Natural Language Processing

Dr. Vaibhav Garg
Collegiate Assistant Professor

vaibhavg@vt.edu
Office# 479 (NVC)

Fall 2024

Bio highlights and humble bragging

Ph.D. in Computer Science from NC State University

Passionate about solving problems having high social impact using NLP :)

Mentored both grad and undergrad students in research

Sanjana Cheerla (Ph.D. student at NCSU) ; also an influencer LOL
Ganning Xu (Undergrad at Georgia Tech)
Arun Gaonkar (SDE 2 at Lexis Nexis)

Won Carla Savage Award for the most awesome Ph.D. student

Won best thesis award during Masters in CS

Recipient of NSA and NSF research grants

Accepted papers in top venues:
ACL and Communications of the ACM

Perhaps I'm the youngest :)

Don't like wearing formal but wa
to be taken seriously!!

Bio highlights and humble bragging

I’m not a nerd and don’t expect a bunch of nerds in my class!!

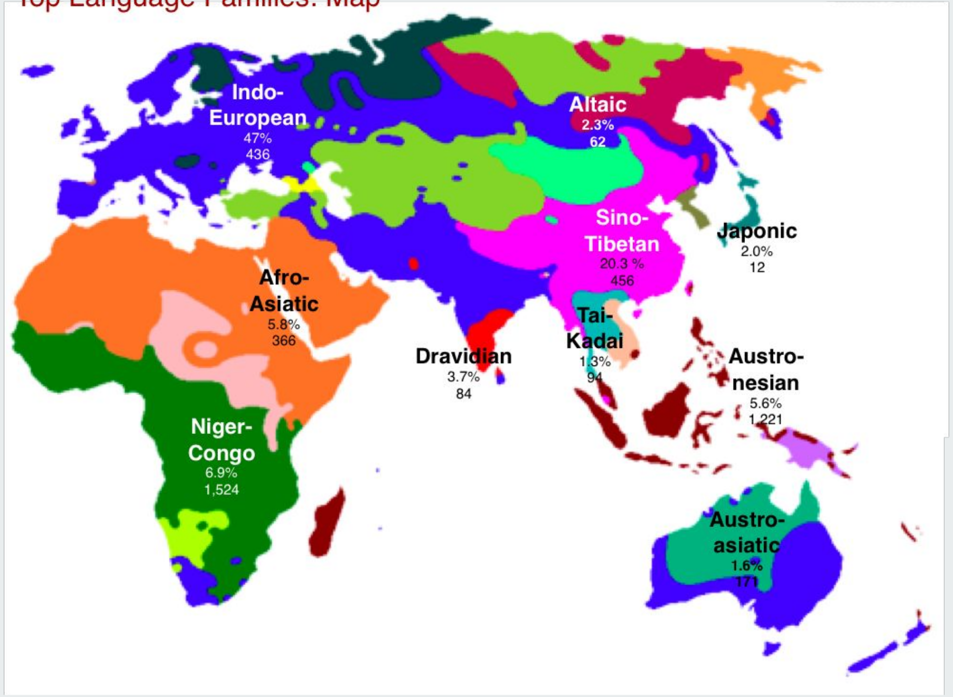
Passionate bollywood dancer
Practice meditation
Badminton player
Excited about movies, filmmaking process, and songs

Now your round of introduction??

Enjoy taking breaks!!

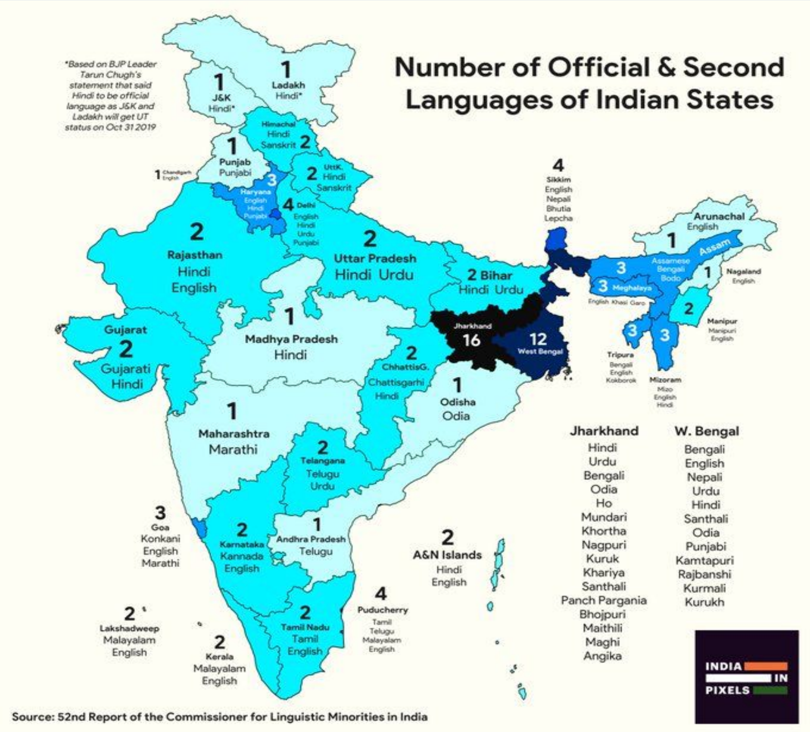
Diversity of languages

Top Language Families: Map



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
3 Haryana English Hindi Punjabi
2 UttK. Hindi Sanskrit
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
1 Nagaland English
3 Meghalaya English Khasi Garo
2 Manipur Manipuri English
3 Tripura Bengali English Kokborok
3 Mizoram Mizo English Hindi
Jharkhand 16
12 West Bengal
Gujarat 2 Gujarati Hindi
1 Madhya Pradesh Hindi
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 Tamil Nadu Tamil English
2 Lakshadweep Malayalam English
2 Kerala Malayalam 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

Any other country or continent with multiple languages??

https://x.com/indiainpixels/status/1299284220687167488

What exactly is natural language?

Some naturally occurring phenomena?

A language created by someone? But then how is it natural?

How is it different from formal logic (if A implies B and B implies C; then A implies C)?

Let’s discuss!

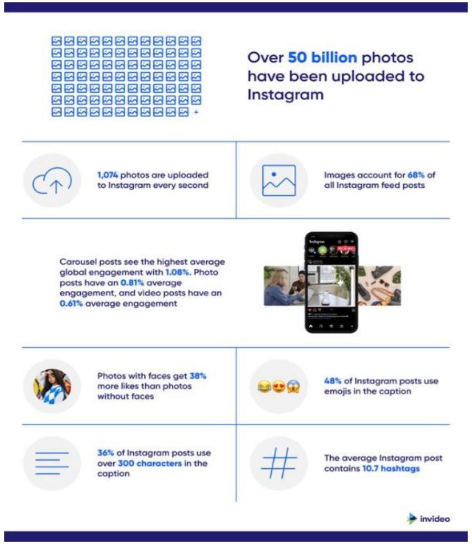
What exactly is natural language?

“A natural language is a human language, such as English or Standard Mandarin, as opposed to a constructed language, an artificial language, a machine language, or the language of formal logic. Also called ordinary language.

The theory of universal grammar proposes that all-natural languages have certain underlying rules that shape and limit the structure of the specific grammar for any given language.”

https://www.thoughtco.com/what-is-a-natural-language-1691422

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

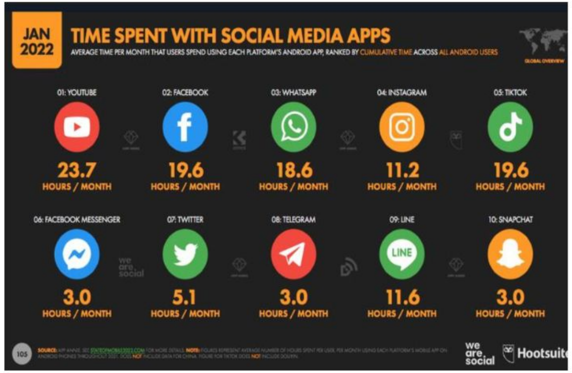
Photos with faces get **38%** more likes than photos without faces

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

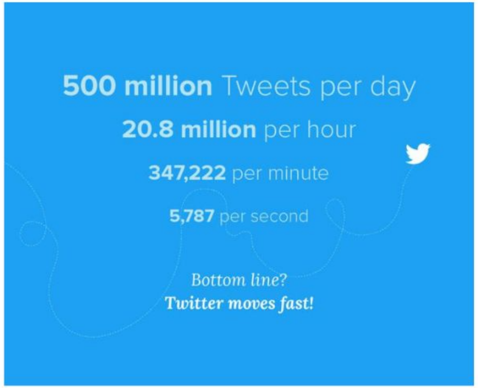
AVERAGE TIME PER MONTH THAT USERS SPEND USING EACH PLATFORM'S ANDROID APP, RANKED BY CUMULATIVE TIME ACROSS ALL ANDROID USERS

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

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.

GLOBAL OVERVIEW
105
we are social
Hootsuit



500 million Tweets per day

20.8 million per hour

347,222 per minute

5,787 per second

Bottom line?
Twitter moves fast!

Evolving language–Phrases commonly used! GenZ slangs used over social media!

NLP vs NLU vs NLG

“Alice is swimming against the **current**” (current is noun)
“The **current** version of the paper is in the folder” (current is adjective)

Is the above processing? or understanding? Or generation?

NLP vs NLU vs NLG

“Alice is swimming against the **current**”
“The **current** version of the paper is in the folder”

Is the above processing? or understanding? Or generation?

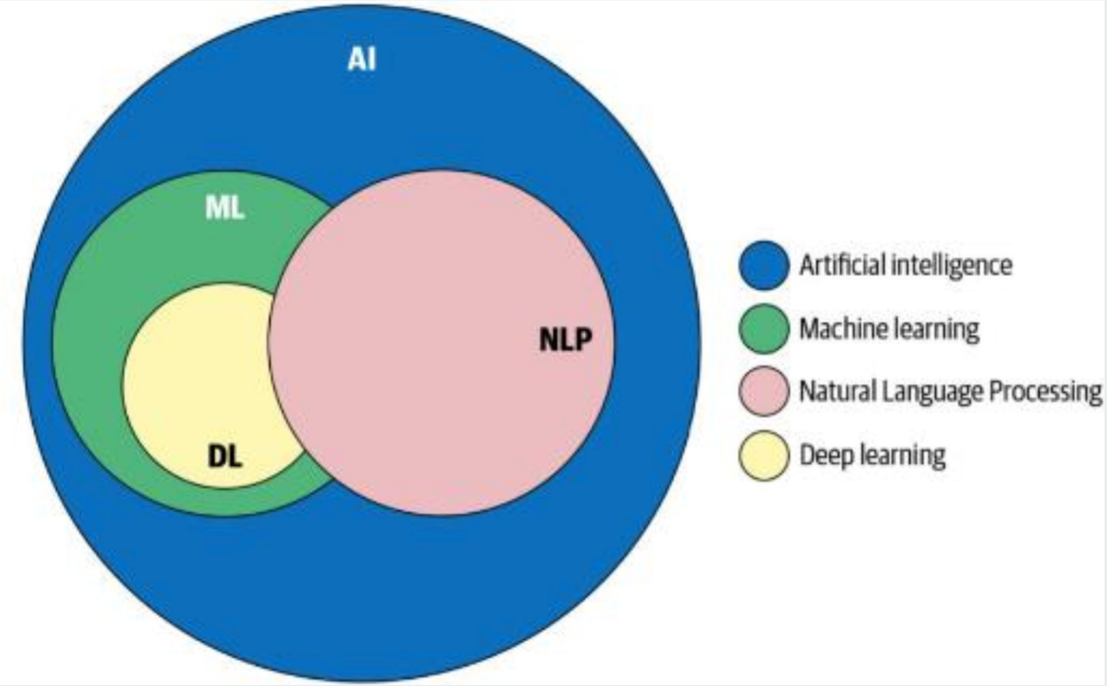
Above is **understanding** based on grammar and semantics involved

NLG involves generating a new text (like QnA and chat bots)

NLP includes both – transforming freely flowing text into structured format which is easy to interpret

https://www.ibm.com/think/topics/nlp-vs-nlu-vs-nlg

NLP



People think only deep learning and LLMs constitute NLP (**Completely wrong!!!!**)

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

Course syllabus

Text normalization and preprocessing
N-gram language models
Naive Bayes, text classification, and sentiment analysis
Vector semantics and embeddings
Parts of speech and named entity recognition
Intro to transformers and LLMs

Fundamentals

Constituency and dependency parsing
Coreference
Semantic role labeling and information extraction
Lexicons for sentiment, affect, and connotation

Word senses and WordNet
Text summarization
Online argumentation

NLP Applications

Smooth functioning of course

Will be using Canvas

Be active to check announcements

Don’t spam with individual emails – first ask on Canvas

Grading policy

Coding assignment 1 will be due in Sept	15%
Coding assignment 2 will be due in Oct	15%
Research paper presentation (Sept and Oct)	15%
Surprise quizzes (3 or 4) and seminars	15%
Final project	40%

Yayyyy!! No exams :)

Allowed: Individual submissions late upto 2 days (total); Group submissions upto 1 day late (total)

Academic integrity

Not allowed to copy the code or cheat in quizzes

Plagiarism check!!

Will also check for LLMs’ redundant code

Severe reduction in grade!!!

Lecture ethics

No usage of phones, tablets, and laptops

No talking and whispering (except in class discussions)

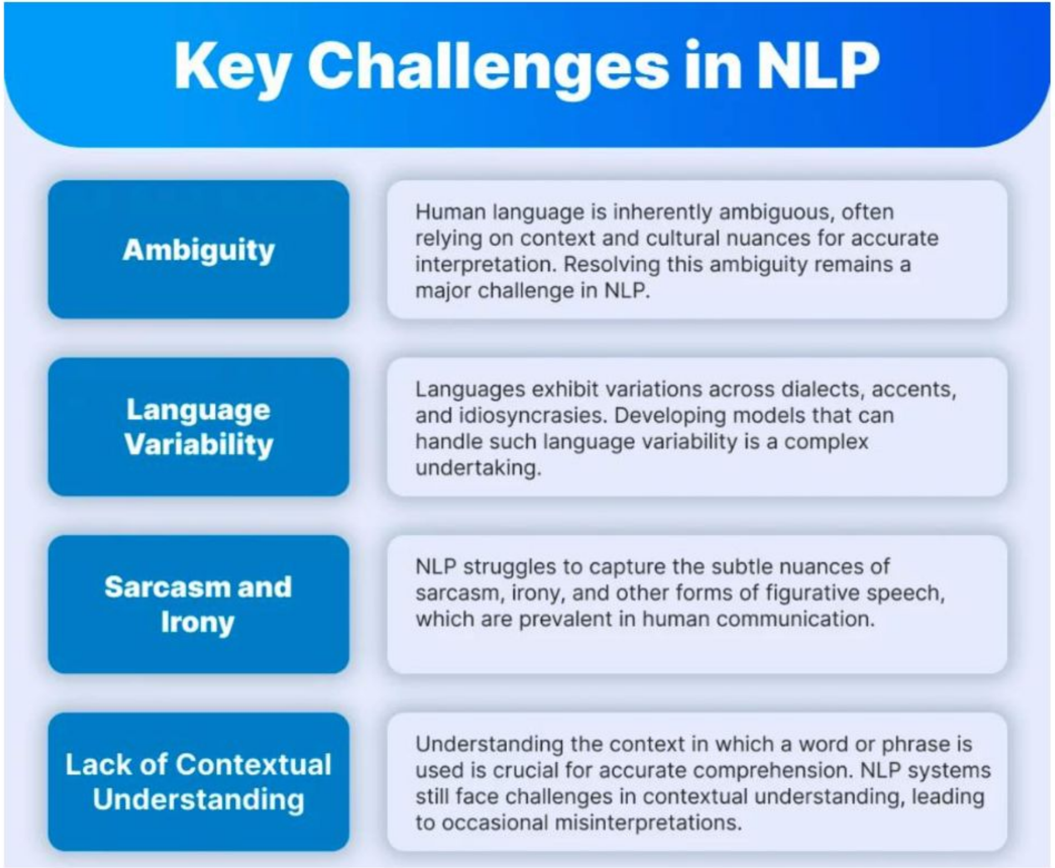
Feel free to interrupt and ask questions

Honesty in answers (discussions and quizzes)

Any questions related to the course structure and logistics?

Challenges

Examples of each?



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

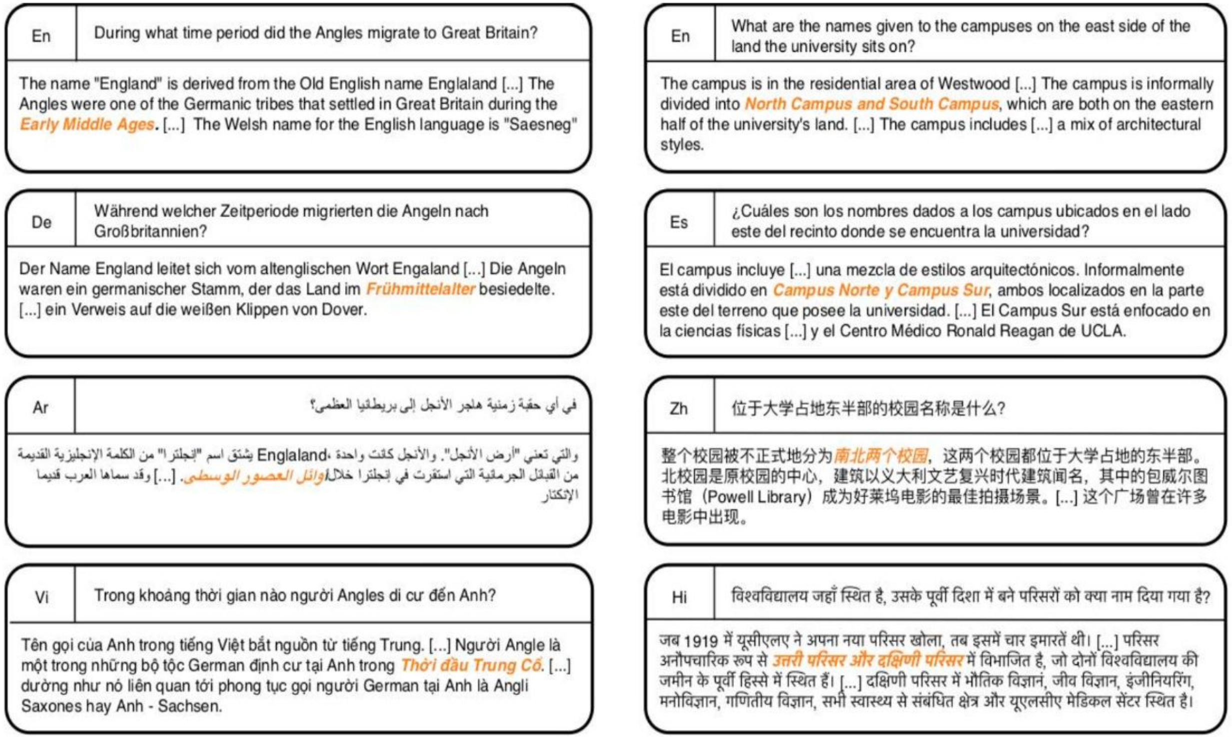
Lexical Ambiguity: This type of ambiguity represents words that can have multiple assertions. For instance, in English, the word “back” can be a noun (back stage), an adjective (back door) or an adverb (back away).

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 خلالأ***وائل 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?

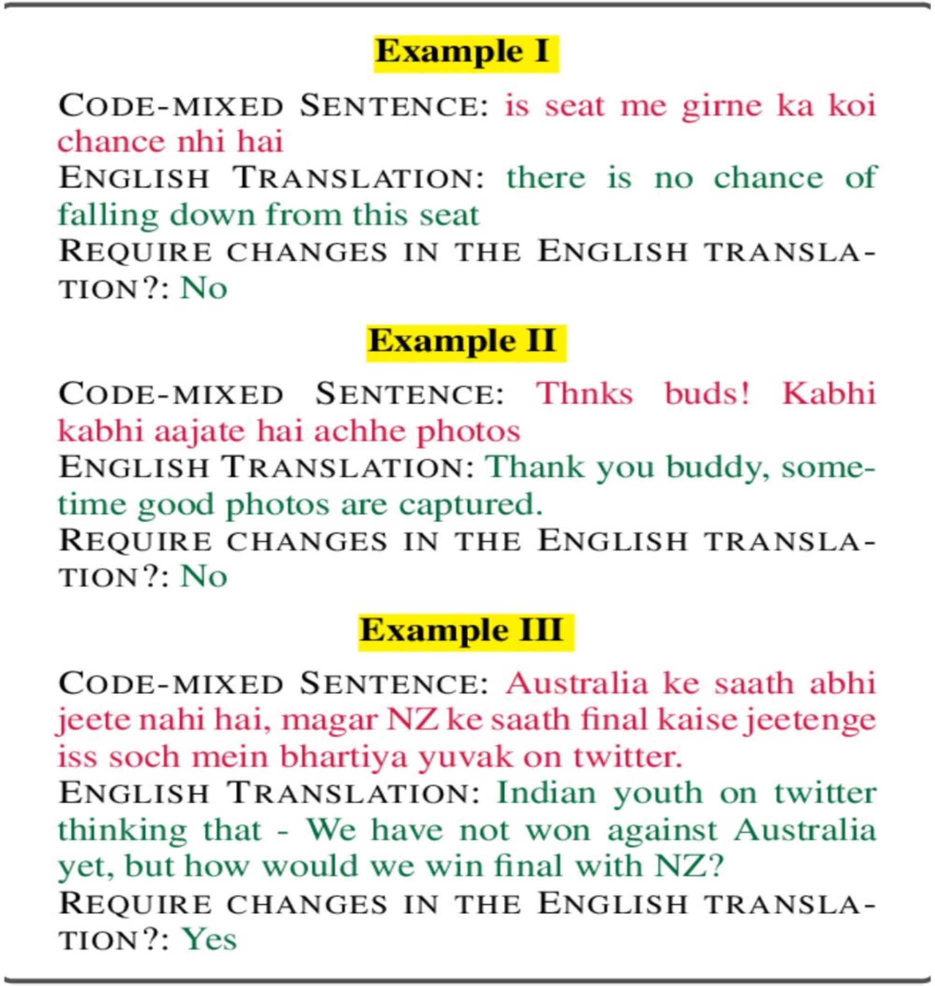
整个校园被不正式地分为***南北两个校园***，这两个校园都位于大学占地的东半部。北校园是原校园的中心，建筑以义大利文艺复兴时代建筑闻名，其中的包威尔图书馆（Powell Library）成为好莱坞电影的最佳拍摄场景。[...] 这个广场曾在许多电影中出现。

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

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

Any takeaways?

Acknowledgements

Dr. Munindar Singh at North Carolina State University
Dr. Nidhi Goel at Mahindra University
Dr. Xuan Wang at Virginia Tech

“Speech and Language Processing” by Daniel Jurafsy and James H. Martin