

Project Mélange:

Microsoft[®] Research

Processing and Understanding Mixed Language

Kalika Bali, Monojit Choudhury, Sunayana Sitaram Microsoft Research India, Bangalore, India {kalikab, monojitc, sunayana.sitaram}@microsoft.com

Code-mixing or Code-Switching is the mixing of two or more languages in a conversation or even an utterance. Kibrisa geldigim ... god Sous la pluie mais beau tout de warum? ich mochte même, chère Ileana! nicht hier Buona giornata a te e a tutti! Coridel Ent merilis full tracklist untuk no me lebante ahorita debut mini album cuz I felt como si me Jessica Jung yg akan kemara por dentro segera rilis bulan Mei mendatang jit fi la fin du mois de dece-mbre kan ljaw bared ktir wttalj

All Speech/NLP Applications break down in presence of Code-switching

Cortana, aaj Hyderabad ka weather kaisa hai? Is it raining ya sunny day hai?

Opinion mining

Adik... sem brape boleh bwak kenderaan? normal parent question – UiTMLendufornia

Why is it challenging?

Problem of Data

Code-mixing is predominantly a spoken phenomenon.
So no large text corpora.

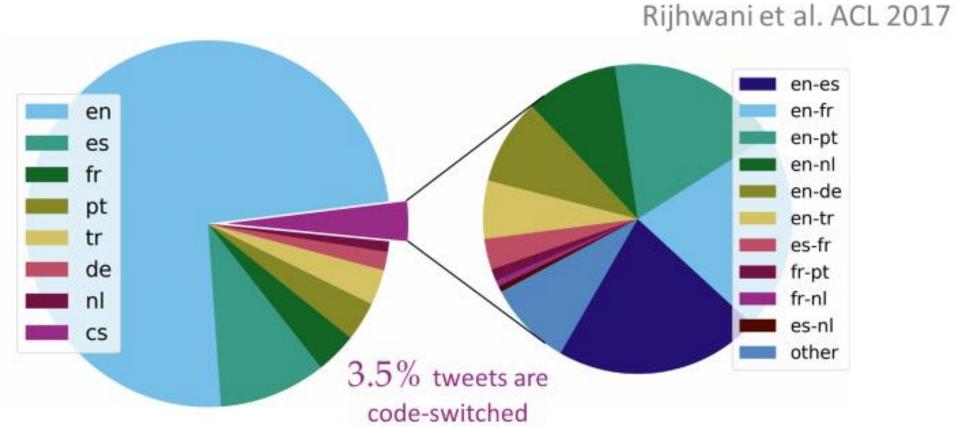
Model Explosion

With n languages, there are O(n²) potential code-mixed pairs.

Reusing Models

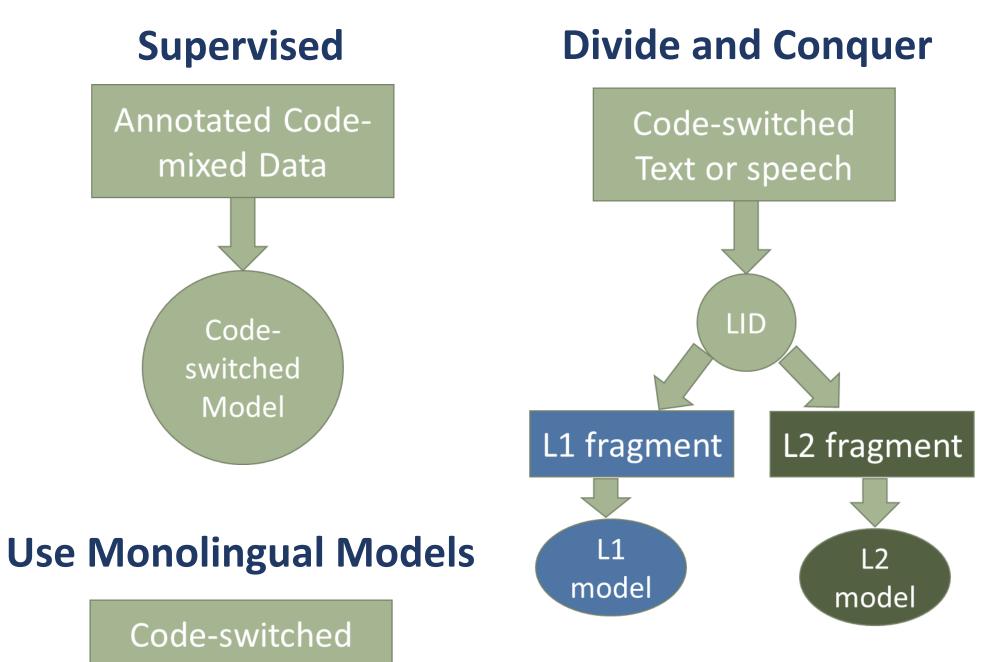
How to exploit the monolingual models and data for code-mixing.

How prevalent is code-switching?



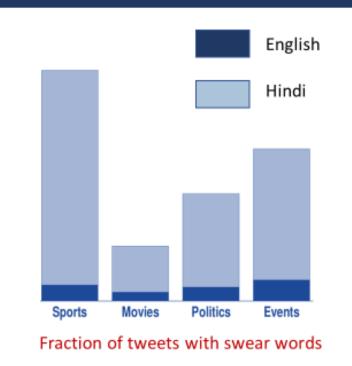
Worldwide language distribution of monolingual and code-switched tweets computed over 50M Tweets (restricted to the 7 languages)

How do we solve it?



Why and when do people code-switch?

We might praise you in English, but *gaali to Hindi me hi denge*! (Rudra et al., EMNLP 2016)

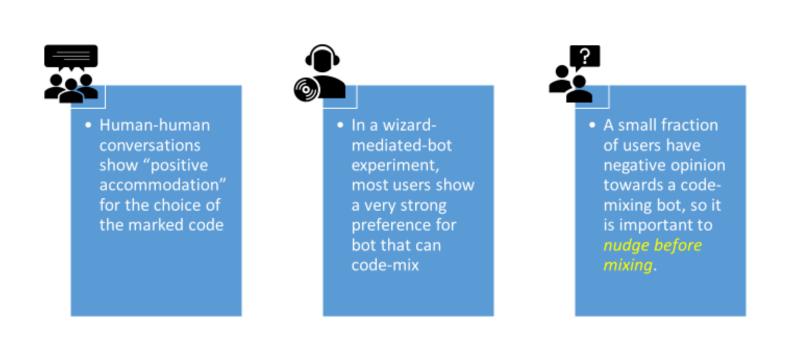


Study of 830K Tweets from Hi-En bilinguals

- The native language, Hindi, is strongly preferred (10 times more) for negativity and swearing
- 2. English is used far more for positive sentiment than negative
- Language change often corresponds with changing sentiment

Should chat-bots code-switch?

Nudge, don't push or assume...



Combination Logic or ML

Text or speech

L1

model

L2

model

Tasks and Applications

Language Identification
Part of Speech Tagging, Parsing
Language Modeling
Machine Translation
Speech Recognition
Speech Synthesis

