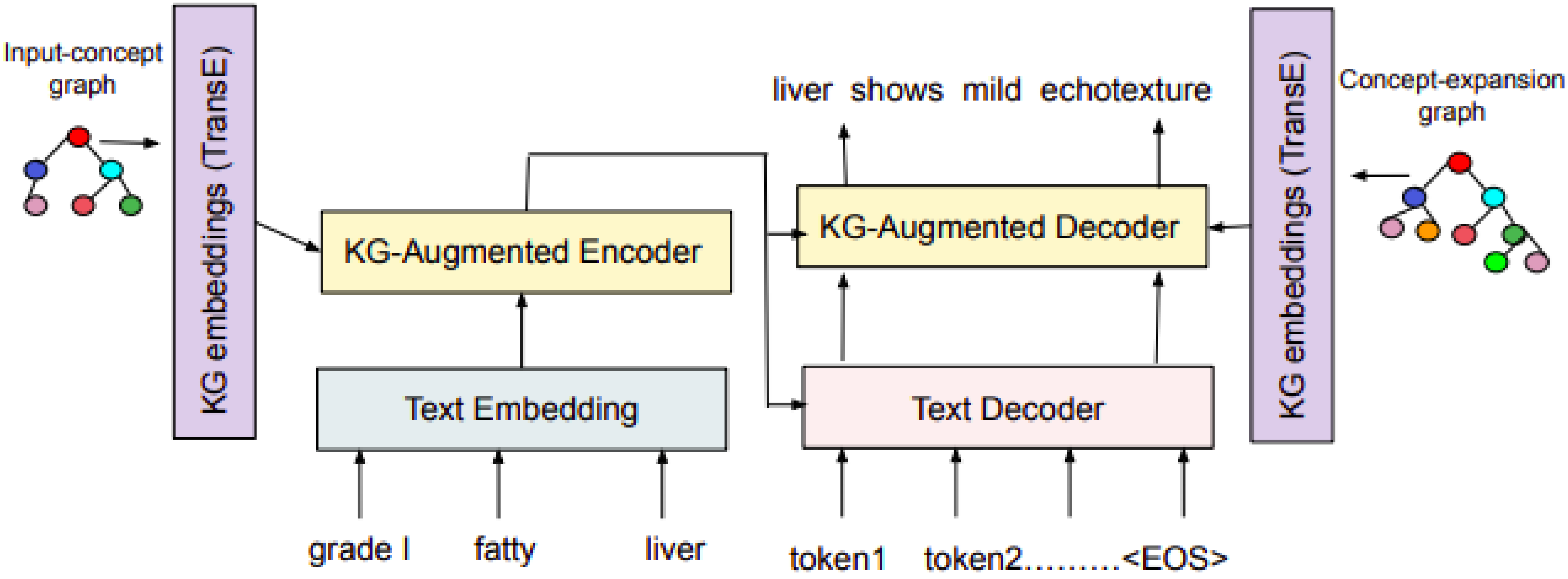
KG-BART Model Input (Concept-expansion Graph Triples):

(calculus,PropertyOf,acute cholecystitis)
(fluid,DefaultPropertyOf,acute cholecystitis)
(cholecystitis,TypeOf,inflammation)
(neck,PartOf,gallbladder)
(wall,PartOf,gallbladder)
(gallbladder,PartOf,upper abdomen)
(>3mm,DefaultPropertyOf,thick)
(thick,PartOf,wall)

Liver Ultrasound KG



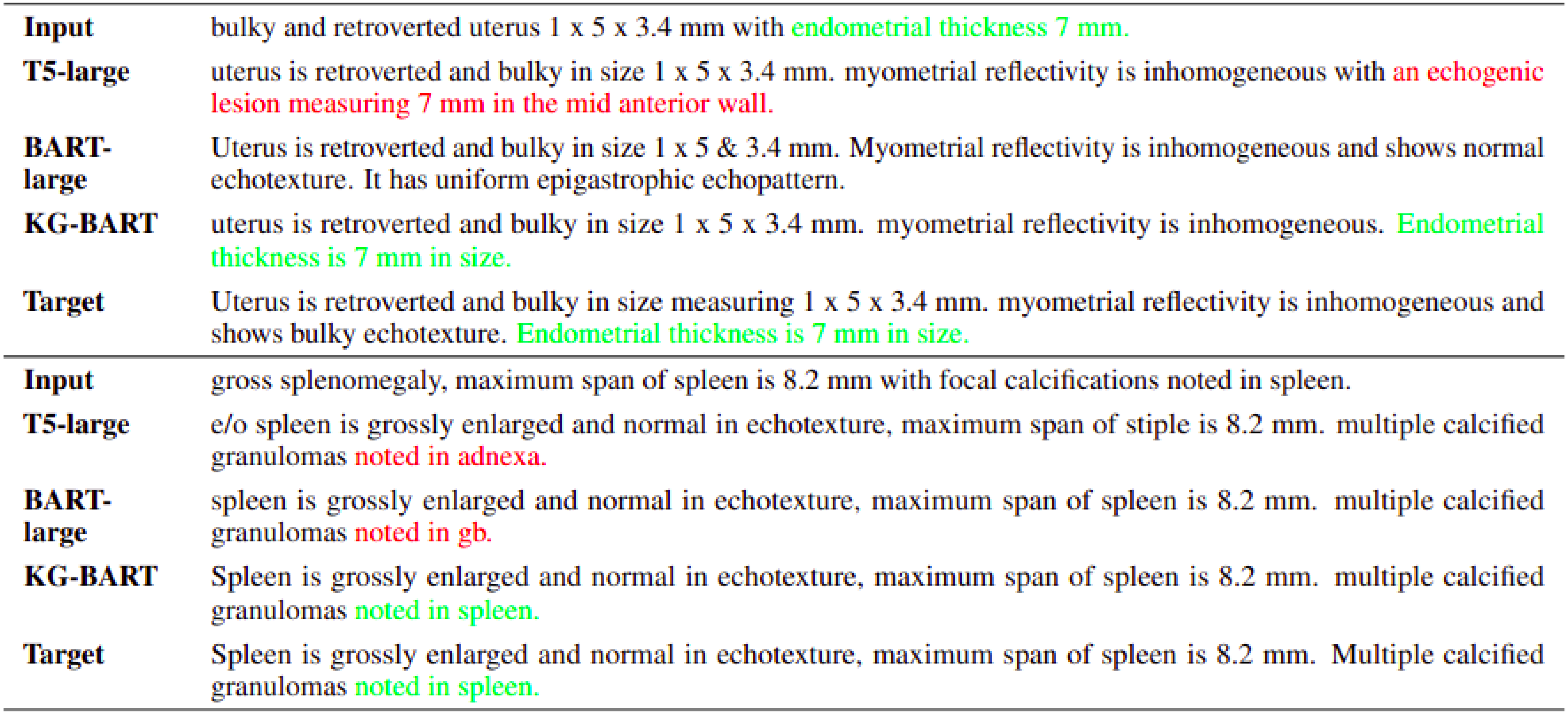
KG-BART (Ultrasound)



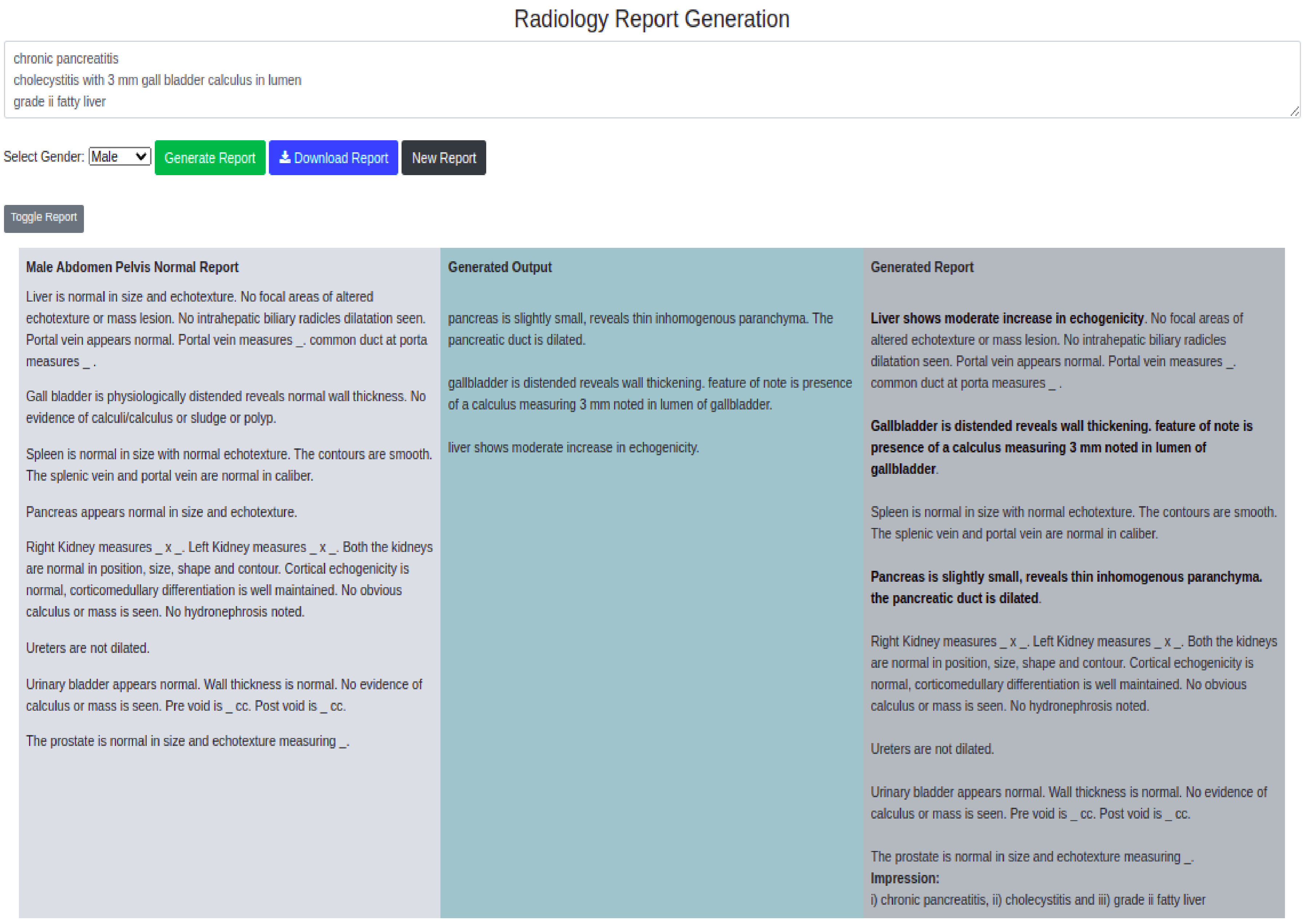
Quantitative Results

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.



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

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

Next

Conversational AI and Empathetic Computer

Harvard Business Review:

https://hbr.org/2015/11/2015-empathy-index

There is a direct link between empathy and commercial success. Businesses are more profitable and productive when they act ethically, treat their staff well, and communicate better with their customers, according to the latest Lady Geek Global Empathy Index. The top 10 companies in the Global Empathy Index 2015 increased in value more than twice as much as the bottom 10 and generated 50% more earnings. Average earnings among the top 10 were up 6% this year, while the average earnings of the bottom 10 dropped 9%. (Last year's empathy index can be found here.)

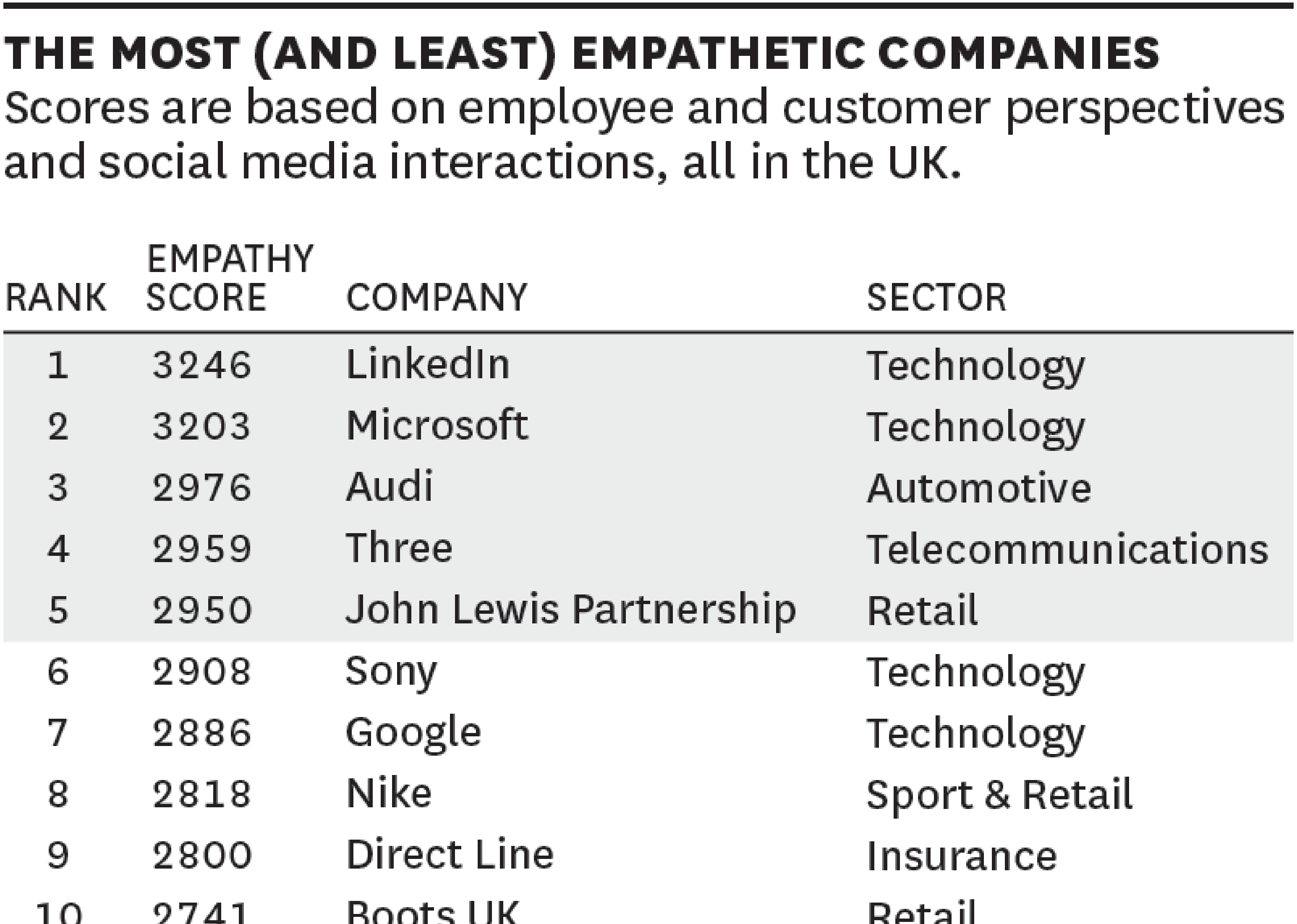
Harvard Business Review:

https://hbr.org/2015/11/2015-empathy-index (1/2)

There is a direct link between empathy and commercial success. Businesses are more profitable and productive when they act ethically, treat their staff well, and communicate better with their customers, according to the latest Lady Geek Global Empathy Index. The top 10 companies in the Global Empathy Index 2015 increased in value more than twice as much as the bottom 10 and generated 50% more earnings. Average earnings among the top 10 were up 6% this year, while the average earnings of the bottom 10 dropped 9%. (Last year’s empathy index can be found here.)

Harvard Business Review:

https://hbr.org/2015/11/2015-empathy-index (2/2)



THE MOST (AND LEAST) EMPATHETIC COMPANIES

Scores are based on employee and customer perspectives and social media interactions, all in the UK.

RANK	EMPATHY SCORE	COMPANY	SECTOR
1	3246	LinkedIn	Technology
2	3203	Microsoft	Technology
3	2976	Audi	Automotive
4	2959	Three	Telecommunications
5	2950	John Lewis Partnership	Retail
6	2908	Sony	Technology
7	2886	Google	Technology
8	2818	Nike	Sport & Retail
9	2800	Direct Line	Insurance
10	2741	Boots UK	Retail

LINK

Wish List

- ***A Conversational AI Agent Must***
  - Understand and respond to Emotion
  - *Decode Sarcasm, Humour*
  - *Understand rare language phenomena like Metaphor, Hyperbole, Irony*
  - *And so on…*

A Multi-task Framework for Hyperbole and Metaphor Detection

Naveen Badathala, Abisek Rajakumar Kalarani, Tejpalsingh Siledar and Pushpak Bhattacharyya, A Match Made in Heaven: A Multi-task Framework for Hyperbole and Metaphor Detection, ACL 2023 Findings, Toronto, July 9-14, 2023.

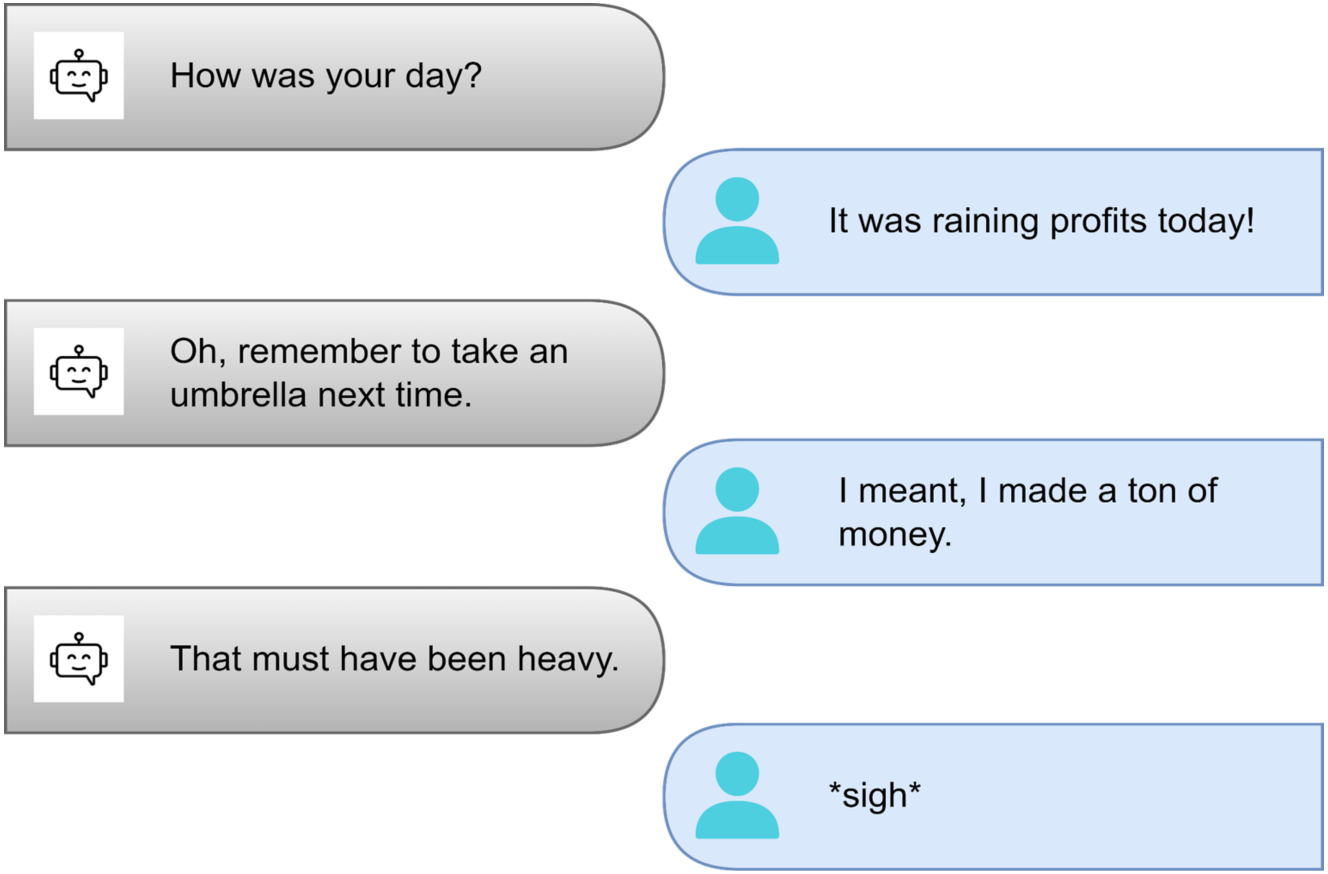
Hyperbole and Metaphor

- A hyperbole is a figurative language in which the literal meaning is exaggerated intentionally
- Metaphor makes implicit comparisons to something that is literally not true

Input	**Hyperbole (0/1)**	**Metaphor (0/1)**
The principal is angry	0	0
I will walk a thousand miles to meet you.	1	0
Life is a journey	0	1
The principal is going to cook us	1	1

Motivation (1/2)

Hyperbole and metaphor are common and their detection is important, e.g., in the following with a chatbot



How was your day?

It was raining profits today!

Oh, remember to take an umbrella next time.

I meant, I made a ton of money.

That must have been heavy.

sigh

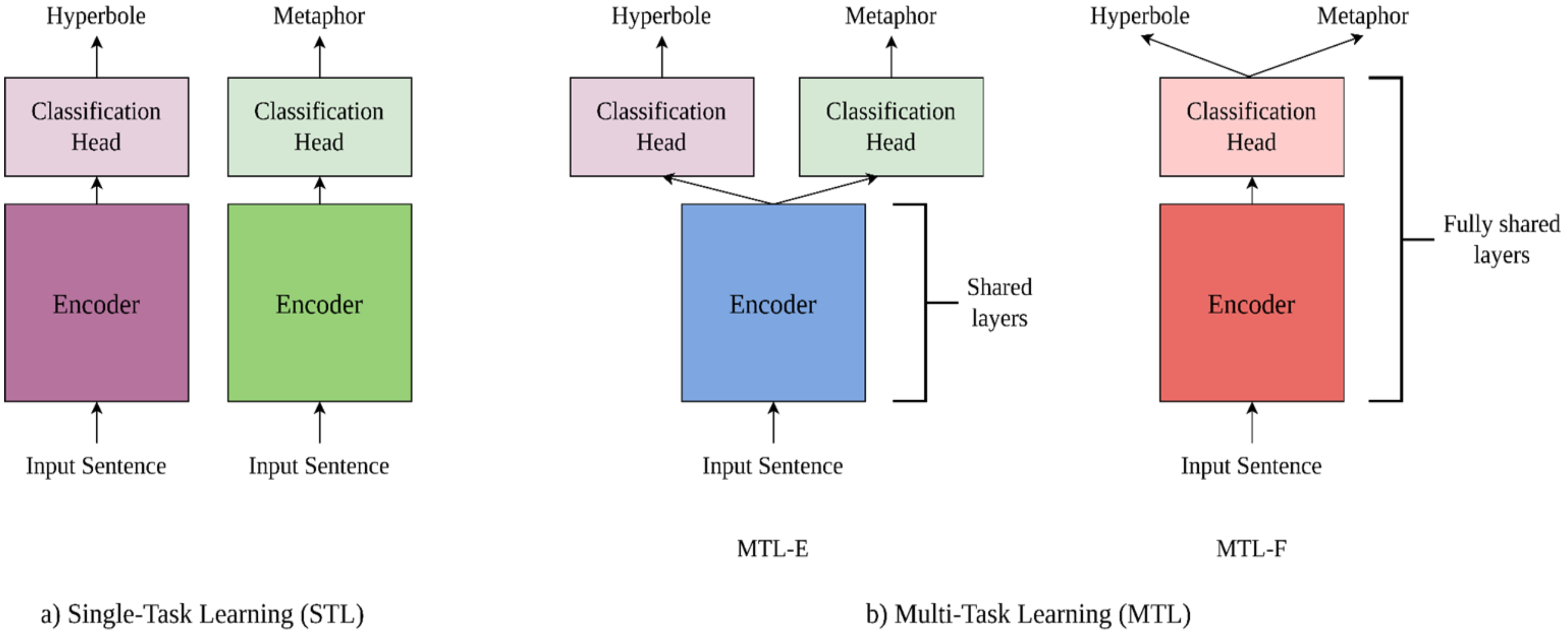
Not a fiction…

Google Assistant/Siri will turn on flashlight in smartphones upon saying **“Lumos maxima**” which is a spell used in the Harry Potter series

Motivation (2/2)

- Shared characteristics of hyperbole and metaphor can be captured at the representational level

Schematic



a) Single-Task Learning (STL)

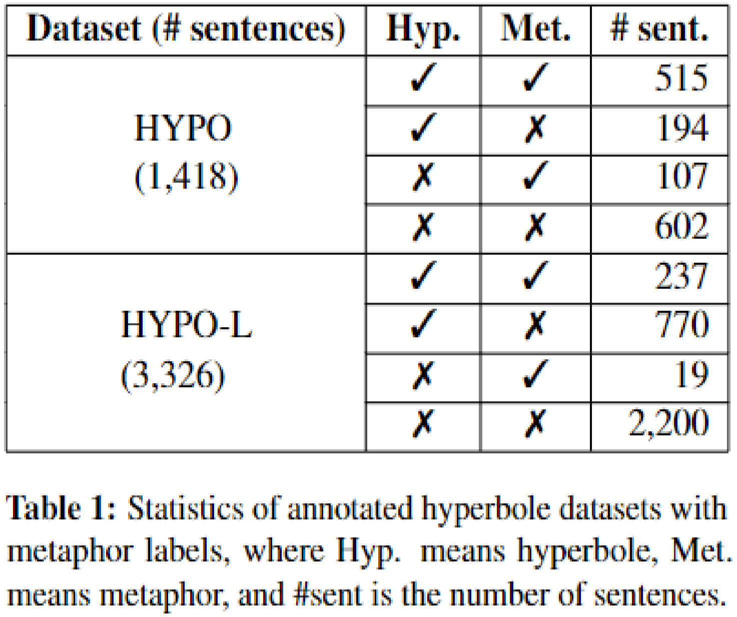
b) Multi-Task Learning (MTL)

Input: Sentence

Output: Hyperbole or Metaphor

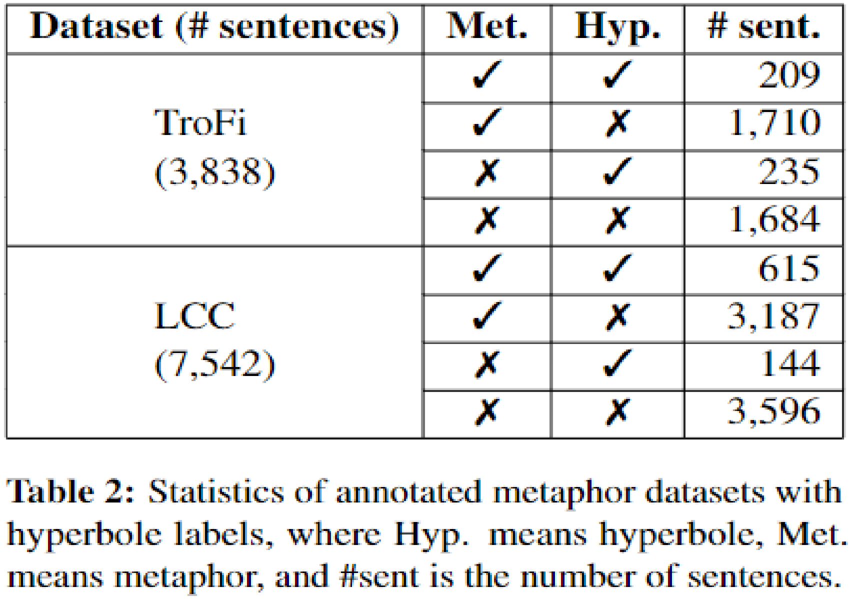
Dataset

- HYPO, HYPO-L (Hyperbole Datasets)
- LCC, TroFi (Metaphor Datasets)



Dataset (# sentences)	Hyp.	Met.	# sent.
HYPO (1,418)	✓	✓	515
	✓	✗	194
	✗	✓	107
	✗	✗	602
HYPO-L (3,326)	✓	✓	237
	✓	✗	770
	✗	✓	19
	✗	✗	2,200

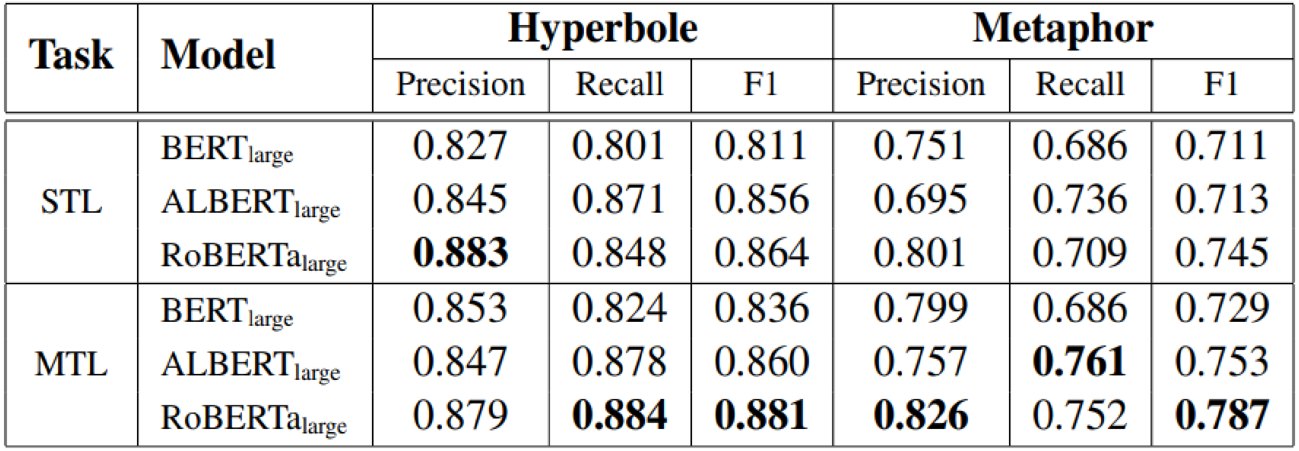
Table 1: Statistics of annotated hyperbole datasets with metaphor labels, where Hyp. means hyperbole, Met. means metaphor, and #sent is the number of sentences.



Dataset (# sentences)	Met.	Hyp.	# sent.
TroFi (3,838)	✓	✓	209
	✓	✗	1,710
	✗	✓	235
	✗	✗	1,684
LCC (7,542)	✓	✓	615
	✓	✗	3,187
	✗	✓	144
	✗	✗	3,596

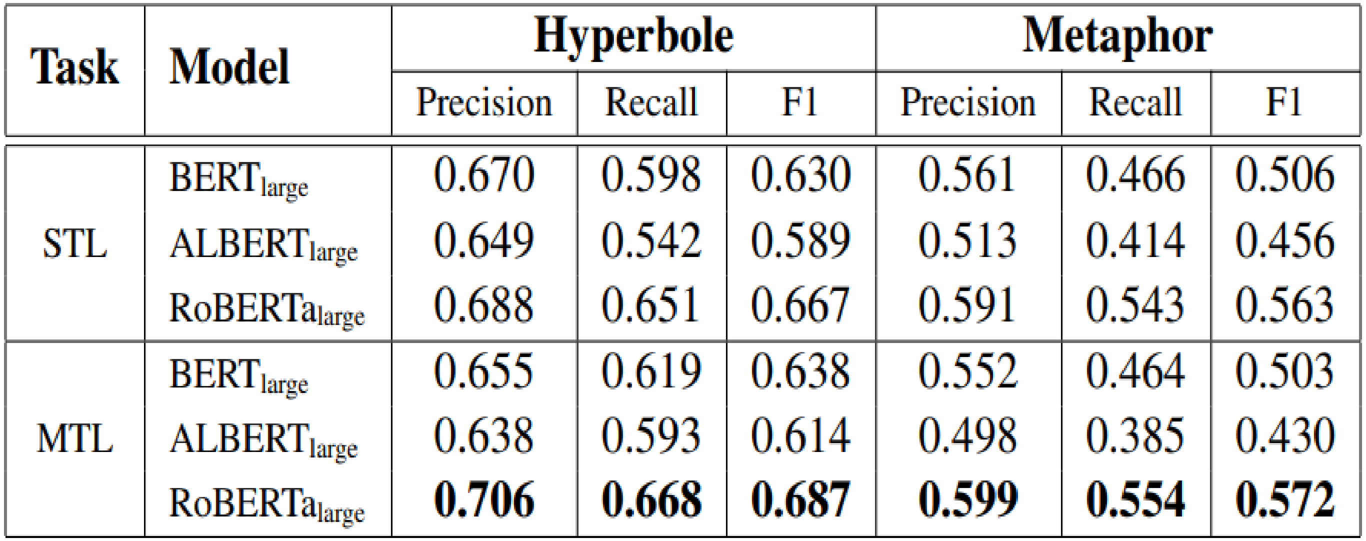
Table 2: Statistics of annotated metaphor datasets with hyperbole labels, where Hyp. means hyperbole, Met. means metaphor, and #sent is the number of sentences.

STL vs. MTL Results - MTL superior (1/2)



Task	Model	Hyperbole			Metaphor		
		Precision	Recall	F1	Precision	Recall	F1
STL	$\text{BERT}_{\text{large}}$	0.827	0.801	0.811	0.751	0.686	0.711
	$\text{ALBERT}_{\text{large}}$	0.845	0.871	0.856	0.695	0.736	0.713
	$\text{RoBERTa}_{\text{large}}$	**0.883**	0.848	0.864	0.801	0.709	0.745
MTL	$\text{BERT}_{\text{large}}$	0.853	0.824	0.836	0.799	0.686	0.729
	$\text{ALBERT}_{\text{large}}$	0.847	0.878	0.860	0.757	**0.761**	0.753
	$\text{RoBERTa}_{\text{large}}$	0.879	**0.884**	**0.881**	**0.826**	0.752	**0.787**

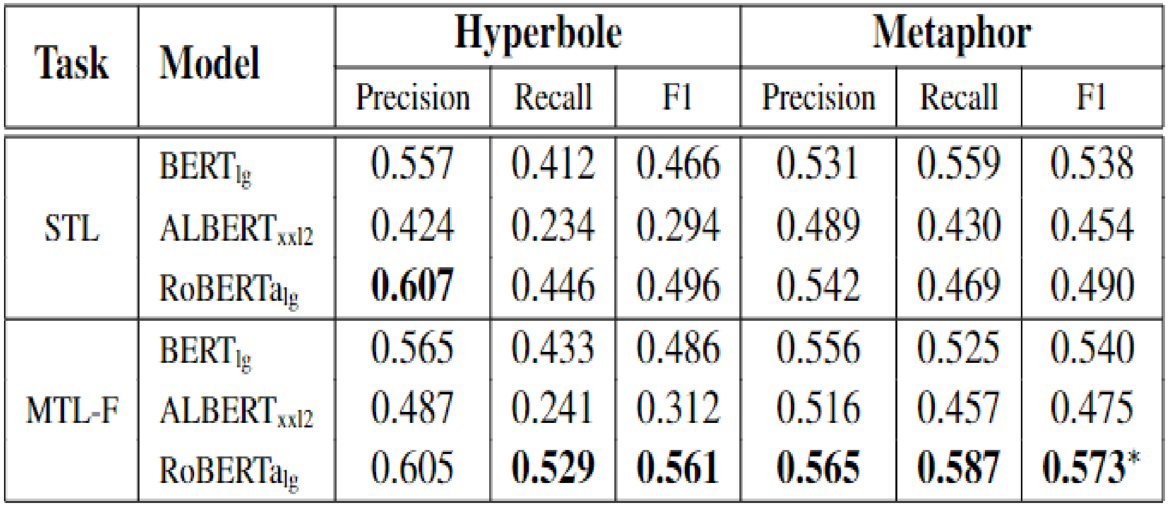
10-fold cross validation over three different runs for hyperbole and metaphor detection task on the **HYPO** dataset.



Task	Model	Hyperbole			Metaphor		
		Precision	Recall	F1	Precision	Recall	F1
STL	$\text{BERT}_{\text{large}}$	0.670	0.598	0.630	0.561	0.466	0.506
	$\text{ALBERT}_{\text{large}}$	0.649	0.542	0.589	0.513	0.414	0.456
	$\text{RoBERTa}_{\text{large}}$	0.688	0.651	0.667	0.591	0.543	0.563
MTL	$\text{BERT}_{\text{large}}$	0.655	0.619	0.638	0.552	0.464	0.503
	$\text{ALBERT}_{\text{large}}$	0.638	0.593	0.614	0.498	0.385	0.430
	$\text{RoBERTa}_{\text{large}}$	**0.706**	**0.668**	**0.687**	**0.599**	**0.554**	**0.572**

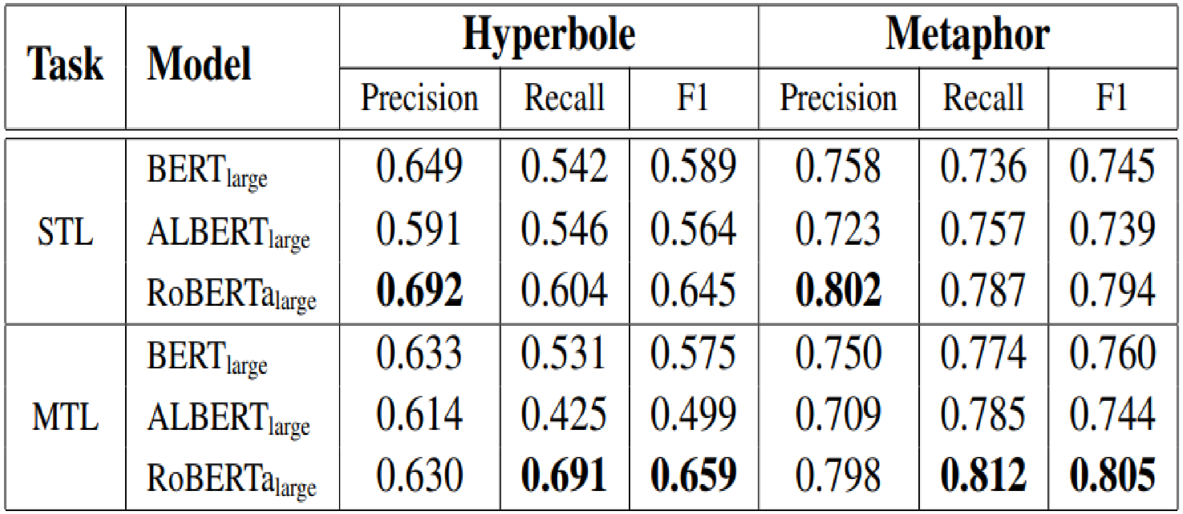
10-fold cross validation over three different runs for hyperbole and metaphor detection task on the **HYPO-L** dataset.

STL vs. MTL Results - MTL superior (2/2)



Task	Model	Hyperbole			Metaphor		
		Precision	Recall	F1	Precision	Recall	F1
STL	BERT$_{lg}$	0.557	0.412	0.466	0.531	0.559	0.538
	ALBERT$_{xxl2}$	0.424	0.234	0.294	0.489	0.430	0.454
	RoBERTa$_{lg}$	**0.607**	0.446	0.496	0.542	0.469	0.490
MTL-F	BERT$_{lg}$	0.565	0.433	0.486	0.556	0.525	0.540
	ALBERT$_{xxl2}$	0.487	0.241	0.312	0.516	0.457	0.475
	RoBERTa$_{lg}$	0.605	**0.529**	**0.561**	**0.565**	**0.587**	**0.573***

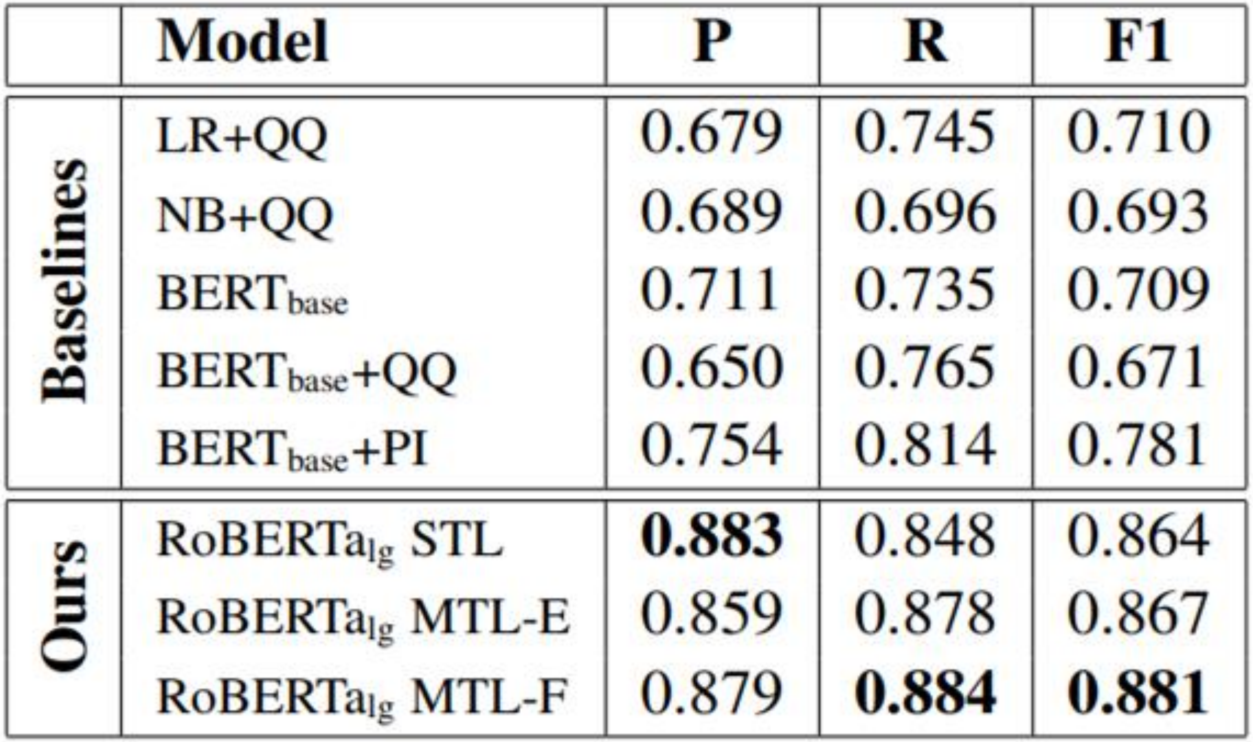
10-fold cross validation over three different runs for hyperbole and metaphor detection task on the label balanced **TroFi** dataset.



Task	Model	Hyperbole			Metaphor		
		Precision	Recall	F1	Precision	Recall	F1
STL	BERT$_{large}$	0.649	0.542	0.589	0.758	0.736	0.745
	ALBERT$_{large}$	0.591	0.546	0.564	0.723	0.757	0.739
	RoBERTa$_{large}$	**0.692**	0.604	0.645	**0.802**	0.787	0.794
MTL	BERT$_{large}$	0.633	0.531	0.575	0.750	0.774	0.760
	ALBERT$_{large}$	0.614	0.425	0.499	0.709	0.785	0.744
	RoBERTa$_{large}$	0.630	**0.691**	**0.659**	0.798	**0.812**	**0.805**

10-fold cross validation over three different runs for hyperbole and metaphor detection task on the label balanced **LCC** dataset.

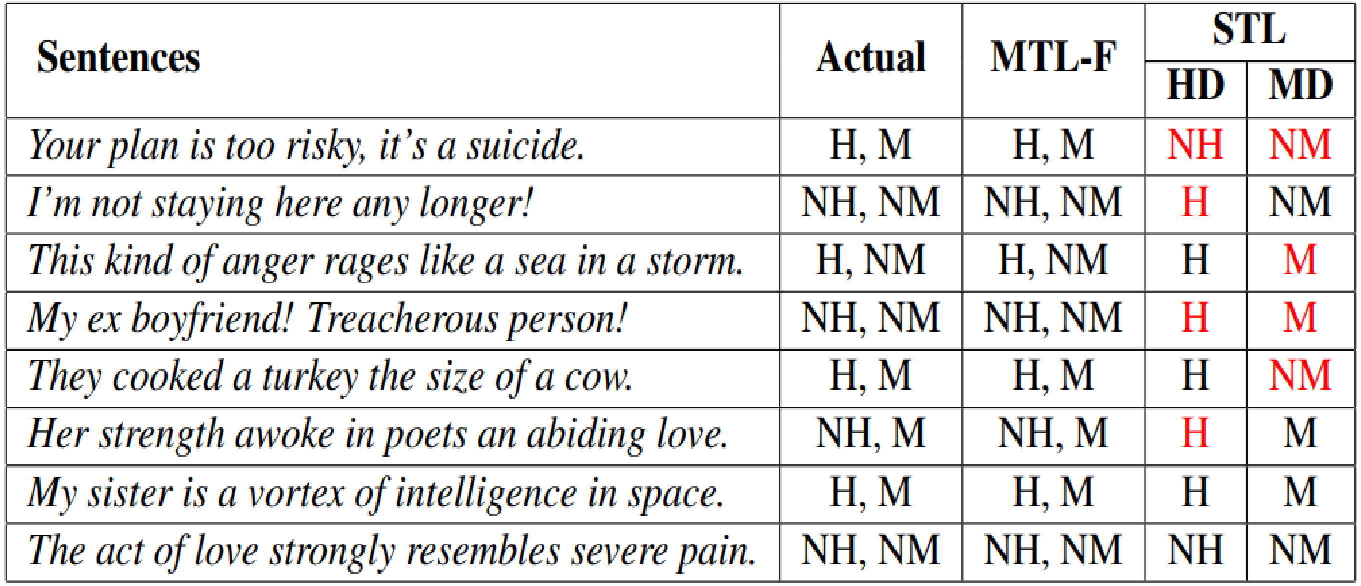
Hyperbole Detection Results - New SOTA!



	Model	P	R	F1
Baselines	LR+QQ	0.679	0.745	0.710
	NB+QQ	0.689	0.696	0.693
	BERT$_{\text{base}}$	0.711	0.735	0.709
	BERT$_{\text{base}}$+QQ	0.650	0.765	0.671
	BERT$_{\text{base}}$+PI	0.754	0.814	0.781
Ours	RoBERTa$_{\text{lg}}$ STL	**0.883**	0.848	0.864
	RoBERTa$_{\text{lg}}$ MTL-E	0.859	0.878	0.867
	RoBERTa$_{\text{lg}}$ MTL-F	0.879	**0.884**	**0.881**

Precision (P), recall (R) and F1 score for baseline models compared to our work on HYPO dataset

Qualitative Analysis

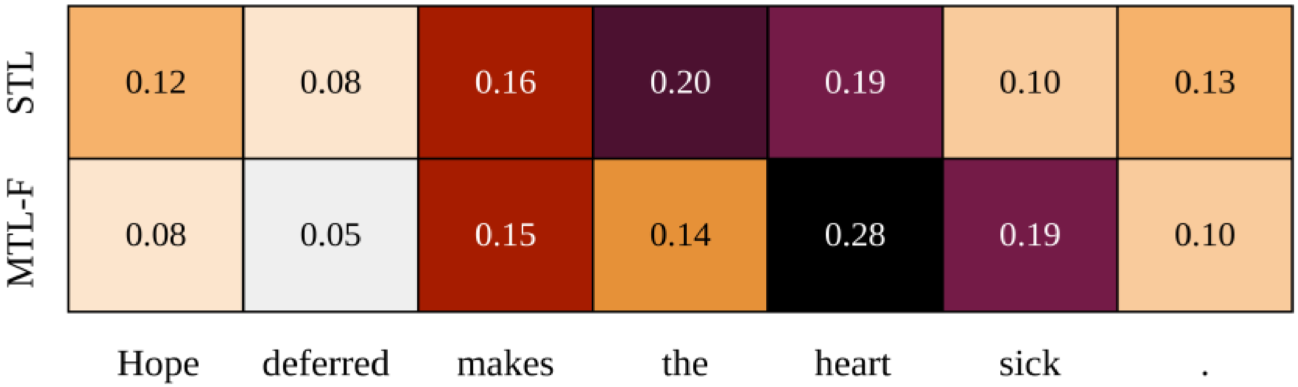


Sentences	Actual	MTL-F	STL	
			HD	MD
Your plan is too risky, it's a suicide.	H, M	H, M	NH	NM
I'm not staying here any longer!	NH, NM	NH, NM	H	NM
This kind of anger rages like a sea in a storm.	H, NM	H, NM	H	M
My ex boyfriend! Treacherous person!	NH, NM	NH, NM	H	M
They cooked a turkey the size of a cow.	H, M	H, M	H	NM
Her strength awoke in poets an abiding love.	NH, M	NH, M	H	M
My sister is a vortex of intelligence in space.	H, M	H, M	H	M
The act of love strongly resembles severe pain.	NH, NM	NH, NM	NH	NM

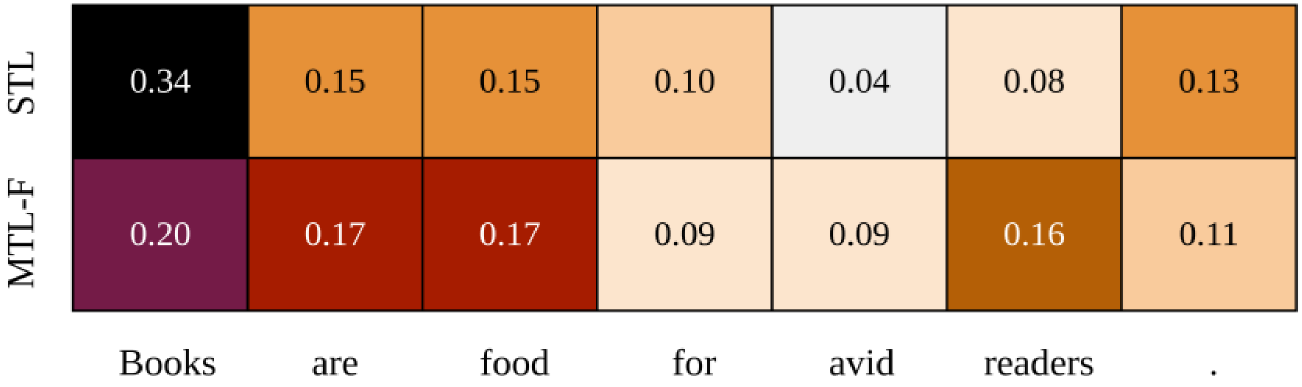
Cases where the MTL-F performs better than the STL. H-Hyperbole, NH-Non Hyperbole, M-Metaphor, NM-Non Metaphor. Red indicates incorrect detection.

- MTL consistently outperforms STL
- We investigate all four combinations of hyperbole and metaphor cases and observe a similar pattern

Attention Analysis:
darker→ higher attention



	Hope	deferred	makes	the	heart	sick	.
STL	0.12	0.08	0.16	0.20	0.19	0.10	0.13
MTL-F	0.08	0.05	0.15	0.14	0.28	0.19	0.10



	Books	are	food	for	avid	readers	.
STL	0.34	0.15	0.15	0.10	0.04	0.08	0.13
MTL-F	0.20	0.17	0.17	0.09	0.09	0.16	0.11

Summary

- Discussed nature of NLP
- ML steps in to do ambiguity resolution
- LLMs ubiquitous
- Depending on resource availability: lightest to heaviest:
- LLM+prompts→ adapater→ fine_tuning→ Pretraining
- ChatGPT currently performing the best
- Applications in health, rare language pehnomena like metaphor

Link

Demos

https://e420-103-21-125-76.ngrok-free.app/

https://www.cfilt.iitb.ac.in/ssmt/speech2speech

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