Natural Language Processing

CS 3216/UG, AI 5203/PG

Week-1

Mahindra™ University

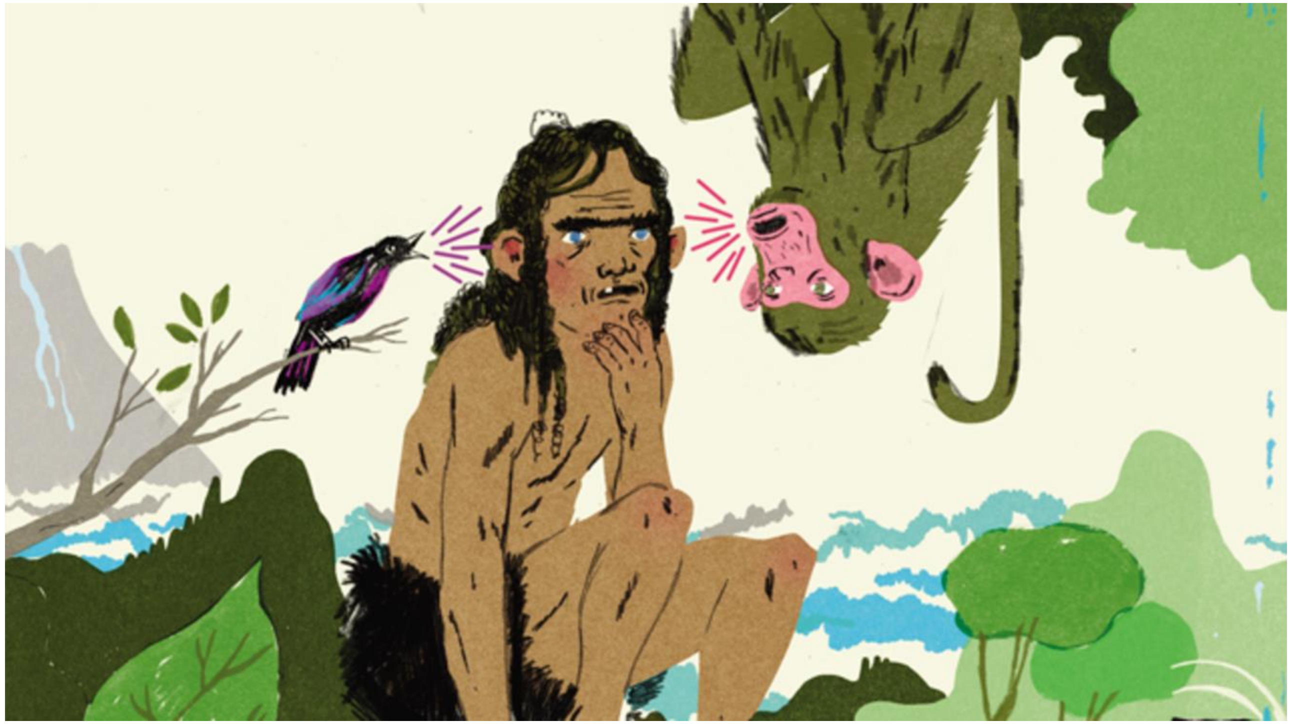
Global Thinkers. Engaged Leaders.

Who am I?

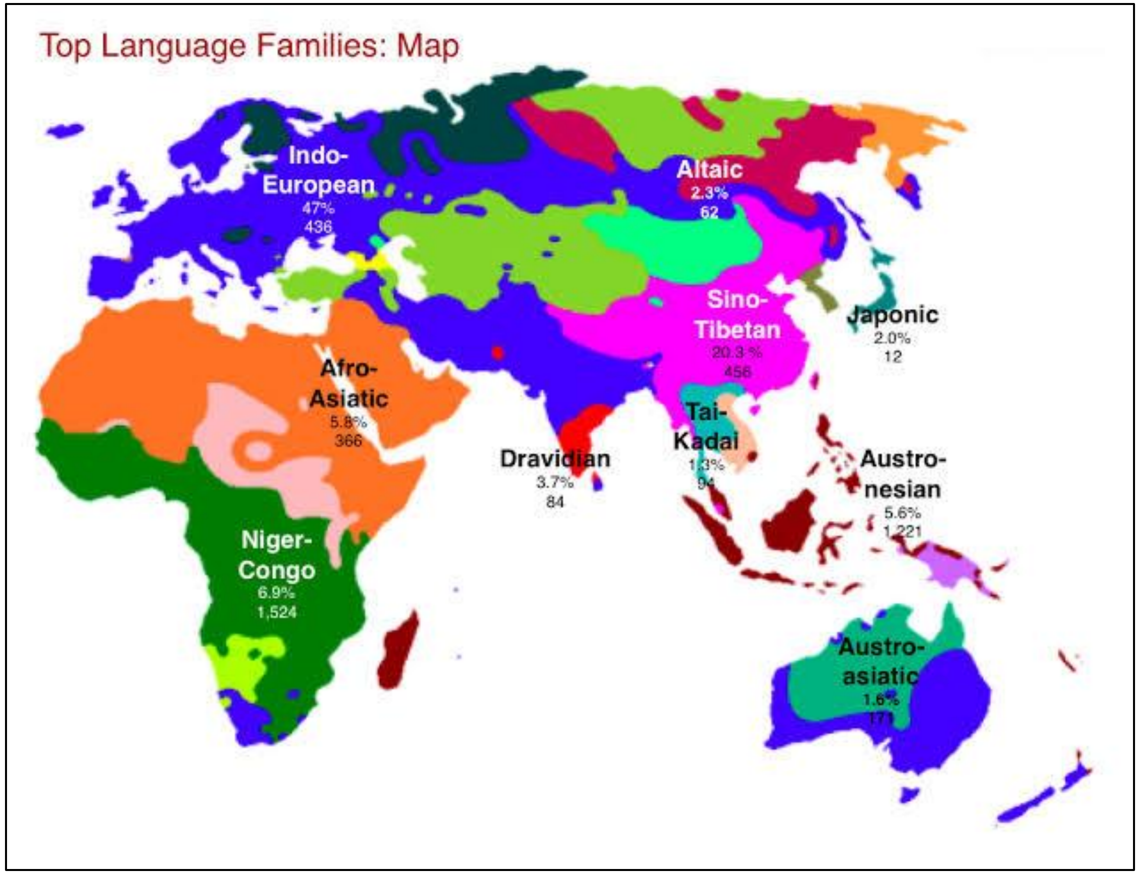
- Current Research at Mahindra University, [VLaNC lab]- NLP, Graphs, Multimodal- Computer Vision, language, Computational Neuroscience, Brain Computer Interfaces, Generative AI and solving real-world problems
- Research Experience
  - Applied researcher in NLP, Knowledge Graphs, Graph Neural Networks ~ 6 years [Research scholar -IIIT-Delhi and Naukri.com]
- Teaching Experience
  - Former Assistant Professor, GGSIPU, New Delhi
  - Former Assistant Professor, DCE, Gurgaon
  - TA for many courses- COA, Privacy and security on Online social media @IIIT, NPTEL, NLP course PGDSAAI@IIIT-Delhi

Nidhi Goyal
(Assistant Professor, Ecole School of Engineering)

What do you see here?

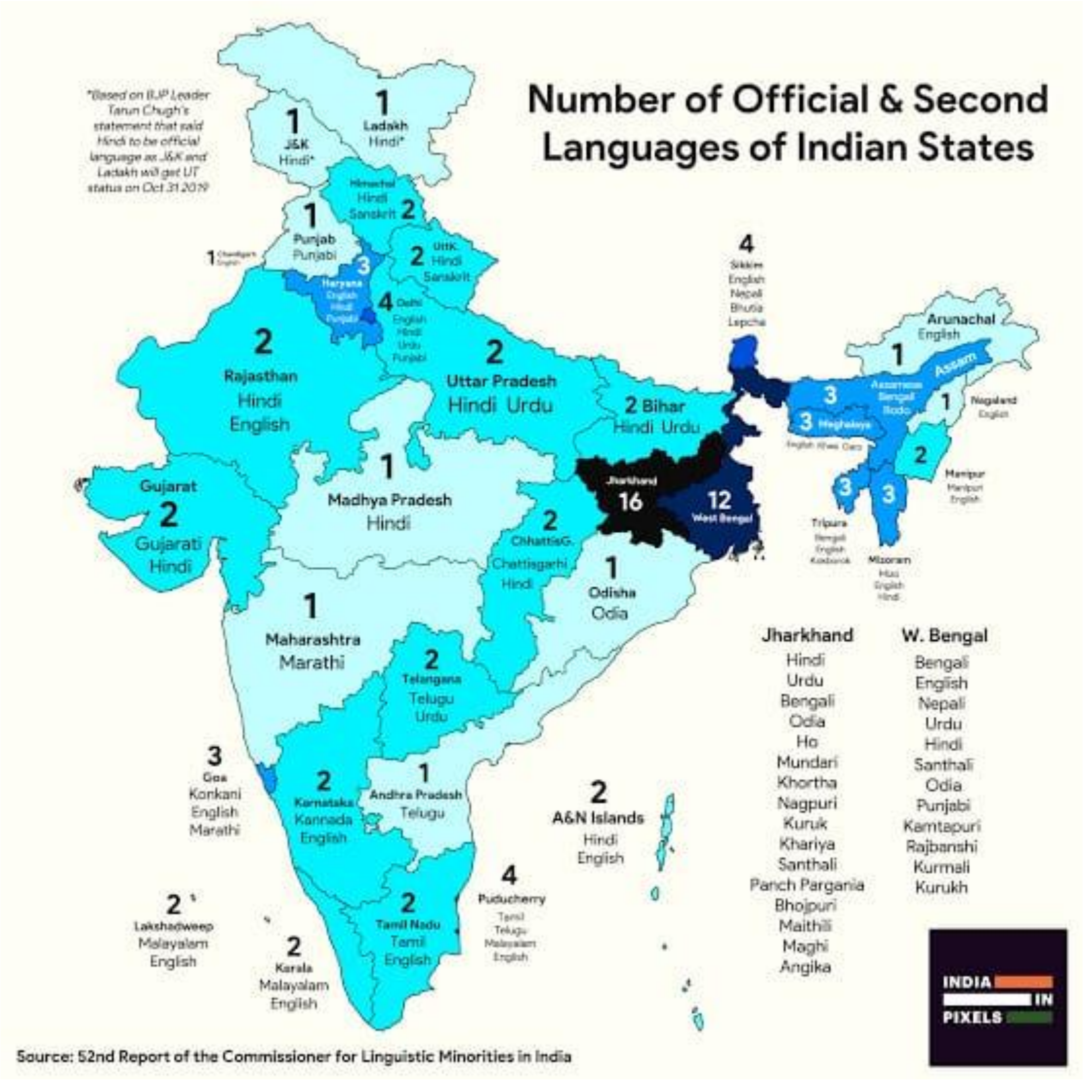


Diversity of languages in this world



https://triangulations.wordpress.com/2013/11/22/language-families/

Variations across Indian languages



Number of Official & Second Languages of Indian States

*Based on BJP Leader Tarun Chugh's statement that said Hindi to be official language as J&K and Ladakh will get UT status on Oct 31 2019
1 J&K Hindi*
1 Ladakh Hindi*
Himachal Hindi Sanskrit 2
1 Punjab Punjabi
1 Chandigarh English
2 Uttk. Hindi Sanskrit
3 Haryana English Hindi Punjabi
4 Delhi English Hindi Urdu Punjabi
2 Rajasthan Hindi English
2 Uttar Pradesh Hindi Urdu
2 Bihar Hindi Urdu
4 Sikkim English Nepali Bhutia Lepcha
Arunachal English 1
Assam
3 Assamese Bengali Bodo
3 Meghalaya English Khasi Garo
1 Nagaland English
2 Manipur Manipuri English
3 Tripura Bengali English Kokborok
3 Mizoram Mizo English Hindi
Gujarat 2 Gujarati Hindi
1 Madhya Pradesh Hindi
Jharkhand 16
12 West Bengal
2 ChhattisG. Chattisgarhi Hindi
1 Odisha Odia
1 Maharashtra Marathi
2 Telangana Telugu Urdu
3 Goa Konkani English Marathi
2 Karnataka Kannada English
1 Andhra Pradesh Telugu
2 A&N Islands Hindi English
4 Puducherry Tamil Telugu Malayalam English
2 Lakshadweep Malayalam English
2 Kerala Malayalam English
2 Tamil Nadu Tamil English
Jharkhand
Hindi
Urdu
Bengali
Odia
Ho
Mundari
Khortha
Nagpuri
Kuruk
Khariya
Santhali
Panch Pargania
Bhojpuri
Maithili
Maghi
Angika
W. Bengal
Bengali
English
Nepali
Urdu
Hindi
Santhali
Odia
Punjabi
Kamtapuri
Rajbanshi
Kurmali
Kurukh
INDIA IN PIXELS

Source: 52nd Report of the Commissioner for Linguistic Minorities in India

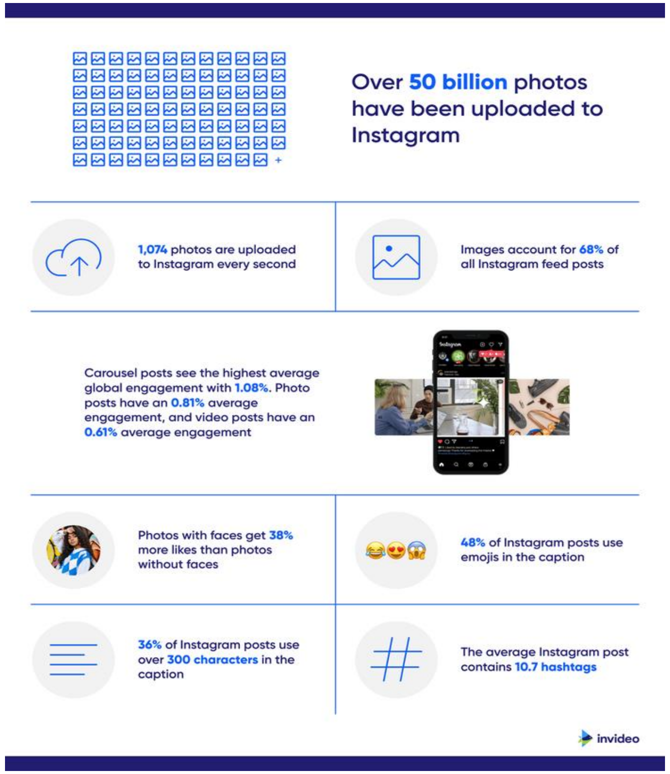
Source: https://twitter.com/indiainpixels/status/1299284220687167488

Natural language

In neuropsychology, linguistics, and philosophy of language,

a natural language is any language that occurs naturally in a human community by a process of use, repetition, and change without conscious planning or premeditation.

Data Data everywhere!!!!



Over **50 billion** photos have been uploaded to Instagram

1,074 photos are uploaded to Instagram every second

Images account for **68%** of all Instagram feed posts

Carousel posts see the highest average global engagement with **1.08%**. Photo posts have an **0.81%** average engagement, and video posts have an **0.61%** average engagement

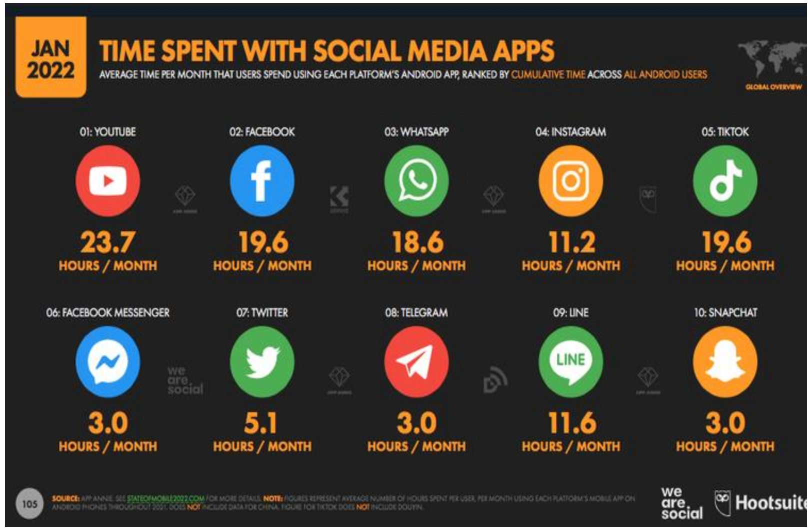
Photos with faces get **38%** more likes than photos without faces

48% of Instagram posts use emojis in the caption 😂😍😱

36% of Instagram posts use over **300 characters** in the caption

The average Instagram post contains **10.7 hashtags**

invideo



JAN 2022

TIME SPENT WITH SOCIAL MEDIA APPS

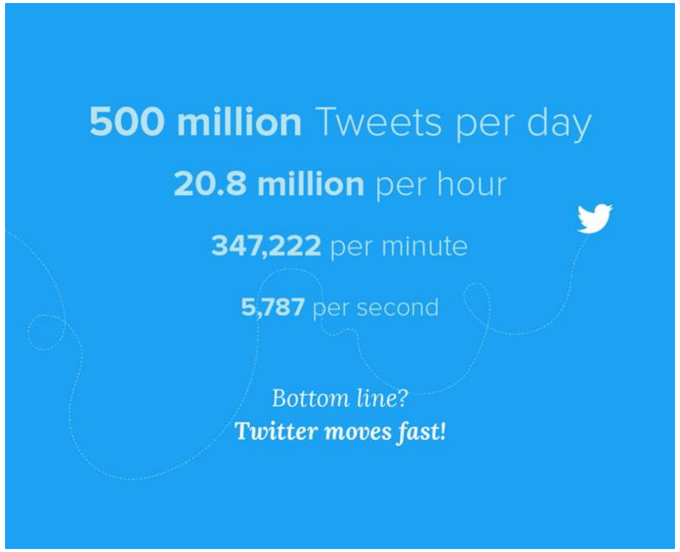
AVERAGE TIME PER MONTH THAT USERS SPEND USING EACH PLATFORM'S ANDROID APP, RANKED BY CUMULATIVE TIME ACROSS ALL ANDROID USERS

GLOBAL OVERVIEW

01: YOUTUBE	02: FACEBOOK	03: WHATSAPP	04: INSTAGRAM	05: TIKTOK
23.7 HOURS / MONTH	19.6 HOURS / MONTH	18.6 HOURS / MONTH	11.2 HOURS / MONTH	19.6 HOURS / MONTH
06: FACEBOOK MESSENGER	**07: TWITTER**	**08: TELEGRAM**	**09: LINE**	**10: SNAPCHAT**
3.0 HOURS / MONTH	5.1 HOURS / MONTH	3.0 HOURS / MONTH	11.6 HOURS / MONTH	3.0 HOURS / MONTH

105 SOURCE: APP ANNIE. SEE STATEOFMOBILE2022.COM FOR MORE DETAILS. NOTE: FIGURES REPRESENT AVERAGE NUMBER OF HOURS SPENT PER USER, PER MONTH USING EACH PLATFORM'S MOBILE APP ON ANDROID PHONES THROUGHOUT 2021. DOES NOT INCLUDE DATA FOR CHINA. FIGURE FOR TIKTOK DOES NOT INCLUDE DOUYIN.

we are social Hootsuite



500 million Tweets per day

20.8 million per hour

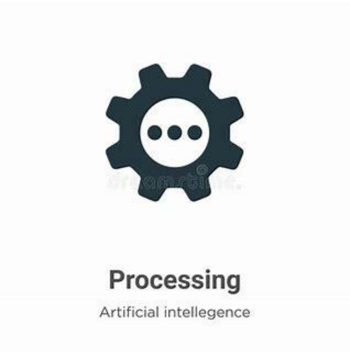
347,222 per minute

5,787 per second

Bottom line?
Twitter moves fast!

Natural language processing?

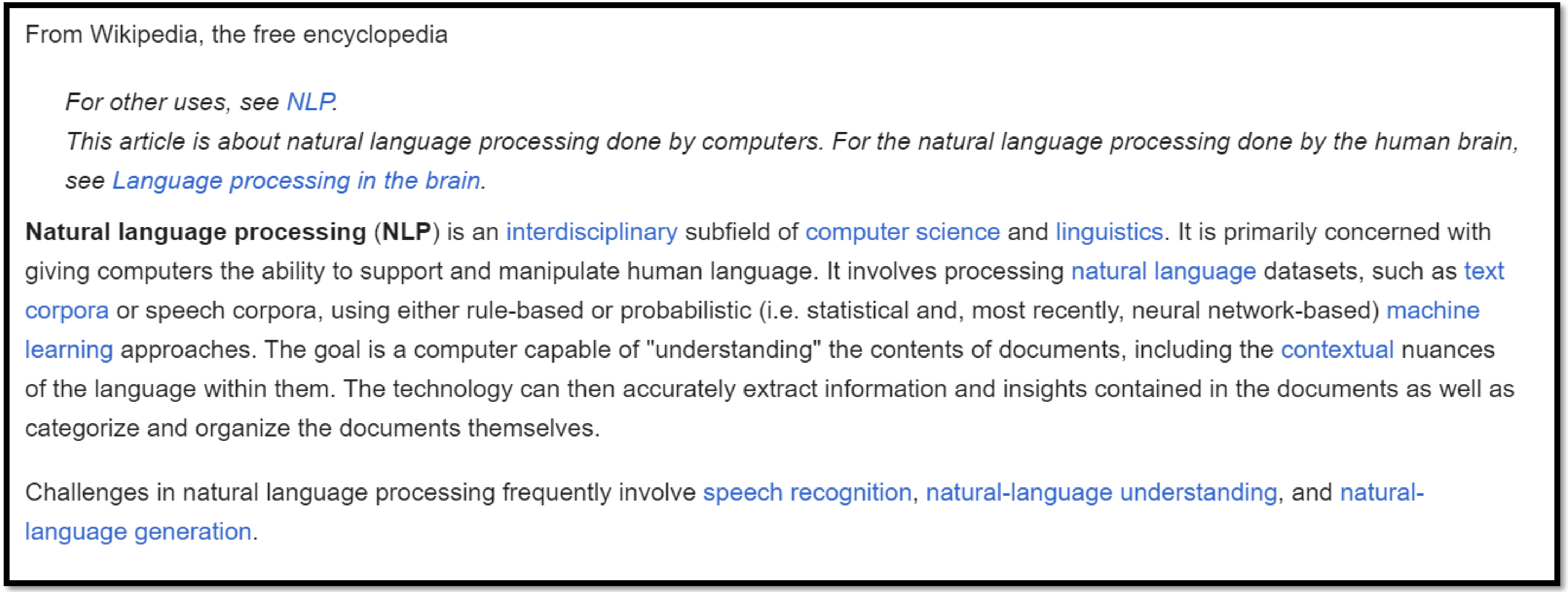
यत्तद् पश्यन्नात्मनि योगसेवया ।
वायं स्थितश्चलति बुद्धिग्राह्यमतीन्द्रियम् तुष्यति
लाभं मन्यते तत्त्वतः
दुःखेन नाधिकं ततः
योगं गुरुणापि योगसं वि



Natural language understanding

Natural language generation

Natural language processing



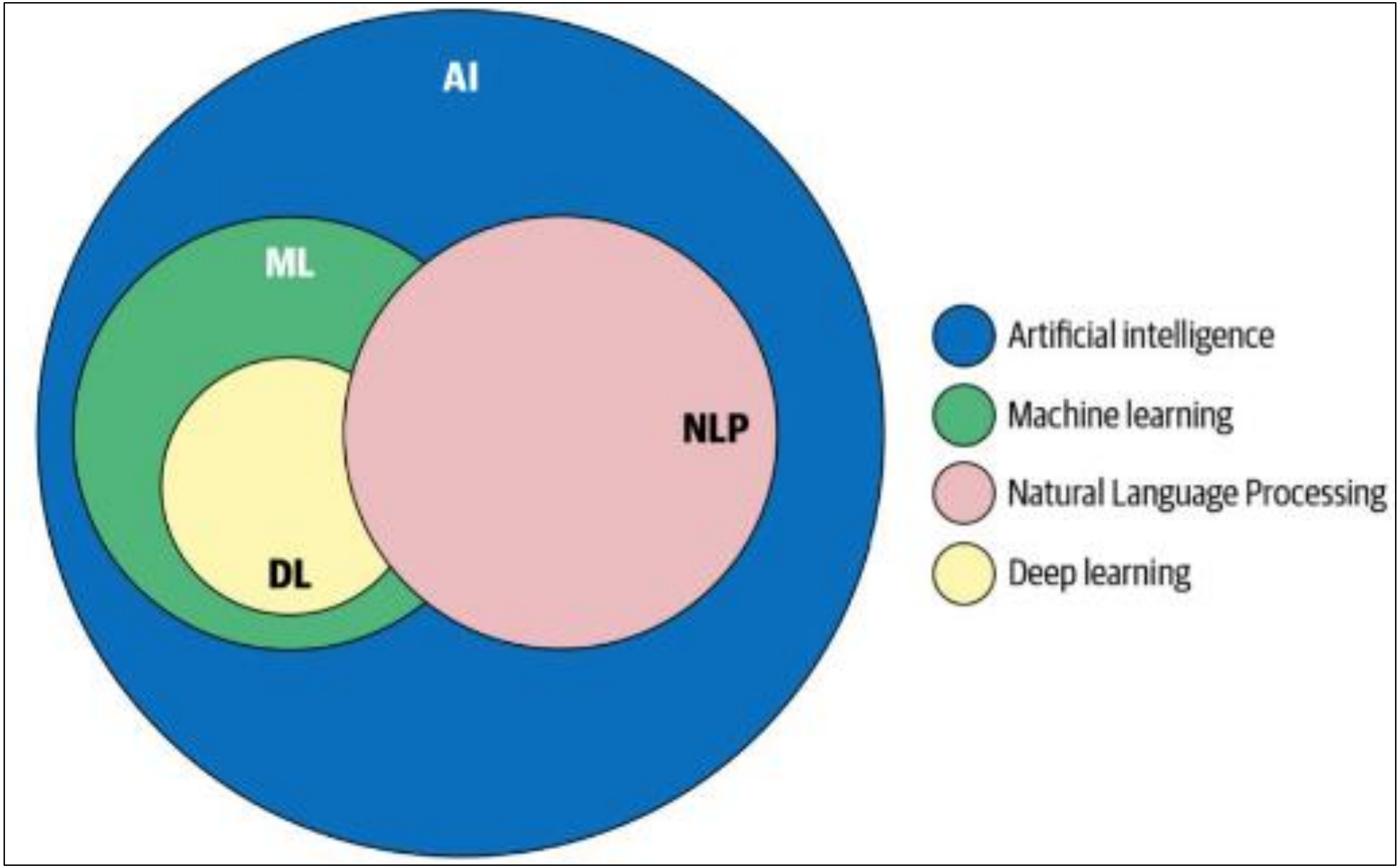
From Wikipedia, the free encyclopedia

For other uses, see NLP.
This article is about natural language processing done by computers. For the natural language processing done by the human brain, see Language processing in the brain.

Natural language processing (**NLP**) is an interdisciplinary subfield of computer science and linguistics. It is primarily concerned with giving computers the ability to support and manipulate human language. It involves processing natural language datasets, such as text corpora or speech corpora, using either rule-based or probabilistic (i.e. statistical and, most recently, neural network-based) machine learning approaches. The goal is a computer capable of "understanding" the contents of documents, including the contextual nuances of the language within them. The technology can then accurately extract information and insights contained in the documents as well as categorize and organize the documents themselves.

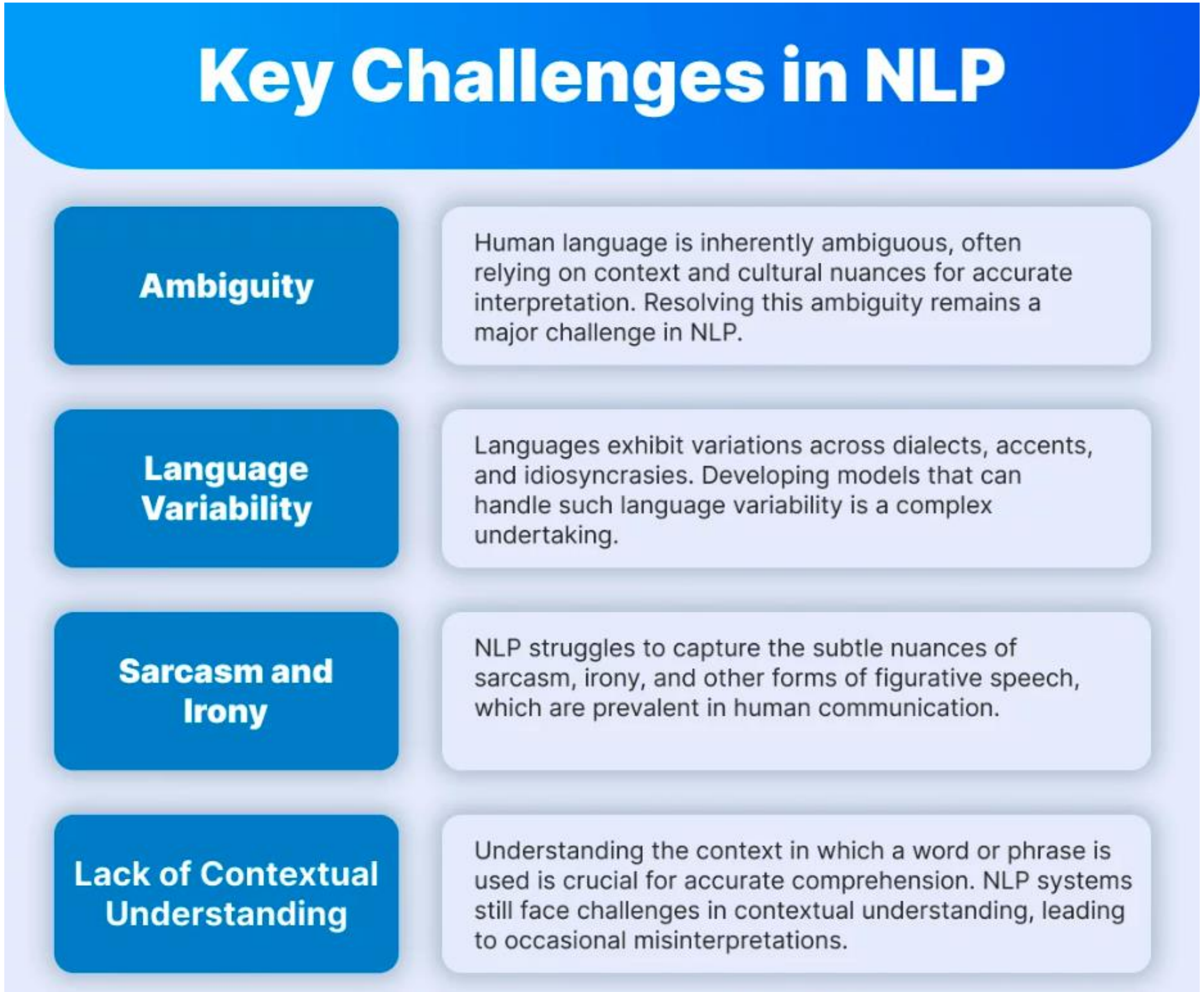
Challenges in natural language processing frequently involve speech recognition, natural-language understanding, and natural-language generation.

https://en.wikipedia.org/wiki/Natural_language_processing



https://www.oreilly.com/library/view/practical-natural-language/9781492054047/

Challenges



Key Challenges in NLP

Ambiguity

Human language is inherently ambiguous, often relying on context and cultural nuances for accurate interpretation. Resolving this ambiguity remains a major challenge in NLP.

Language Variability

Languages exhibit variations across dialects, accents, and idiosyncrasies. Developing models that can handle such language variability is a complex undertaking.

Sarcasm and Irony

NLP struggles to capture the subtle nuances of sarcasm, irony, and other forms of figurative speech, which are prevalent in human communication.

Lack of Contextual Understanding

Understanding the context in which a word or phrase is used is crucial for accurate comprehension. NLP systems still face challenges in contextual understanding, leading to occasional misinterpretations.

Ambiguity in NLP

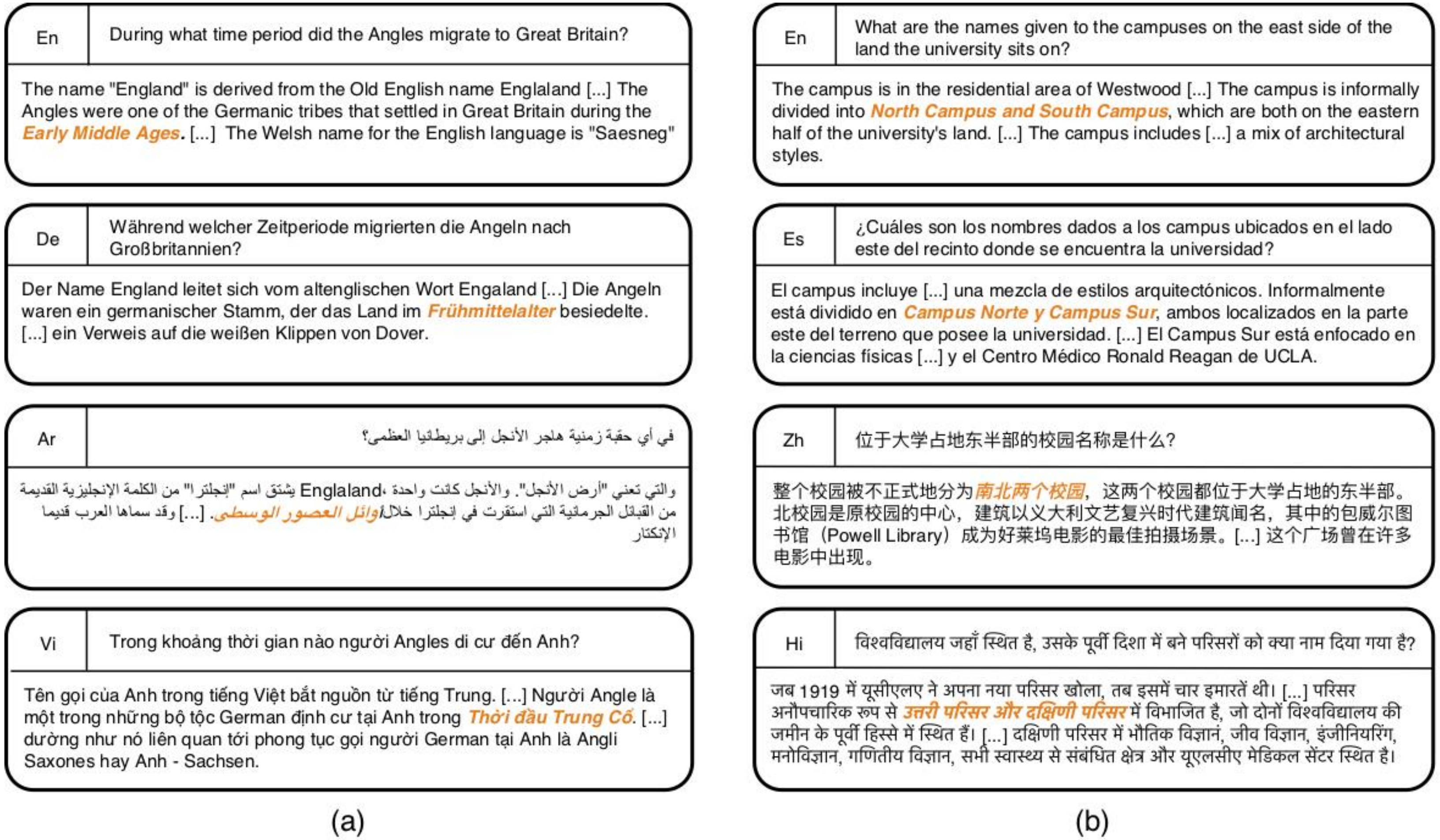
Lexical Ambiguity: This type of ambiguity represents words that can have multiple assertions. For instance, in English, the word “back” can be a noun (back stage), an adjective (back door) or an adverb (back away).

Syntactic Ambiguity: This type of ambiguity represents sentences that can be parsed in multiple syntactical forms. Take the following sentence: “ I heard his cell phone ring in my office”. The propositional phrase “in my office” can be parsed in a way that modifies the noun or on another way that modifies the verb.

Semantic Ambiguity: This type of ambiguity is typically related to the interpretation of sentence. For instance, the previous sentence used in the previous point can be interpreted as if I was physically present in the office or as if the cell phone was in the office.

Metonymy: Arguably, the most difficult type of ambiguity, metonymy deals with phrases in which the literal meaning is different from the figurative assertion. For instance, when we say “Samsung us screaming for new management”, we don’t really mean that the company is literally screaming (although you never know with Samsung these days ;)).

Multilingual



En

During what time period did the Angles migrate to Great Britain?

The name "England" is derived from the Old English name Englaland [...] The Angles were one of the Germanic tribes that settled in Great Britain during the ***Early Middle Ages***. [...] The Welsh name for the English language is "Saesneg"

De

Während welcher Zeitperiode migrierten die Angeln nach Großbritannien?

Der Name England leitet sich vom altenglischen Wort Engaland [...] Die Angeln waren ein germanischer Stamm, der das Land im ***Frühmittelalter*** besiedelte. [...] ein Verweis auf die weißen Klippen von Dover.

Ar

في أي حقبة زمنية هاجر الأنجل إلى بريطانيا العظمى؟

والتي تعني "أرض الأنجل". والأنجل كانت واحدة ،Englaland يشتق اسم "إنجلترا" من الكلمة الإنجليزية القديمة
من القبائل الجرمانية التي استقرت في إنجلترا خلال***أوائل العصور الوسطى***. [...] وقد سماها العرب قديما
الإنكتار

Vi

Trong khoảng thời gian nào người Angles di cư đến Anh?

Tên gọi của Anh trong tiếng Việt bắt nguồn từ tiếng Trung. [...] Người Angle là một trong những bộ tộc German định cư tại Anh trong ***Thời đầu Trung Cổ***. [...] dường như nó liên quan tới phong tục gọi người German tại Anh là Angli Saxones hay Anh - Sachsen.

(a)

En

What are the names given to the campuses on the east side of the land the university sits on?

The campus is in the residential area of Westwood [...] The campus is informally divided into ***North Campus and South Campus***, which are both on the eastern half of the university's land. [...] The campus includes [...] a mix of architectural styles.

Es

¿Cuáles son los nombres dados a los campus ubicados en el lado este del recinto donde se encuentra la universidad?

El campus incluye [...] una mezcla de estilos arquitectónicos. Informalmente está dividido en ***Campus Norte y Campus Sur***, ambos localizados en la parte este del terreno que posee la universidad. [...] El Campus Sur está enfocado en la ciencias físicas [...] y el Centro Médico Ronald Reagan de UCLA.

Zh

位于大学占地东半部的校园名称是什么?

整个校园被不正式地分为***南北两个校园***，这两个校园都位于大学占地的东半部。北校园是原校园的中心，建筑以义大利文艺复兴时代建筑闻名，其中的包威尔图书馆（Powell Library）成为好莱坞电影的最佳拍摄场景。[...] 这个广场曾在许多电影中出现。

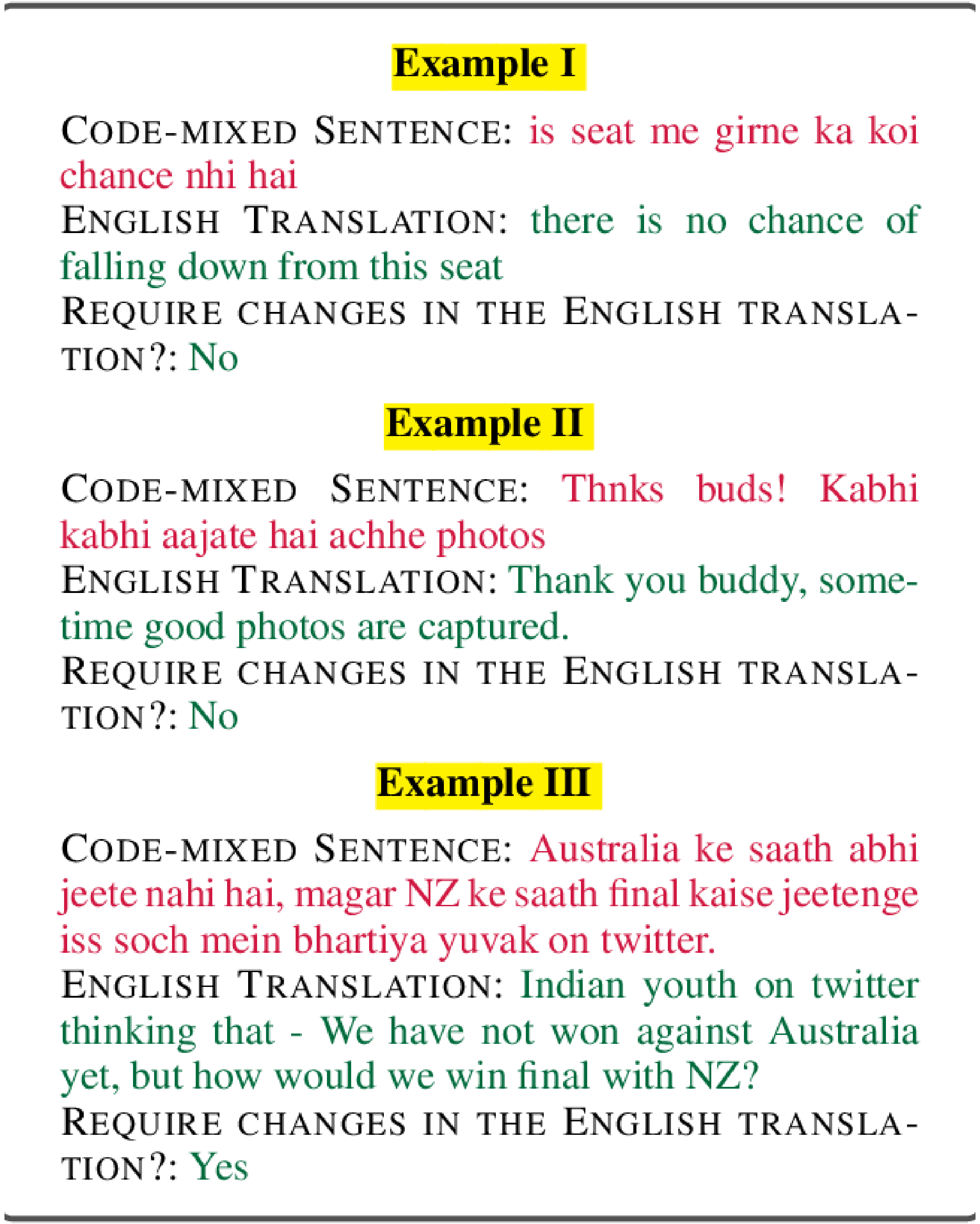
Hi

विश्वविद्यालय जहाँ स्थित है, उसके पूर्वी दिशा में बने परिसरों को क्या नाम दिया गया है?

जब 1919 में यूसीएलए ने अपना नया परिसर खोला, तब इसमें चार इमारतें थी। [...] परिसर अनौपचारिक रूप से ***उत्तरी परिसर और दक्षिणी परिसर*** में विभाजित है, जो दोनों विश्वविद्यालय की जमीन के पूर्वी हिस्से में स्थित हैं। [...] दक्षिणी परिसर में भौतिक विज्ञानं, जीव विज्ञान, इंजीनियरिंग, मनोविज्ञान, गणितीय विज्ञान, सभी स्वास्थ्य से संबंधित क्षेत्र और यूएलसीए मेडिकल सेंटर स्थित है।

(b)

Code-mixed data



Example I

Code-mixed Sentence: is seat me girne ka koi chance nhi hai

English Translation: there is no chance of falling down from this seat

Require changes in the English translation?: No

Example II

Code-mixed Sentence: Thnks buds! Kabhi kabhi aajate hai achhe photos

English Translation: Thank you buddy, sometime good photos are captured.

Require changes in the English translation?: No

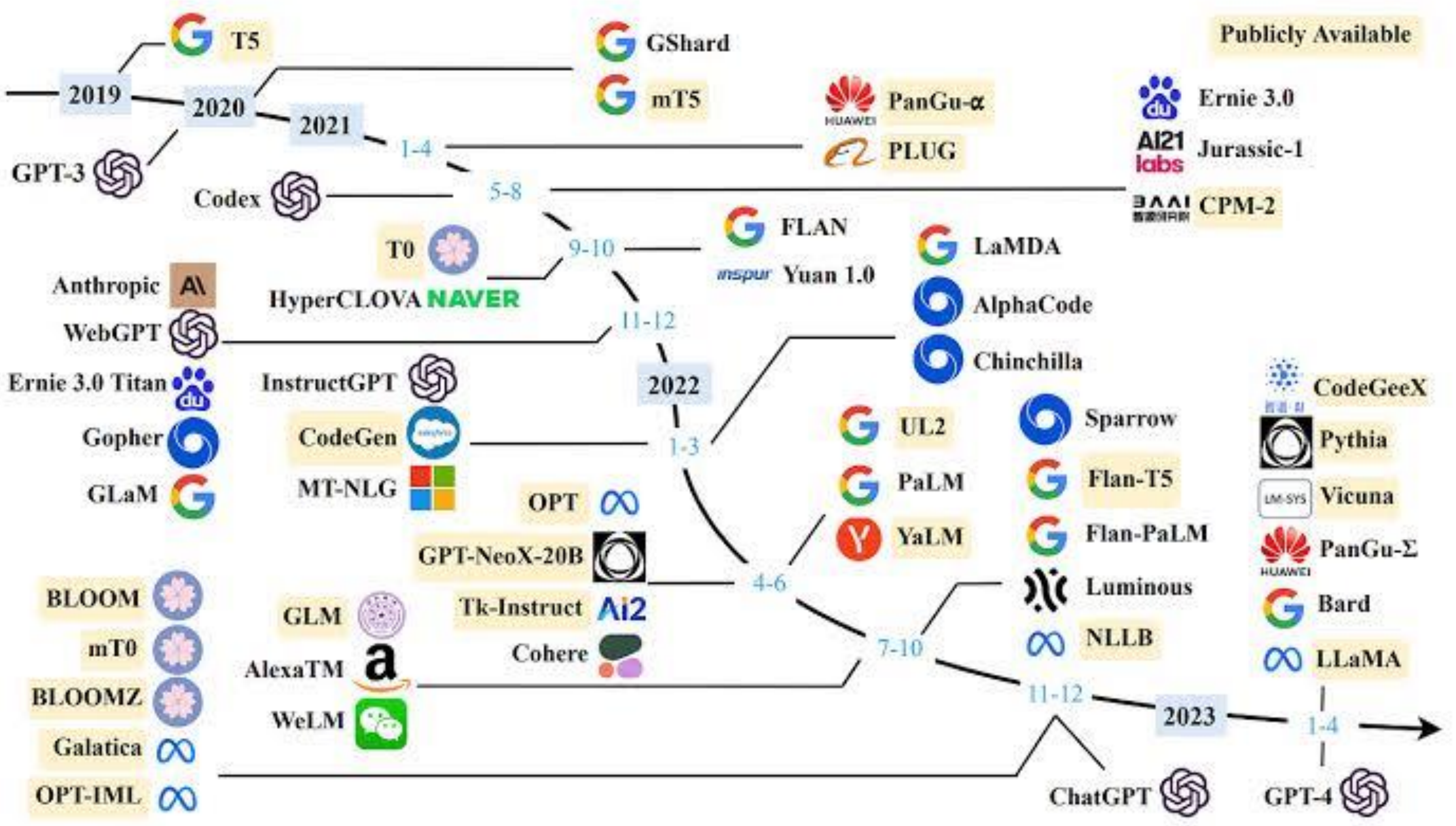
Example III

Code-mixed Sentence: Australia ke saath abhi jeete nahi hai, magar NZ ke saath final kaise jeetenge iss soch mein bhartiya yuvak on twitter.

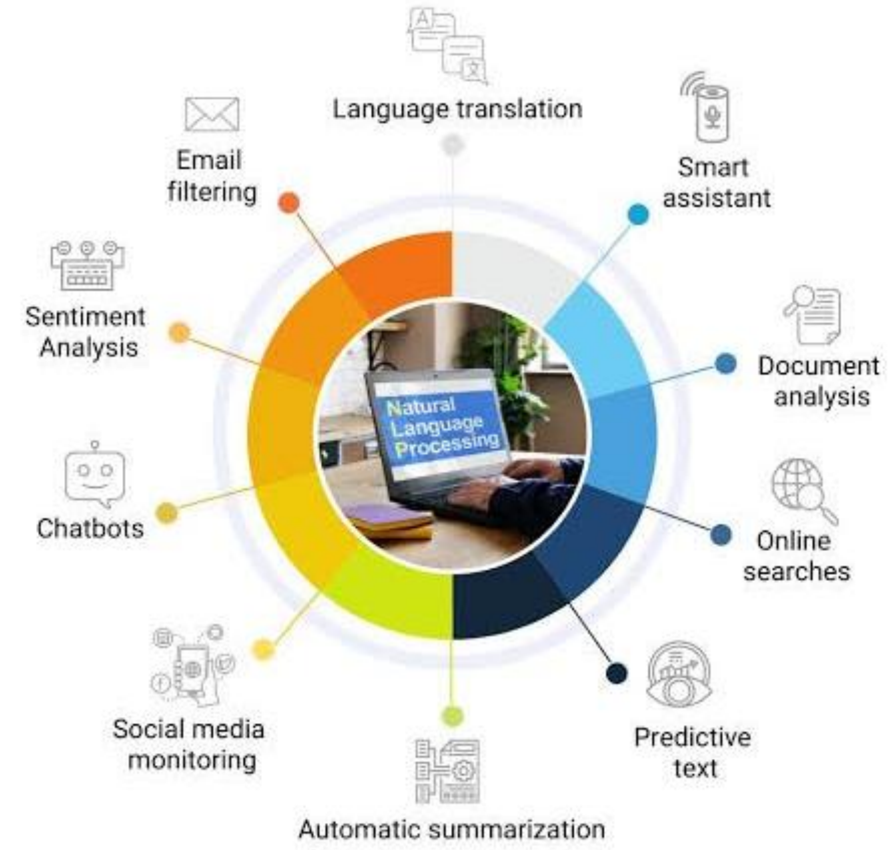
English Translation: Indian youth on twitter thinking that - We have not won against Australia yet, but how would we win final with NZ?

Require changes in the English translation?: Yes

NLP research and advanced topics

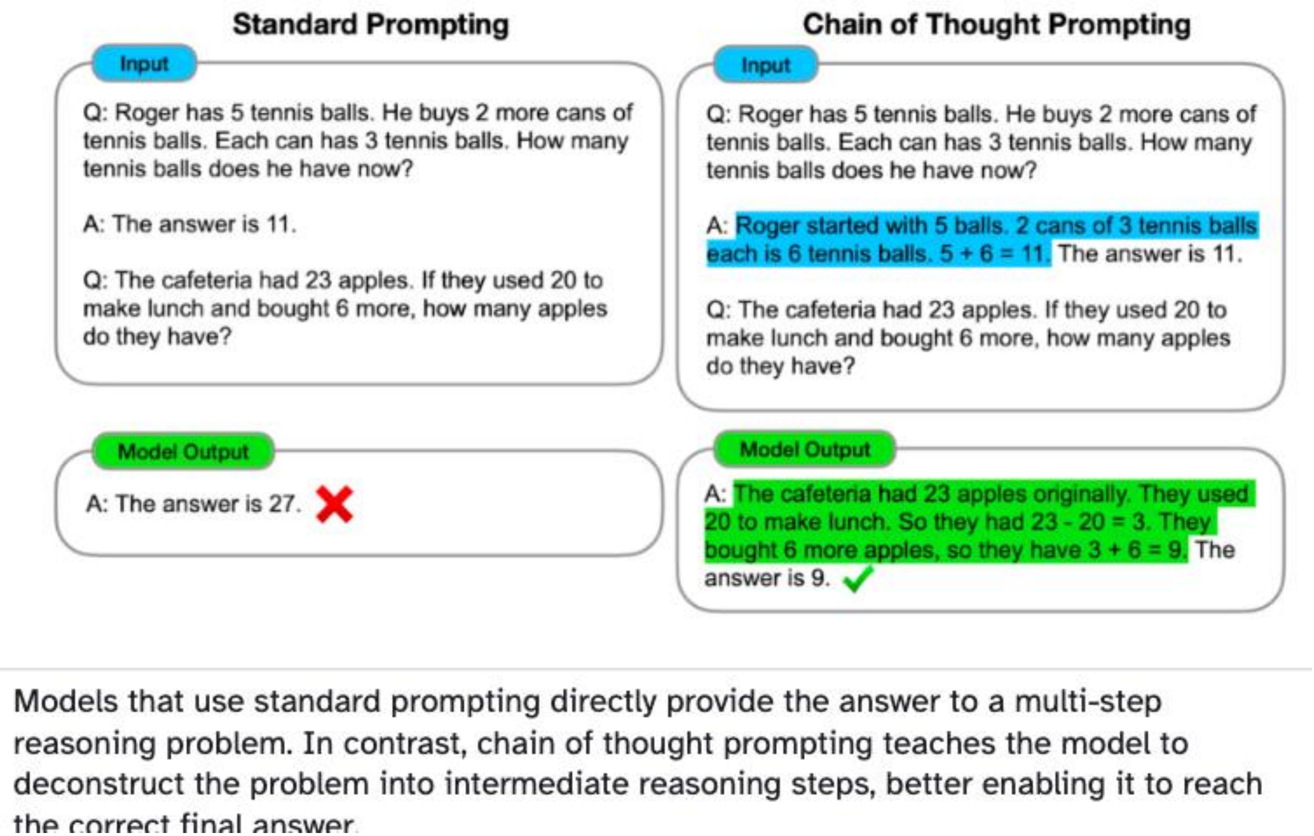


NLP Applications



https://www.youtube.com/watch?v=lj0bFX9HXeE&t=3s

Applications!



Standard Prompting

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

A: The answer is 27. ✗

Chain of Thought Prompting

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. 5 + 6 = 11. The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had 23 - 20 = 3. They bought 6 more apples, so they have 3 + 6 = 9. The answer is 9. ✓

Models that use standard prompting directly provide the answer to a multi-step reasoning problem. In contrast, chain of thought prompting teaches the model to deconstruct the problem into intermediate reasoning steps, better enabling it to reach the correct final answer.

https://blog.research.google/2023/01/google-research-2022-beyond-language.html

Lecture Ethics

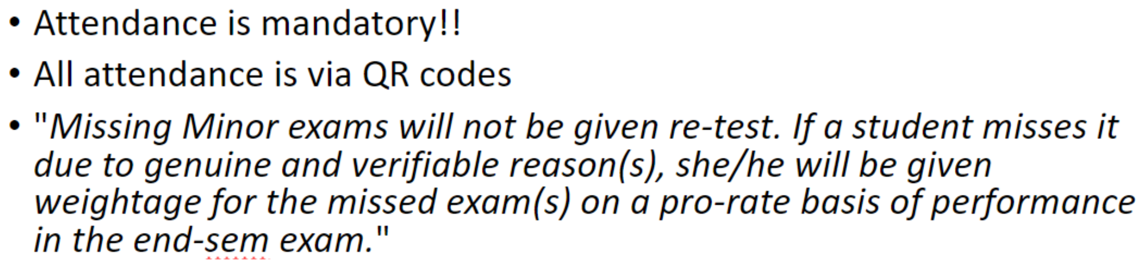
sincerity
fairness
clarity
openness
transparency
honesty
believability
truth
accuracy
directness
forthrightness

Smooth functioning of Course

- Slack
    - Create your account on slack/Communication on slack
    - Be active to check announcements!!
    - Don't spam with individual mails.
        - First communication point for doubts on slack
        - Discuss among batchmates (if doubt already resolved)
        - Instructors and TAs (Always for help!!!)

Course logistics

- Slack for assignments, doubts, course announcements, deadlines, etc.
- Follow Course page for resources, materials, logistics, etc.
- Use email judiciously for important/urgent communication
- Don't come without a scheduled slot.
- Don't want to make you wait (in case another meeting)!!Therefore, Calendar invite is must!!
- Make use of Office hours (Don't spam through individual mails!!!)



- Attendance is mandatory!!
- All attendance is via QR codes
- *"Missing Minor exams will not be given re-test. If a student misses it due to genuine and verifiable reason(s), she/he will be given weightage for the missed exam(s) on a pro-rate basis of performance in the end-sem exam."*

Grading Policy

Type of Evaluation	% Contribution in Grade
Minor 1 and Minor 2	30
Endsem	25
In class activity/quiz / Class Attendance	10
Assignment (6)	15
Project	20

Cheating Policy

  - everything you turn in must be your own work, and
  - you must note the names of anyone you collaborated with on each problem (the only exceptions are the instructors and TAs), and the nature of the collaboration (e.g., “X helped me,” “I helped X,” “X and I worked it out together.”).
- If you find material in published literature (e.g., on the Web) that is helpful in solving a problem, you must cite it and explain the answer in your own words.
- The project is to be completed by a team; you are not permitted to discuss any aspect of your project with anyone other

References

https://research.google/research-areas/

https://research.ibm.com/topics/natural-language-processing

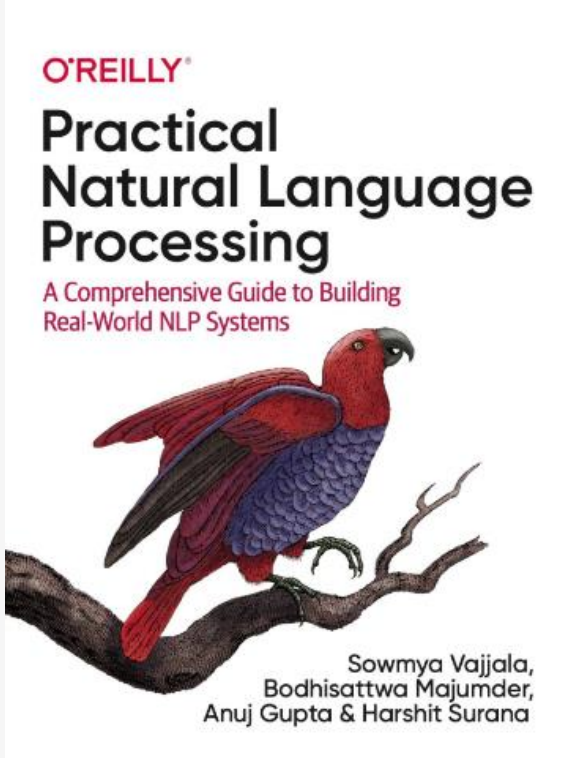
https://openai.com/research/better-language-models

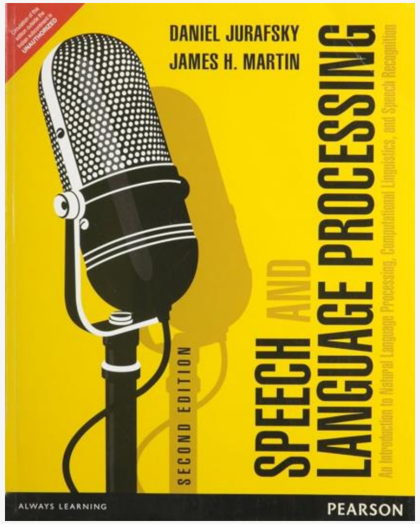
Reference materials

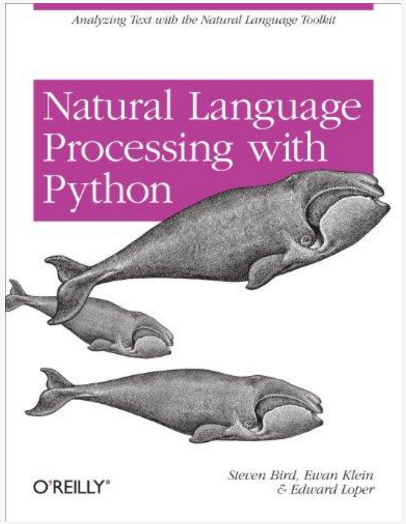
- https://vlanc-lab.github.io/mu-nlp-course/

- Lecture notes

- (A) Speech and Language Processing by Daniel Jurafsky and James H. Martin
- (B) Natural Language Processing with Python. (updated edition based on Python 3 and NLTK 3) Steven Bird et al. O’Reilly Media


O'REILLY
Practical Natural Language Processing
A Comprehensive Guide to Building Real-World NLP Systems
Sowmya Vajjala, Bodhisattwa Majumder, Anuj Gupta & Harshit Surana


DANIEL JURAFSKY
JAMES H. MARTIN
SECOND EDITION
SPEECH AND LANGUAGE PROCESSING
An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition
ALWAYS LEARNING
PEARSON


Analyzing Text with the Natural Language Toolkit
Natural Language Processing with Python
O'REILLY
Steven Bird, Ewan Klein & Edward Loper