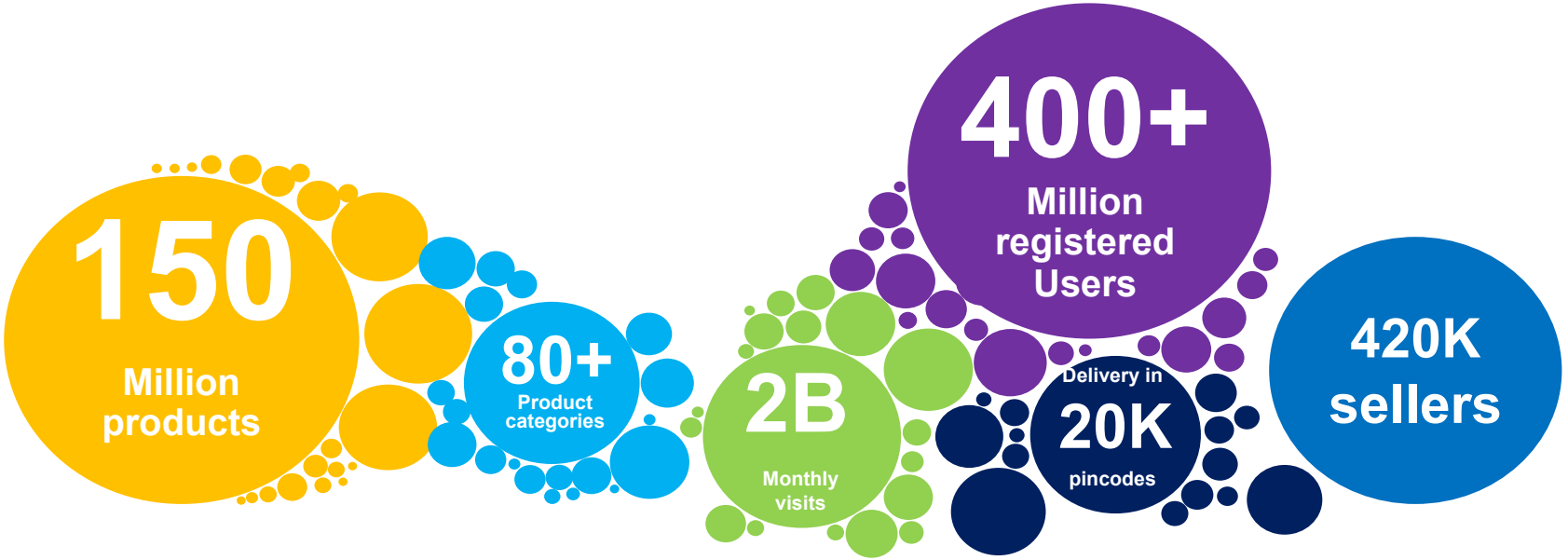
ACM India at a Glance

- **ACM**: world's largest educational and scientific computing society
  - **Mission**: advancing computing as science and profession
  - **Members**: ~100,000 worldwide, ~13000 in India
  - Comprising students, faculty, professionals
- **ACM India Chapters**: ~200 student chapters, ~20 professional chapters
- **ACM-W India**: empowering women in computing
- **Research Initiatives**
  - Student research: ARCS Symposium, best doctoral dissertation, partial travel grant, PhD clinic and Anveshan Setu
  - SIG research conferences: CODS-COMAD, ISEC
- **ACM India Annual Event**
  - Discuss recent trends in technology and celebrate India's achievements in computing
- **Education Initiatives**
  - Summer and winter schools: ~2 week full-time course on technology area
  - Compute: Symposium on computing education
  - Teaching Partner Program: External experts offering a course
  - CSpathshala: inculcate computational thinking in schools
- **Learning and Professional Development**
  - Eminent Speaker Program
  - Industry Webinars, Education Webinars
  - Minigraphs: Comprehensive coverage of a tech area
  - ACM global resources: Digital Library, ACM Learning Center
- **Individual Professioal Awards**
  - Acknowledge and celebrate outstanding contributions
- **ACM Membership in India**
  - Student? student member form
  - Professional? professional member form

Onboarding the next 500Mn users

Mayur Datar
Chief Data Scientist, Flipkart

Fast facts / Scale of Flipkart business



Flipkart

Let's Talk about Scale - BBD 2021!

Flipkart

Total visits in TBBD'21 is ***3X the population of the United States***

Credit extended can fund ***11 Chandrayan launches to the moon!***

Fastest Delivery in just ***10 mins through Flipkart Quick!***

All footwear boxes sold and stacked together are ***100X the height of Mt. Everest***

Total saplings sold can produce ***37000 ltrs of Oxygen Daily!***

Total light bulbs sold can light up ***5 Eiffel Towers!***

Fun Facts

WE SOLD

Height of all smartphones sold and stacked is **> 1000 Burj Khalifas**

120,000 chocolate bars sold in **24 hours**

Flights sold can make **2400 trips around Earth**

3.6 million plays of **TBBD'21 Games**

Solving for Bharat

Scale and diversity

- 1.3 Billion people
- 22 scheduled languages + hundreds of dialects
- Culture changes from state to state

Low literacy levels

- Jeans, jins, jute, shoes, …

Poor infrastructure

- Internet speeds
- Background noises

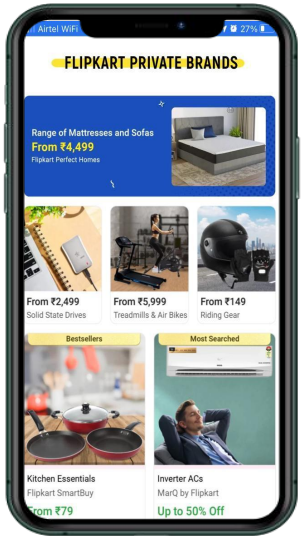
Low purchasing power

- Thin margins, cost of fulfillment
- Fraud

Making e-commerce accessible to the N500 mn customers

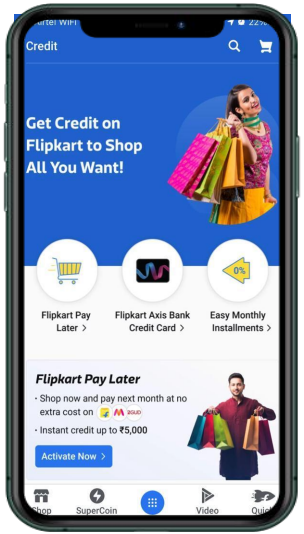
AFFORDABILITY

Affordable Products



Quality Selection Made Affordable

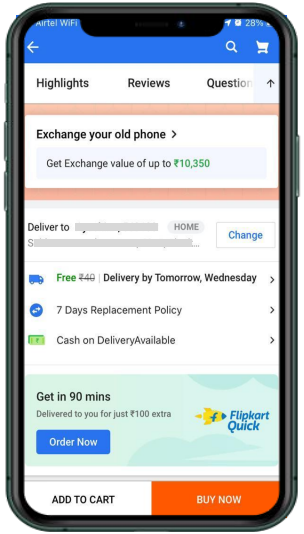
Affordable Credit



Digital Financial Inclusion to Address Credit Needs

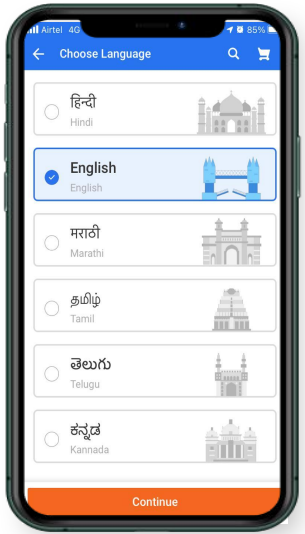
ACCESS

Physical Access



Free, Fast and Reliable Delivery

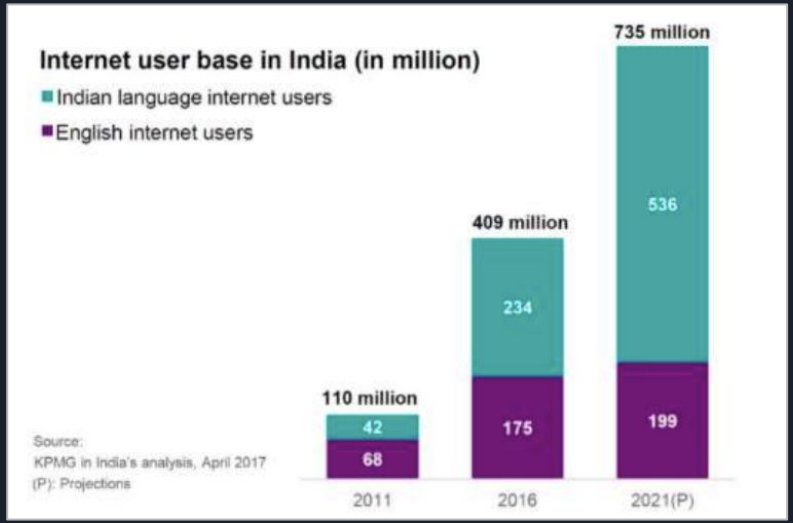
Digital Access



Vernacular and Voice to Onboard Next Wave of Shoppers

Vernacular growth in India

- Only 10% of Indian population is versed in English
- Internet users with Indic languages growing at fast pace - next 200M users



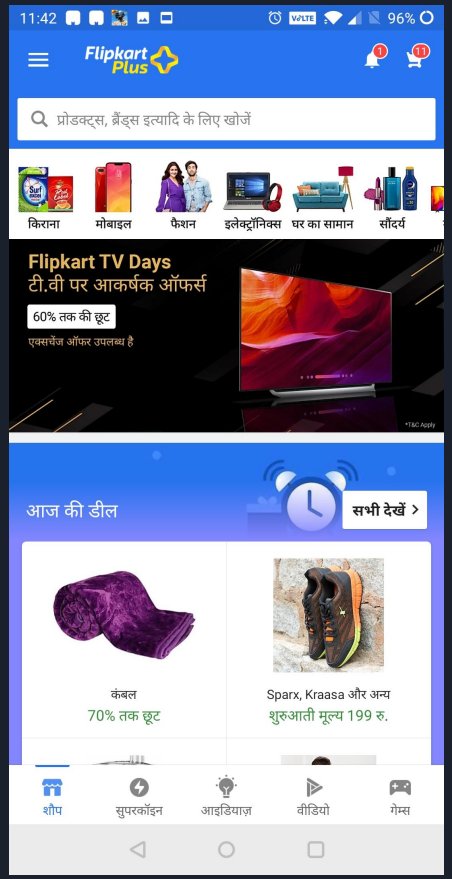
Challenges faced by Vernac users:

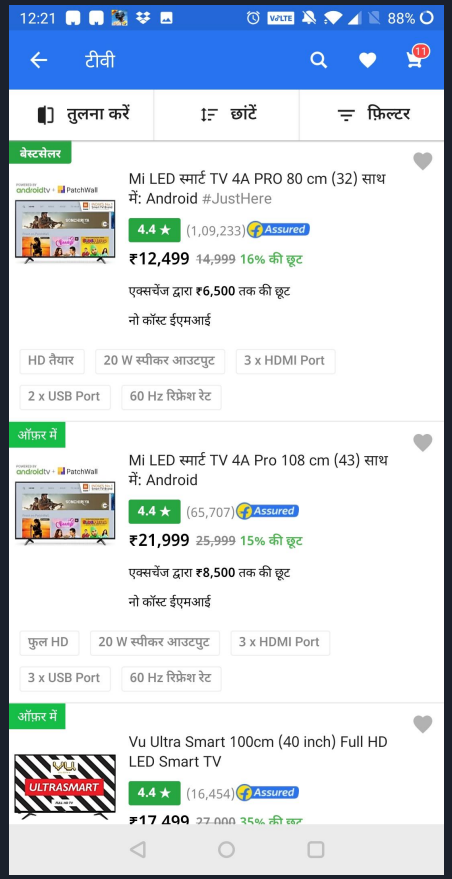
- Inability to comprehend English content
- Lack of trust and confidence

Provide the product pages in regional languages to build familiarity and trust for vernac user base

- **Talk to your customers in their language.**

Mission: Scale to 11 Indian languages!





Areas for Machine Translation

- Catalog Attributes
- Product Descriptions
- User Reviews
- Search Queries
- Addresses
- Customer service chats
- ....

Areas for Machine Translation

- Catalog Attributes
- **Product Descriptions**
- **User Reviews**
- **Search Queries (User path, real time)**
- Addresses
- Customer service chats

....

Challenges

- Data challenges:
  - Low resource constraints for Indian language pairs
    - Lack of large amounts of parallel corpora, comparable/strong baselines, trained models/code for Indian language pairs as compared to the efforts for European languages (such as EuroParl [Koehn, 2005] and Paracrawl [Bañón et al. 2020])
  - Lack of in-domain (ecommerce) data

- Noisy data:
  - Improper sentence structure - ex. Light but do it is work, no good no buy
  - Spelling mistakes, Missing punctuations, Pronouns or articles are dropped
  - Emojis, Missing spaces, repeating characters
  - Code-mixed data
    - nice product gajab camera quality and dhansu product
    - osm product. lekin price bahot jyada hai

Challenges: Colloquial Translations

Product Description	Colloquial Translation(Good)	Non Colloquial Translation(Bad)
A good quality product for you	आपके लिए एक अच्छी क्वालिटी का प्रोडक्ट	आपके लिए एक अच्छी गुणवत्ता का उत्पाद
Our Boots are available in eye catching looks and designs.	हमारे बूट्स आई कैचिंग लुक्स और डिज़ाइन में उपलब्ध हैं।	हमारे जूते आंख को पकड़ने वाले रूप और डिजाइन में उपलब्ध हैं।
And to Filter Ultraviolet Rays.	और अल्ट्रावायलेट रे को फ़िल्टर करने के लिए है।	और पराबैंगनी किरणों को फ़िल्टर करने के लिए है।
You Maniacs! you won't want to miss this awesome collection of 1970S and 1980s-inspired figures from the Planet of the Apes films!	आप मेनियाक हैं! आप 1970S और 1980s के इस शानदार कलेक्शन को मिस नहीं करना चाहेंगे - प्लेनेट ऑफ़ द एप्स से प्रेरित फिगर्स!	आप पागल! आप 1970S और 1980 के दशक प्रेरित फिल्मों के ग्रह से प्रेरित आंकड़ों के इस भयानक संग्रह को याद नहीं करना चाहेंगे!

Model Training

We train our System in 4 phases

1. Translation Pre-training using a transformer model
2. In-domain data generation and Synthetic Data filtering
3. Transfer Learning for Colloquial translations
4. Model re-training

Transformer based approach

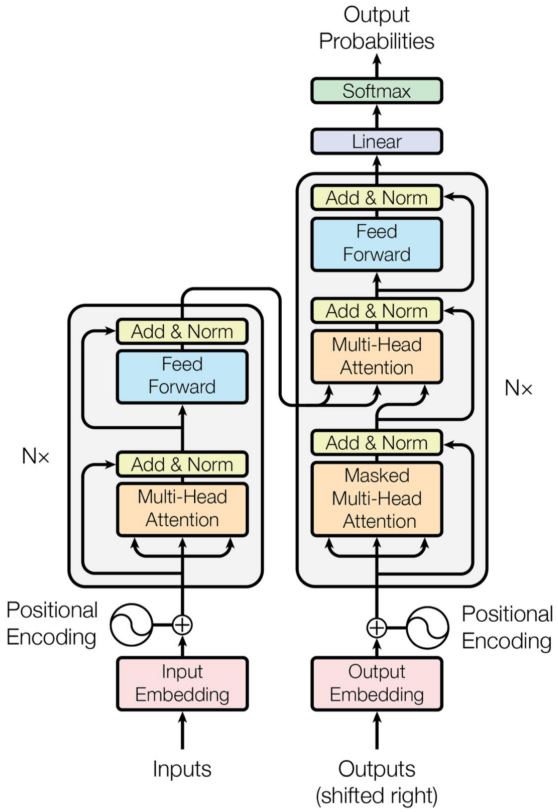
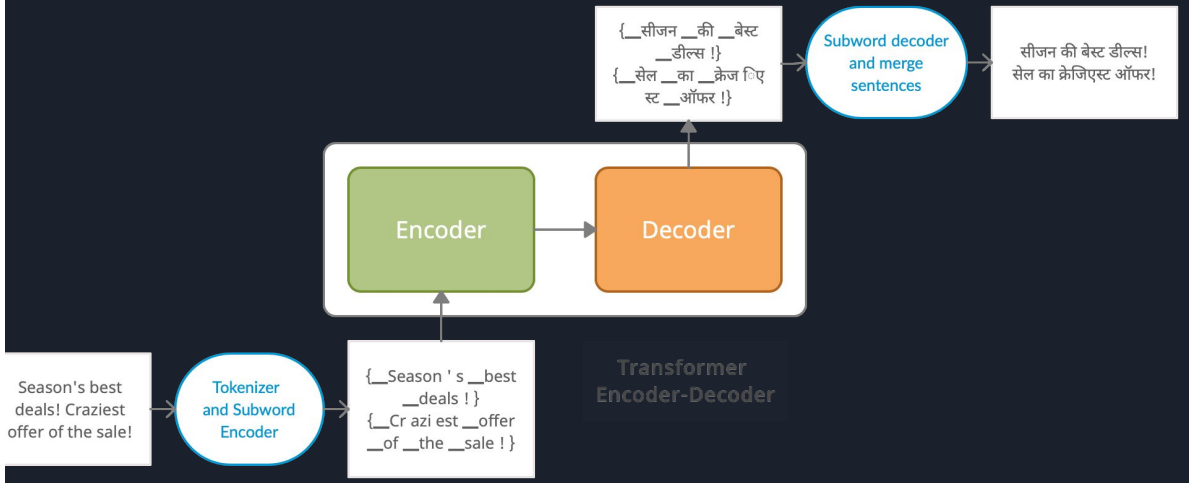
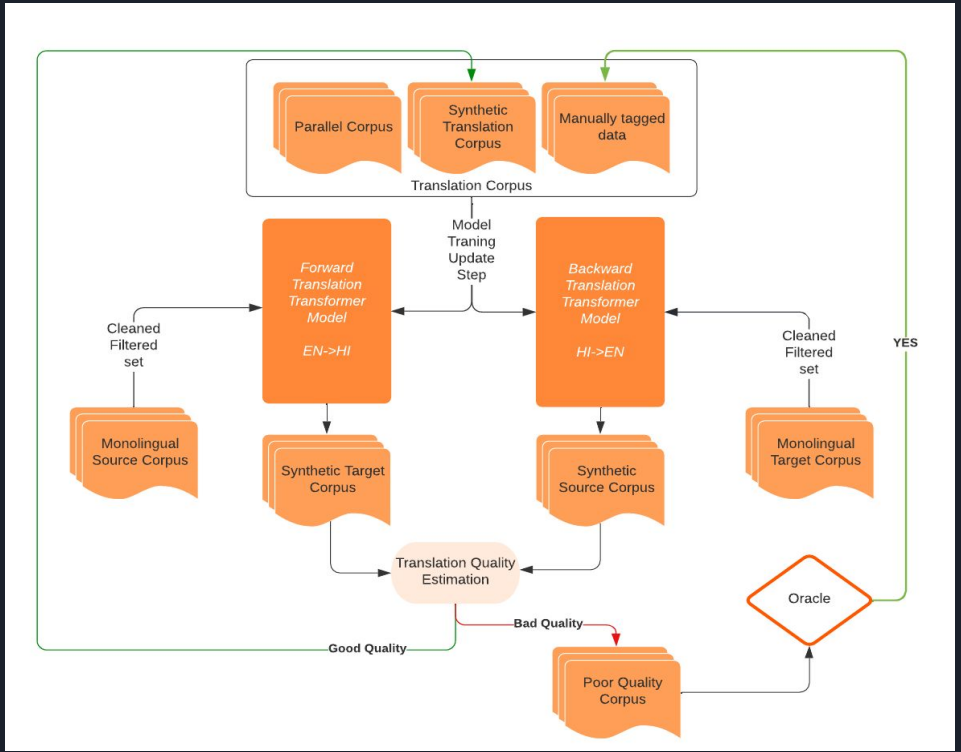


Figure 1: The Transformer - model architecture.



Solving for lack of data challenge

1. Clean and Filter abundant monolingual data
2. Generate high quality synthetic translation Corpus
3. Sample bad quality synthetic translations and correct the pairs using manual taggers
4. Re-train Translation models with new Translation Corpus



Solving for colloquial language challenge

- Model translation quality score using BERT based models trained on manually tagged in-domain data

  (Buy this keychain from Confident , क्योंकि ये किचेन कॉन्फिडेंट से ।) - 0.20

  (Package Includes:, पैकेज में शामिल है:) - 0.94

- Limitation: Non-colloquial translations are still scored very high

  (Printed Key Chain, मुद्रित कुंजी श्रृंखला) - 0.94 not colloquial!

- Transformer LM scoring - *lower perplexity is better*

28.8	यह उत्पाद अच्छा है	434.0	टेम्प र्ड ग्लास दृश्य ता का त्याग किए बिना आपकी डिवाइस स्क्रीन पर खरोंच को रोक देगा
20.0	ये उत्पाद अच्छा है	232.8	टेम्प र्ड ग्लास बिना दृश्य ता का बलिदान किए आपके डिवाइस स्क्रीन पर खरोंच से बचा एगा
9.6	ये प्रोडक्ट अच्छा है	128.4	टेम्प र्ड ग्लास वि जि बिलिटी का त्याग किए बिना आपके डिवाइस स्क्रीन पर स् क्रै च से बचा एगा

Question: How to threshold perplexity? - *no relative sense in different ppl scores*

Results: Product Description Translation

Bleu Results

Model	Hindi	Kannada	Tamil	Telugu	Bengali	Marathi
State of the art external model	29.29	31.82	31.86	30.78	24.33	28.86
FK Model	**42.88**	**32.69**	**44.48**	**39.88**	**31.92**	**32.69**

Results: Product Description Translation

Bleu Results

Model	Hindi	Kannada	Tamil	Telugu	Bengali	Marathi
State of the art external model	29.29	31.82	31.86	30.78	24.33	28.86
FK Model	**42.88**	**32.69**	**44.48**	**39.88**	**31.92**	**32.69**

Manual evaluation Results

Hindi	FK Model V1	State of the art external models
Good	41.30%	4.60%
Can be better	55.24%	17.18%
Bad	**3.46%**	**78.21%**

Telugu	FK Model V1	State of the art external models
Good	16.33%	12.88%
Can Be Better	79.31%	80.66%
Bad	**4.38%**	**6.46%**

Tamil	FK Model V1	State of the art external models
Good	34.23%	16.91%
Can be better	48.41%	59.61%
Bad	**17.34%**	**23.46%**

Bengali	FK Model V1	State of the art external models
Good	33.70%	11%
Can be better	53.93%	78.42%
Bad	**8.25%**	**10.58%**

Results: Review Translation

BLEU Results (for various lengths of reviews)

BBD test_set	1-5 words	6-10 words	11-20 words	21-30 words	31-50 words	Overall
counts	2153	2012	2040	2017	2079	10301
State of the art external model	24.00	21.60	24.32	26.91	29.26	27.17
FK Model	**24.87**	**25.82**	**26.83**	**28.01**	**30.15**	**28.47**

Results: Review Translation

BLEU Results (for various lengths of reviews)

BBD test_set	1-5 words	6-10 words	11-20 words	21-30 words	31-50 words	Overall
counts	2153	2012	2040	2017	2079	10301
State of the art external model	24.00	21.60	24.32	26.91	29.26	27.17
FK Model	**24.87**	**25.82**	**26.83**	**28.01**	**30.15**	**28.47**

Manual evaluation Results

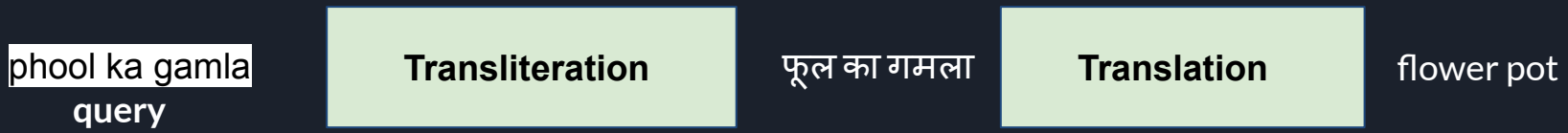
	Good	Can be better	Bad
FK Model	30.7%	59.0%	**10.3%**
State of the art external model	7.40%	78.35%	**14.25%**

Challenge: Code mixed Search Translation

Hinglish Query	English Translation	Issue
semsung phone 8 hajar vala	samsung phone for 8 thousand rupees	Brand mispelling
redami 5a ki baitiry	redmi 5a battery	Brand mispelling
goldan kalar ka aie linar	golden colour eye liner	Spell errors
lenovo k6 ka kabar	lenovo k6 cover	Spell errors
makhamal wala seutar	velvet sweater	Spell errors
resmi 7a kwr	redmi 7a cover	Spell errors
250 tak ki biluthuth	bluetooth upto 250	Spell errors
baal hawa machine	hair dryer machine	Articulation Gap
juta bina dori wala	shoe without lace	Articulation Gap
teen char sau wali saree sadi	saree for 300-400 rupees	Free flowing text
kam keemat mein four g mobile	4g mobile in low price	Free flowing text

Solution

Standard workflow



Convert to Hindi text in Devanagari

Translate from Hindi to English

- **Problem: Error compounding**
- **Solution: Build direct Hinglish => English translation model**
    - Challenge: Need of larger training set to capture query variations

Solution

Standard workflow

phool ka gamla
query

Transliteration

फूल का गमला

Translation

flower pot

Convert to Hindi text in Devanagari

Translate from Hindi to English

New workflow

phool ka gamla
query

Transformer (Encoder-Decoder)
Code-mix Translation model

flower pot
English

Synthetic data augmentation to capture query variations

Solution (Contd): Data augmentation

- Labeled Data
- Auto encoder
- Masking
- Drop Char
- Word Order Permute

Results

Setting	BART-base	T5-base	Bert2Bert
Baseline	31.88	32.8	19.88
+AutoEncoder	+1.37	+1.13	+7.13
+CharDrop	+1.32	+1.14	+1.8
+Masking	+1.14	+1.43	+5.02
+WordOrderPermute	-3.1	-7.4	+5.6

Bleu scores comparing various data augmentation approaches

Results for code-mixed search translation

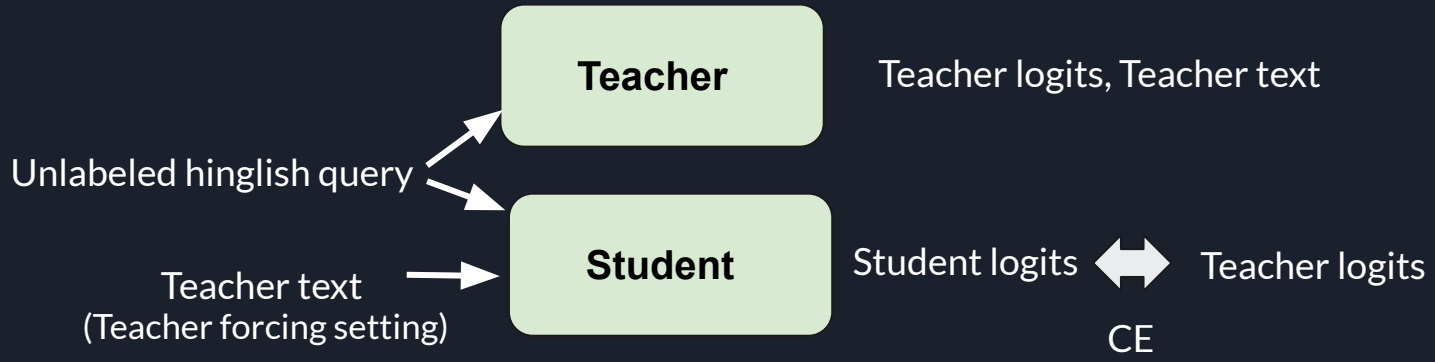
Model	Accuracy	% Good	% Bad
Live model	62.9	46.6	37.1
Transformer v3	**78.7**	**61.3**	**21.2**

Model	Head	Torso	Tail
Live model	72.3	62.9	53.3
Transformer v3	**92.9**	**79**	**64**

Another challenge: Latency

Solution: Knowledge Distillation for NMT

- Well known technique: Pseudo labeling [1]
- For search queries, we explored following distillation scheme



- Matching probabilities transfer more knowledge between teacher and student [2]

1. Pre-trained summarization distillation, Sam et.al., 2020
2. Sequence level knowledge distillation, Yoon et. al, 2016

Latency reduction with Knowledge Distillation

Model	# layers	KD loss	BLEU	p95
BART teacher	6	-	51.3	~ 200 ms
Student (Q.)	1	CE	49.7	~ 47 ms
Student (Q.)	1	JS	50.8	~ 47 ms
OpenNMT Student (Q.)	1	-	51.1	~ 14 ms
Only OpenNMT (Q.)	1	-	38.5	~ 14 ms

Results for Knowledge Distillation

Learnings

1. Lack of parallel data + domain data
   - Data generation with Forward/Back-translation + Data filtering + Transfer Learning

2. Noisy Data
   - Synthetic noisy training data + Sub-word modeling

3. Colloquial vs Non-Colloquial Translations
   - Transfer Learning + Colloquial data filtering

4. Latency for real-time translations
   - Knowledge Distillation