Code Mixing computationally bahut challenging hai

Computational Approaches to Code Mixing of Indian Languages on Online Social Networks.

Prashant Kodali

LTRC
LANGUAGE TECHNOLOGIES RESEARCH CENTER
IIIT HYDERABAD

IIIT Hyderabad

Precog
IIIT HYDERABAD

Comprehensive Viva Panel

Prof. Dipti M Sharma

Dr. Radhika Mamidi

PhD Advisors

Dr. Manish Shrivastava

Dr. Ponnurangam Kumaraguru (PK)

Agenda

01 What is Code Mixing?

02 Typological Frameworks for Code Mixing

03 Metrics of Code Mixing

04 Challenges in Processing Code Mixing

05 Data, Resources, Tasks, Computational Approaches

06 Gaps Identified

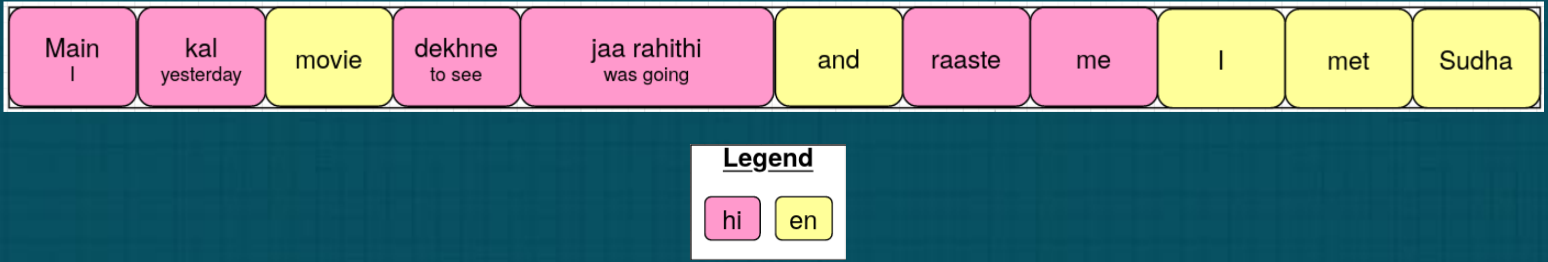
What is Code mixing

Code-Switching is **“juxtaposition within the same speech exchange of passages of speech belonging to two different grammatical systems or subsystems”**

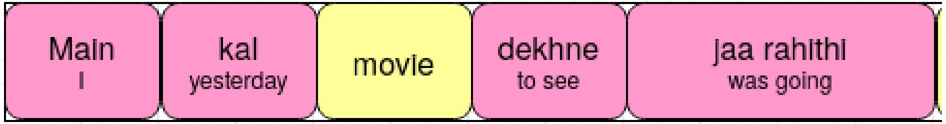
Gumprez, 1982

“juxtaposition within the same speech exchange of passages of speech belonging to two different grammatical systems or subsystems”

Gumprez, 1982



Word Borrowing or Code Mixing?



A continuum in the manner in which a lexical item transfers from one to another of two languages in contact.

Code Mixing is not just about filling lexical gaps.

Variety in "juxtaposition of two systems"

Main I	kal yesterday	movie	dekhne to see	jaa rahithi was going	and	raaste	me	I	met	Sudha

Intra Sentence

Hum we	khelne to play	jaa rahein hai. are going	Why	don't	you	join	us?

Inter Sentence

Legend

hi en

Variety in "juxtaposition of two systems"

Main I	kal yesterday	movie	dekhne to see	jaa rahithi was going	and	raaste	me	I	met	Sudha

Intra Sentence

Hum we	khelne to play	jaa rahein hai. are going	Why	don't	you	join	us?

Inter Sentence

It's	ok	guys.	India	gelchina win or	odipoina loss	mana our	EMI lu EMIs	maname we	katukovali have to pay

nee you	naale tomorrow	evng evening	shoppingnu to shop	varunnundo? coming?

+ Word Level

Legend

hi	en	ml	te	MIX	NE

A Continum ......

Word Borrowing .... Intra Sentence .... Inter Sentence .... Word Level

In this work we use **Code Mixing, Code Switching** interchangeably to denote these phenomena.

Code Mixing ⟺ **Code Switching**

Why should NLP pipelines handle code mixing

Why should NLP pipelines handle code mixing

Scale Rijhwani et al 2017

- Estimated that 3.5% of tweets are code-mixed
- More common in non-English speaking cities like Istanbul (12%)
- European vs Indian Context?

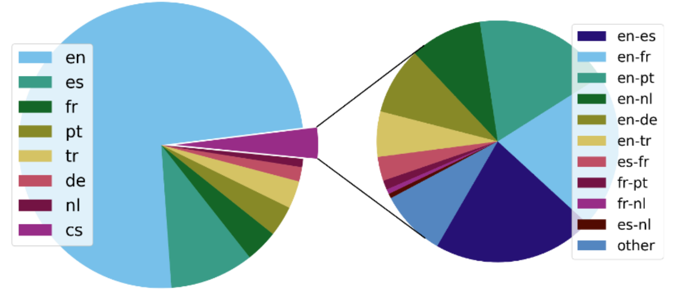


Figure 5: Worldwide distribution of monolingual and CS tweets (left and right charts respectively)

Why should NLP pipelines handle code mixing

Scale

Social , Psychological & Conversational Factors

- Switch language to express ___________ .

I may talk in English but *gaali toh Hindi mein hi denge* : A study of English-Hindi Code-Switching and Swearing Pattern on Social Networks

Prabhat Agarwal*, Ashish Sharma*, Jeenu Grover*, Mayank Sikka*, Koustav Rudra*, Monojit Choudhury†
**Department of Computer Science and Engineering*
Indian Institute of Technology Kharagpur
Kharagpur, WB 721302, India
{prabhat.agr2010,ashishshrma22,groverjeenu,mayanksikka95, krudra5} @gmail.com
†Microsoft Research India,
Bangalore, Karnataka 560027, India
monojitc@microsoft.com

Why should NLP pipelines handle code mixing

Scale

Social, Psychological & Conversational Factors

- Switch language to express __________ .

- Used in interpersonal, informal settings and Interactions. Online Forums, chats where code mixing manifests frequently.

Snapshot from a predominantly Telugu speakers subreddit

Why should NLP pipelines handle code mixing

Scale

Social, Psychological & Conversational Factors

Utility in Human Computer Interactions

- Search Engines, Translators
- Chatbots
- Educational Resources

Do Multilingual Users Prefer Chat-bots that Code-mix? *Let's Nudge and Find Out!*

HCI 2020

ANSHUL BAWA, Microsoft Research, India
PRANAV KHADPE, Microsoft Research, India
PRATIK JOSHI, Microsoft Research, India
KALIKA BALI, Microsoft Research, India
MONOJIT CHOUDHURY, Microsoft Research, India

Despite their pervasiveness, current text-based conversational agents (chatbots) are predominantly monolingual, while users are often multilingual. It is well-known that multilingual users mix languages while interacting with others, as well as in their interactions with computer systems (such as query formulation in text-/voice-based search interfaces and digital assistants). Linguists refer to this phenomenon as *code-mixing* or *code-switching*. Do multilingual users also prefer chatbots that can respond in a code-mixed language over those which cannot? In order to inform the design of chatbots for multilingual users, we conduct a mixed-method user-study ($N = 91$) where we examine how conversational agents, that code-mix and reciprocate the users' mixing choices over multiple conversation turns, are evaluated and perceived by bilingual users. We design a human-in-the-loop chatbot with two different code-mixing policies – (a) *always code-mix* irrespective of user behavior, and (b) *nudge* with subtle code-mixed cues and reciprocate only if the user, in turn, code-mixes. These two are contrasted with a monolingual chatbot that never code-mixed. Users are asked to interact with the bots, and provide ratings on perceived naturalness and personal preference. They are also asked open-ended questions around what they (dis)liked about the bots. Analysis of the chat logs, users' ratings, and qualitative responses reveal that multilingual users strongly prefer chatbots that can code-mix. We find that self-reported language proficiency is the strongest predictor of user preferences. Compared to the *Always code-mix* policy, *Nudging* emerges as a low-risk low-gain policy which is equally acceptable to all users. Nudging as a policy is further supported by the observation that users who rate the code-mixing bot higher typically tend to reciprocate the language mixing pattern of the bot. These findings present a first step towards developing conversational systems that are more human-like and engaging by virtue of adapting to the users' linguistic style.

Typological Frameworks Of Code Mixing

Can I arbitrarily mix tokens from different languages to generate code mix utterances?

Appears to be distinction between **an acceptable mix** vs **an unacceptable mixing.**

Ex. 1. I do research in code mixing

Ex. 2. main **code mixing** mein **research** karta hoon.

Ex. 3. I do **shodh karya** on code mixing.

Ex. 4. * main do code mixing pe shodh karya.

Can I arbitrarily mix tokens from different languages to generate code mix utterances?

Appears to be distinction between **an acceptable mix** vs **an unacceptable mixing.**

Ex. 1. I do research in code mixing

Ex. 2. main **code mixing** mein **research** karta hoon.

Ex. 3. I do **shodh karya** on code mixing.

Ex. 4. * main do code mixing pe shodh karya.

"…. a **cline** of acceptability.....".

- Neither an open ended process – lexically or grammatically
- Not necessarily a "yes" or "no" judgement.

Tow ards Structuring Code Mixing : An Indian Perspective ,Kachru, 1985

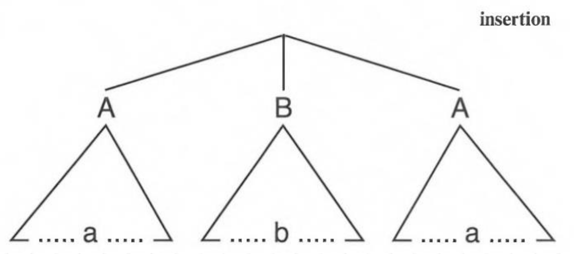
Are there rules to distinguish between "natural" and "unnatural" code mix utterances?

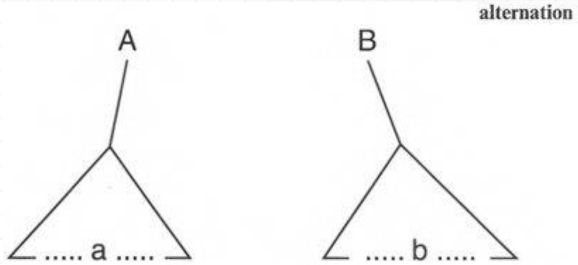
Constraint Based Theories :

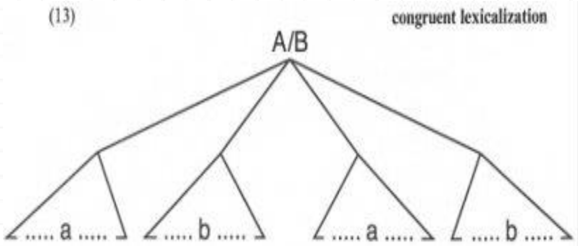
Two or more languages are interacting.

What are the **constraints** on these interactions to generate "natural" code mix sentences?

Categorization of Constraint Based Theories







insertion
of material from a language
into
a structure from the other language.

Ex: "main **window shopping** ke liye jaa raha hoon"

alternation
between structures from languages

Ex: Usne bola ki **one in hand is better than two in a bush.**

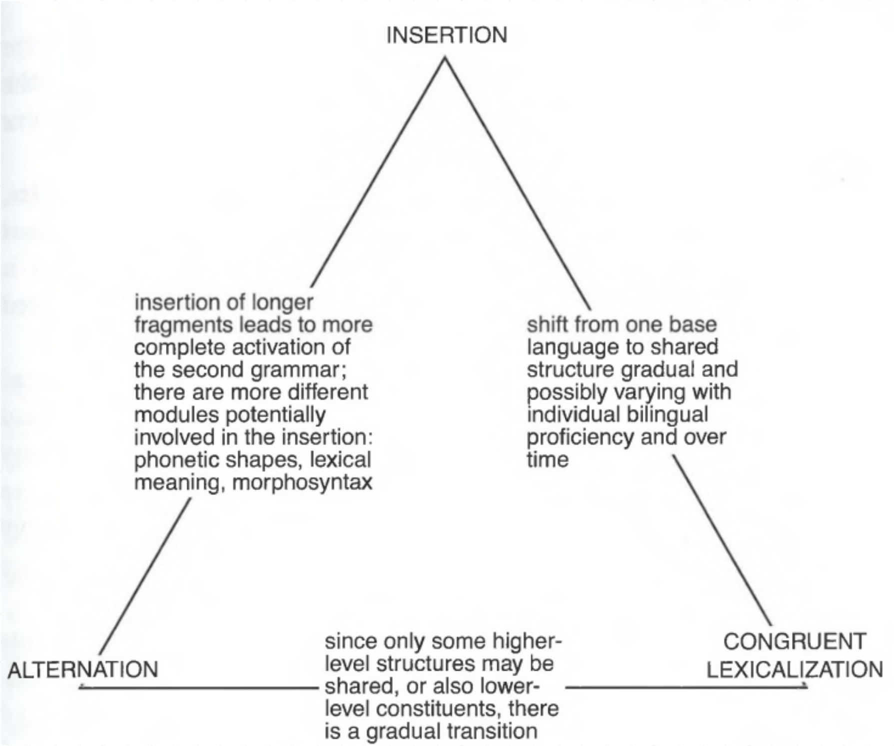
congruent lexicalization
of material from different lexical inventories into a shared grammatical structure.

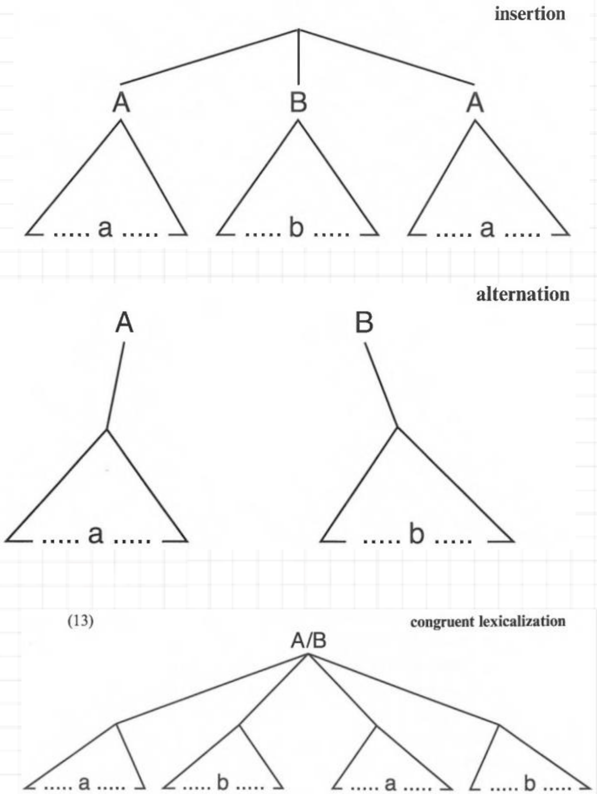
Ex. I want to **neladeesify** them.

Gloss : Neeladiyatam == Confront

Ref : Bilingual Speech A Typology of Code-Mixing, Pieter Muysken, 2001

Interaction between these categories





Ref : Bilingual Speech A Typology of Code-Mixing, Pieter Muysken, 2001

Typological Frameworks – In Conclusion

- Code Mixing isn't a open-ended system. Distinction between natural and un-natural code mixing

- Abstraction of Insertion – Alternation – Congruent Lexicalisation for covering the gamut of code-mixing.

- Implication for computational tools –
    - Models should be multilingual.
    - Utility of grammatical constraints to generarte synthetic code mix sentences

Metrics of Code Mixing

Metrics of code mixing

To capture

- Degree of Code Mixing
- Nature of Code Mixing

01

Ratio Based

Ratio of number of tokens belonging to different languages

02

Time – Course Measures

Temporal Distribution of switch points

03

Memory Based

Time Series view of switch spans

Metrics of code mixing

To capture

- Degree of Code Mixing
- Nature of Code Mixing

Limitations

- Only Language ID tags considered
- Do not capture
  - **"naturalness"**
  - **syntactic variation**

01 Ratio Based

Ratio of number of tokens belonging to different languages

02 Time – Course Measures

Temporal Distribution of switch points

03 Memory Based

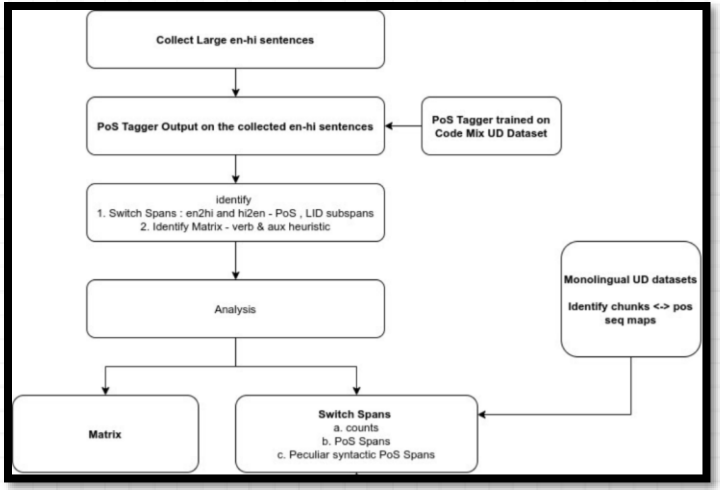
Time series view of switch spans

Syntactic Variety in Switching.

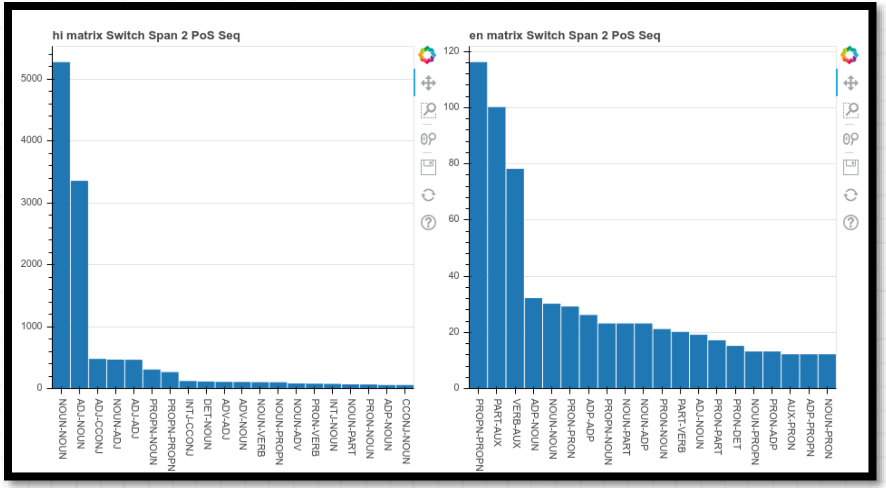
- In a corpus, which syntactical units (PoS, Chunks) are switched?
- Do they impact efficacy of a computational pipeline?
- Is example 2 more acceptable/natural than example 1?

Do we have Quantitative Measure to encode this notion?

Our on-going work to compute Syntactic Measure of Code Mixing



Pipeline for Syntactic Analysis of Code Mixing



Analysing the natrue of sw itched spans – a syntactic perspective

C_{ON} : Open Class – Noun family – [ADJ, NOUN]
C_{OV} : Open class – Verb family – [ADV, VERB]
C_{OO} : Open class – others – [INTJ]
C_{Closed} : Closed class – [ADP, AUX, DET, NUM, PART, PRON, SCONJ, CCONJ]
C_{oth} : Others – [PUNCT, SYM, X]

Syntactic measure of Code Mixing
- Ratio of the switched syntactic categories of tokens belonging to L1 and L2.

Data
Resources
Tasks

Data, Resources & Tasks

Name	Language Mix	Source of dataset	Purpose of Dataset
LINCE Benchmark [36]	hi-en, es-en, ne-en, MSA-Egyptian Arabic	Tweets, Facebook, Conversational	LID POS NER MT
GLUECoS Benchmark [37]	en-es, en-hi	Tweets, Facebook, Translated monolingual datasets	LID POS POS NER Sentiment Analysis NLI QA
Sentiment Analysis [38]	en-hi	Tweets	Sentiment Analysis
Semeval-2020 Sentiment Analysis [39]		Tweets	Sentiment Analysis
Machine Translation [40]	en-hi	Social Media	MT
Aggression Detection Shared Task [41]	en-hi	Facebook, Twitter	Aggression Detection
Hate Speech Detection [42]	en-hi	Tweets	Hate speech detection
Stance Detection [43]	en-hi	Tweets	Stance Detection
Stance Detection [44]	en-hi	Tweets	
Stance Detection [45]	en-ka	Facebook	
Sarcasm Detection [46]	en-hi	Tweets	Sarcasm Detection
Humor Detection [47]	en-hi	Tweets	Humor Detection
Code Mixed Goal Oriented Conversation Systems [48]	en-hi en-gu en-ta en-be	Translated Monolingual Dataset	Conversational Datasets
Sentiment Analysis [49]	en-te	Tweets	Sentiment Analysis
ICON 2015-2016 Contest [50]	en-hi en-be en-te	Tweets, Facebook	POS, LID
Sentiment Analysis [51]	hi-en bn-en	Tweets	Sentiment Analysis
FIRE 2013-16 Tasks [52]	en,hi,ba,gu,ml,ta,te	Tweets, Facebook, Gutenberg Project	Transliterated Search, Code Mix Cross Script QA, IR on Code mix hi-en tweets
Information Retrieval [53]	en-hi	Tweets	IR
FIRE 2020 Dravidian Code Mixed [54]	en-ta	YouTube Comments	Sentiment Analysis
Offenseval Dravidian [55]	en-ta en-ma en-ka	YouTube Comments	Offensive Language Detection

To understand

- Tasks that have been attempted for code mix
- Language Pairs addressed
- Scale of available data

We collate

- Publicly available code mix datasets for
    - Indian Language Pairs,
    - different tasks

Language Pairs

Language Pair	Number of sentences
en-hi	89,338
en-ta	45,472
en-be	14,625
en-gu	12,094
en-ml	9,291
en-ka	4,675
en-te	1,617

- en-hi has most number of datasets, for various tasks
- Recent uptick in en-Dravidian Language Pairs
- Disparity in the language pairs.
- All are sourced from Facebook, Twitter, YouTube comments, Movies scripts.
- For en-hi language pair

~ 16% of sentences from these datasets are monolingual

Our Work for Code Mix Data Collection

- **Objectives**
  - Collect corpus for Indian code mix language paris.
  - Characterize the collected corpus – LID based measures and syntactically.
  - A toolkit to replicate the data collection exercise along with prescription of data collection strategies that work well in our experiments.

- **Methodology**
  - A sentence level
    - Binary Classifier – Code mix or not?
    - If code mix – What is the language mix?
  - Mine frequently occurring code-mix spans which could become query terms for Online Social Network APIs.
  - Training data
    - Collect publicly available datasets for different language pairs.
    - Synthetic data for language pairs with very less data.
  - For curating a test set – apply combination of heuristics + existing LID tools to create such

Tasks

- Sentiment and Stance Detection have highest number of datasets
- Recently, Hate and Offensive Speech Detection have attracted researchers attention.
- Code mix generation and translation has also attracted attention in last couple of years.

Task	Number of datasets
LID	2
PoS	2
NER	2
Shallow Parsing, Dependency Parsing	2
Sentiment	5
Stance	3
Sarcasm	2

Task	Number of datasets
Humour	1
Hate	1
Offensive	1
Aggression Detection	1
Information Retreival	1
MT , Dialouge Generation	4

Benchmarks

Litmus Test for Code Mixing Processing?

English-Hindi				
Corpus	**Sent (Train)**	**Sent (Dev)**	**Sent (Test)**	**Sent (All)**
Fire LID (D)	2631	500	406	3537
UD POS (D)	1384	215	215	1814
FG POS (R)	2104	263	264	2631
IIITH NER (R)	2467	308	309	3084
SAIL Sentiment (R)	10080	1260	1261	12601
QA (R)	250	-	63	313
NLI (R)	1040	130	130	1300
English-Spanish				
Corpus	**Sent (Train)**	**Sent (Dev)**	**Sent (Test)**	**Sent (All)**
EMNLP 2014	10259	1140	3014	14413
Bangor POS	2192	274	274	2758
CALCS NER	27366	3420	3421	34208
Sentiment	1681	211	211	2103

Tasks	Corpus Authors	Languages	Training			Development			Test		
			CMI	Posts	Tokens	CMI	Posts	Tokens	CMI	Posts	Tokens
LID	Molina et al. (2016)	SPA-ENG	8.491	21,030	253,221	7.062	3,332	40,391	8.264	8,289	97,341
	Solorio et al. (2014)	NEP-ENG	20.322	8,451	122,952	17.079	1,332	19,273	19.754	3,228	46,559
	Mave et al. (2018)	HIN-ENG	10.222	4,823	95,224	10.122	744	15,446	9.930	1,854	36,052
	Molina et al. (2016)	MSA-EA	2.567	8,464	171,872	3.185	1,116	21,978	3.849	1,663	33,504
POS	Singh et al. (2018b)	HIN-ENG	21.449	1,030	22,993	15.293	160	3,476	18.910	299	6,541
	Soto and Hirschberg (2017)	SPA-ENG	24.191	27,893	217,068	24.040	4,298	33,345	24.282	10,720	82,656
NER	Aguilar et al. (2018)	SPA-ENG	5.567	33,611	404,428	4.398	10,085	122,656	5.867	23,527	281,579
	Singh et al. (2018a)	HIN-ENG	20.117	1,243	21,065	19.913	314	5,364	19.733	522	8,945
	Aguilar et al. (2018)	MSA-EA	–	10,103	204,296	–	1,122	22,742	–	1,110	21,414
SA	Patwa et al. (2020)	SPA-ENG	20.643	12,194	186,602	21.553	1,859	28,202	20.528	4,736	72,006

GLUECoS Benchmark

ACL 2020

LINCE Benchmark

LREC 2020

Limitations

1. Limited Language pairs
2. Small dataset size
3. Limited tasks

Benchmarks

How well are the models performing?

Data	Baseline	Unsup. MUSE	Sup. MUSE	BiSkip	SOTA
FIRE En-Hi	93.21	94.53	94.92	93.98	N/A[9]
	BiCVM	**GCM**	**mBERT**	**Mod. mBERT**	
	95.24	93.64	95.87	**96.6**	
EMNLP En-Es	**Baseline**	**Unsup. MUSE**	**Sup. MUSE**	**BiSkip**	**SOTA**
	92.95	92.86	93.39	92.79	94.0
	BiCVM	**GCM**	**mBERT**	**Mod. mBERT**	
	91.47	92.42	95.97	**96.24**	

Table 3: LID results (F1)

Data	Baseline	Unsup. MUSE	Sup. MUSE	BiSkip	SOTA
SAIL En-Hi	50.44	48.37	51.27	48.84	56.9
	BiCVM	**GCM**	**mBERT**	**Mod. mBERT**	
	49.56	50.01	58.24	**59.35**	
Sentiment En-Es	**Baseline**	**Unsup. MUSE**	**Sup. MUSE**	**BiSkip**	**SOTA**
	50.62	58.73	58.44	60.4	64.6
	BiCVM	**GCM**	**mBERT**	**Mod. mBERT**	
	62.62	62.89	66.03	**69.31**	

Table 6: Sentiment Analysis results (F1)

- Models struggle to perform well on semantic tasks – Sentiment, NLI

While doing well on Syntactic tasks like LID, NER, PoS.

- Huge performance gap between similar Monolingual task and Code Mix task.

- Multilingual Transformer Based models outperform word embeddings based models.

Computational Approaches to Code mixing

- Char, Sub word level models

Universal Dependency Parsing for Hindi-English Code-switching

NAACL 2018

Irshad Ahmad Bhat
LTRC, IIIT-H,
Hyderabad, India
irshad.bhat@iiit.ac.in

Riyaz Ahmad Bhat
Interaction Labs,
Bangalore, India
rbhat@interactions.com

Manish Shrivastava
LTRC, IIIT-H,
Hyderabad, India
m.shrivastava@iiit.ac.in

Dipti Misra Sharma
LTRC, IIIT-H,
Hyderabad, India
dipti@iiit.ac.in

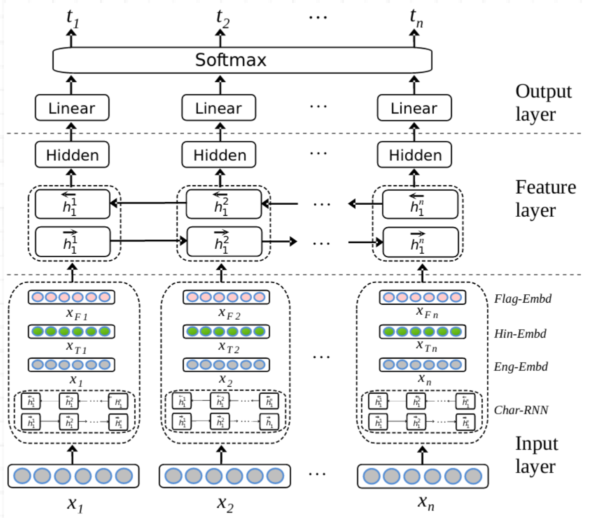


Figure 1: Language identification network

Computational Approaches to Code mixing

- Char, Sub word level models
- Transfer Learning – Zero / Few Shot
  - Monolingual Corpora / Resources
  - Multilingual Transformer based Models

Joining Hands: Exploiting Monolingual Treebanks for Parsing of Code-mixing Data

EACL 2017

Irshad Ahmad Bhat, Riyaz Ahmad Bhat, Manish Shrivastava and Dipti Misra Sharma

LTRC, IIIT-H, Hyderabad, India

{irshad.bhat,riyaz.bhat,m.shrivastava,dipti}@iiit.ac.in

How multilingual is Multilingual BERT?

ACL 2019

Telmo Pires* **Eva Schlinger** **Dan Garrette**

Google Research

{telmop,eschling,dhgarrette}@google.com

4.3 Code switching and transliteration

Code-switching (CS)—the mixing of multiple languages within a single utterance—and transliteration—writing that is not in the language's standard script—present unique test cases for M-BERT, which is pre-trained on monolingual, standard-script corpora. Generalizing to code-switching is similar to other cross-lingual transfer scenarios, but would benefit to an even larger degree from a shared multilingual representation. Likewise, generalizing to transliterated text is similar to other cross-script transfer experiments, but has the additional caveat that M-BERT was not pre-trained on text that looks like the target.

	Corrected	Transliterated
Train on monolingual HI+EN		
M-BERT	86.59	50.41
Ball and Garrette (2018)	—	77.40
Train on code-switched HI/EN		
M-BERT	90.56	85.64
Bhat et al. (2018)	—	90.53

Table 6: M-BERT's POS accuracy on the code-switched Hindi/English dataset from Bhat et al. (2018), on script-corrected and original (transliterated) tokens, and comparisons to existing work on code-switch POS.

Computational Approaches to Code mixing

- Char, Sub word level models

- Transfer Learning – Zero / Few Shot

  - Monolingual Corpora / Resources

  - Multilingual Transformer based Models

  - Cross Lingual Word Embeddings

- Synthetic Code Mix Data

Word Embeddings for Code-Mixed Language Processing

EMNLP 2018

Adithya Pratapa, Monojit Choudhury, Sunayana Sitaram
Microsoft Research, India
{t-pradi,monojitc,sunayana.sitaram}@microsoft.com

Embedding	Sentiment			POS	
	CM Overall	SemEval 2014	TASS 2016	CM Overall	at SP
None	54.4 (1.3)	64.5 (0.6)	61.4 (1.0)	84.5 (0.3)	74.0 (0.7)
BiCCA	57.6 (3.0)	64.6 (1.0)	59.5 (1.8)	**84.7** (0.8)	**75.0** (1.8)
BiCVM	**64.3** (1.3)	66.8 (1.0)	61.9 (1.0)	82.0 (0.5)	70.6 (1.7)
BiSkip	61.5 (1.7)	66.6 (0.9)	**63.9** (1.2)	84.4 (0.7)	73.8 (0.9)
χ-gCM-Skip	62.0 (1.9)	**67.4** (1.3)	63.2 (1.5)	**84.8** (0.6)	74.0 (0.6)
ρ-gCM-Skip	**64.6** (2.0)	**67.7** (1.4)	**63.8** (2.2)	**84.9** (0.7)	**75.3** (1.7)

Table 1: The performance of different pre-trained embeddings on Sentiment (F1 score) and POS tasks (Accuracy). The reported values are mean and deviation (in parentheses) values computed over multiple runs.

Challanges For Processing Code Mixing

Challenges For Code Mix Processing

- Data. Data. And more data

- Richer Representations – for any task

- Variety in Code Mixing patterns

Language Pair	Number of sentences
en-hi	89,338
en-ta	45,472
en-be	14,625
en-gu	12,094
en-ml	9,291
en-ka	4,675
en-te	1,617

Challenges For Code Mix Processing

- Data. Data. And more data
- Pre-processing – Specific to Code Mix Pipelines.
    - LID – a tool that doesn't expect set of possible Languages apriori.

```
from litcm import LIT
lit = LIT(labels=['hin', 'eng'], transliteration=True)
```

Ref : LITCM_LID_Tool

    - Transliteration – Romanized text to native script and vice versa
    - Spelling Normalization
    - Syntactic Analysis

i	i	en
thght	thought	en
mosam	मौसम	hi
dfrnt	different	en
hoga	होगा	hi
bs	बस	hi
fog	fog	en
h	है	hi

Ref : CSNLITool

Challenges For Code Mix Processing

- Data. Data. And more data
- Pre-processing
  - LID – a tool that doesnt expect set of possible Languages apriori.
  - Transliteration – romanised text to native script and vice versa
  - Spelling Normalisation
  - Syntactic Analysis

- Attention to Diverse Language Pairs – tools that aren't language pair specific
  - en-hi , en-be sab theek hai.
  - But **en-te**, **en-ka**, **hi – te**, **en-hi-be** jaise language pairs ka kya? ?

Challenges For Code Mix Processing

An end-to-end Pipeline that addresses and incorporates these issues.

- Data. Data. And more data
- Pre-processing
  - LID – a tool that doesnt expect set of possible Languages apriori.
  - Transliteration – romanised text to native script and vice versa
  - Spelling Normalisation
  - Syntactic Analysis
- Attention to Diverse Language Pairs – tools that aren't language pair specific

An example of such a pipeline

Code Mix Machine Translation

- Data – Utility in large scale pre-training

- Pre-processing
  - LID – To assess the nature of code mix generated by the model.
  - Transliteration – Converting Hindi words into Devanagari script .
  - Spelling Normalization – Evaluation. Ex : "hain" , "hai"
  - Syntactic Analysis – controlling the nature of generated code mix output

- Is the generated output "acceptable" code mix?

- Attention to Diverse Language Pairs – tools that aren't language pair specific

Code Mix Machine Translation – End-to-End Pipeline– A Step Forward

CoMeT: Towards Code-Mixed Translation Using Parallel Monolingual Sentences. CALCS (NAACL) 2021

Code Mix Machine Translation – A Step in that direction

- Data – Utility in large scale pre-training
- Pre-processing
  - LID
  - Transliteration
  - Spelling Normalization
  - Syntactic Analysis
- Is the generated output "acceptable" code mix?
- Attention to Diverse Language Pairs – tools that aren't language pair specific

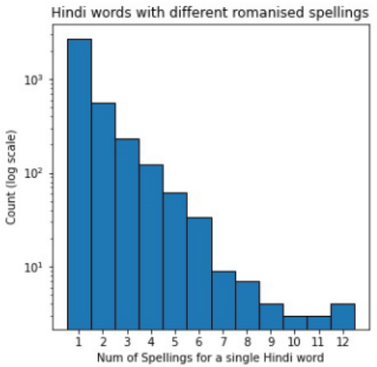


Figure 1: Multiple roman spellings for the same Hindi Word. These spelling variations can cause the BLEU score to be low, even if the correct Hindi word is predicted.

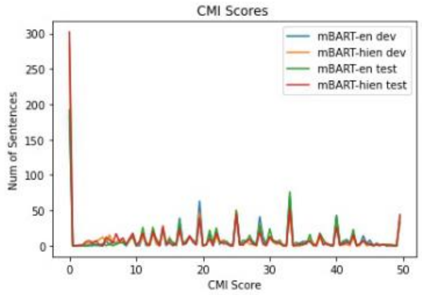


Figure 2: Code Mixing Index(CMI) for the generated translation of dev and test set .

Model	Validation Set		Test Set	
	BLEU	$\text{BLEU}_{\text{normalized}}$	BLEU	$\text{BLEU}_{\text{normalized}}$
mBART-en	**15.3**	18.9	**12.22**	–
mBART-hien	14.6	**20.2**	11.86	–

Gaps Identified
&
Current Work

Gaps Identified

Data Collection, Pre-Processing	Transfer Learning, Representation for CM	End-to-end pipeline	Other
Sentence Level CM Classifier	Analysis of large multlingual LMs for Code mixing	Machine Translation, Generation	Bias of Models trained on Code mix Data
Syntactic Analysis of Code Mix data	Modifying Multilingual LLMs for richer CM Representations	Benchmarks of CM	

Legend

Currently Working On

Part of Future Work

Infavourable Results.
Needs reformulation

Publications

- **CoMeT: Towards Code-Mixed Translation Using Parallel Monolingual Sentences.**
  - Venue : *Fifth Workshop on Computational Approaches to Linguistic Code-Switching, NAACL '21*
  - Authors : Devansh Gautam, Prashant Kodali, Kshitij Gupta, Anmol Goel, Manish Shrivastava, Ponnurangam Kumaraguru

- **Battling Hateful Content in Indic Languages HASOC'21**
  - Venue : To be presented at FIRE '21
  - Authors : Aditya Kadam, Anmol Goel, Jivitesh Jain, Jushaan Singh Kalra, Mallika Subramanian, Manvith Reddy, Prashant Kodali, TH Arjun, Manish Shrivastava, Ponnurangam Kumaraguru

Limitations

- Primary focus on code mix text from Online Social Networks. Speech as source of code mix data is not addressed in this study.

- Aims to formulate computational pipelines capable of procesing code mix sentences. Other aspects of Code mixing – Grammatical theories, socio-linguistic analysis is not the primary area of contribution.

Acknowledgements

- Collaborators – The interactions which challanged my understanding of topics
- Lab mates – whose constant feedback shaped this work and for being constant source of inspiration.
- Researchers who chose to FOSS their data and code base
- Teachers who taught me basics of CL / NLP .

Thank You!

Questions?