

Processing Code-Mixed Text

Mohd Sanad Zaki Rizvi
Monojit Choudhury

Microsoft
monojitc@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
warum? *ich mochte*
nicht hier

Sous la pluie mais beau tout de
même, chère Ileana!
Buona giornata a te e a tutti!

no me lebante ahorita
cuz I felt como si me
kemara por dentro

Coridel Ent merilis
full tracklist untuk
debut mini album
Jessica Jung yg akan
segera rilis bulan Mei
mendatang

jit fi la fin du mois de dece-mbre *kan*
ljaw bared ktir wttalj





Outline of the tutorial

- Importance of code-mixing and multilingual processing for CSS
- Hands-on-session on Word-level Language Detection
- Advanced text processing



Project Mélange

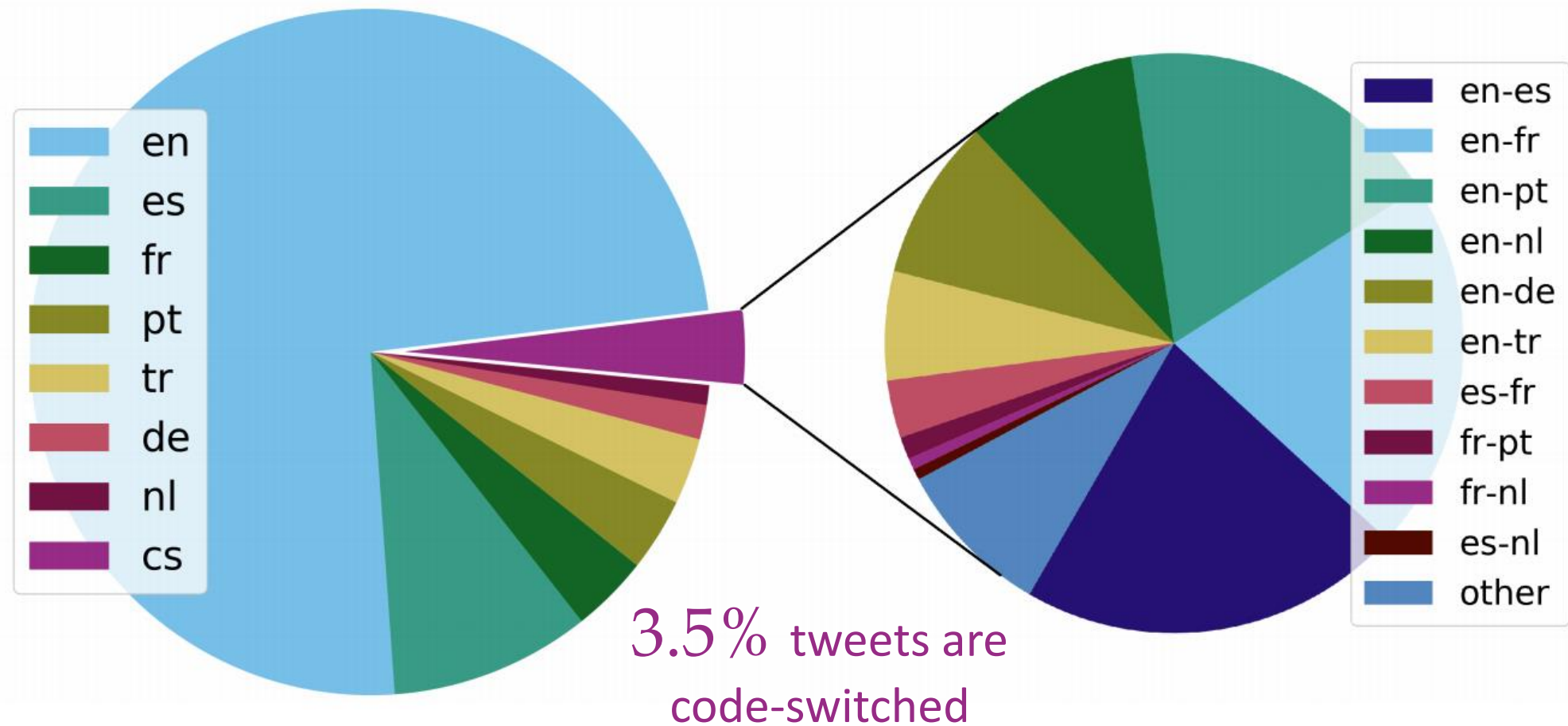
Established: January 1, 2012

Acknowledgments

Kalika Bali, Sunayana Sitaram, Mohit Jain, Tanuja Ganu, Ishani Mandol,
Anirudh Srinivasan, Simran Khanuja, Sanket, Sebastin Santy, Adithya Pratap,
Anshul Bawa, Shruti Rijhwani, Gayatri Bhat, Koustav Rudra, Rafia Begum,
Royal Sequeria, Yogarshi, Spandana Gella

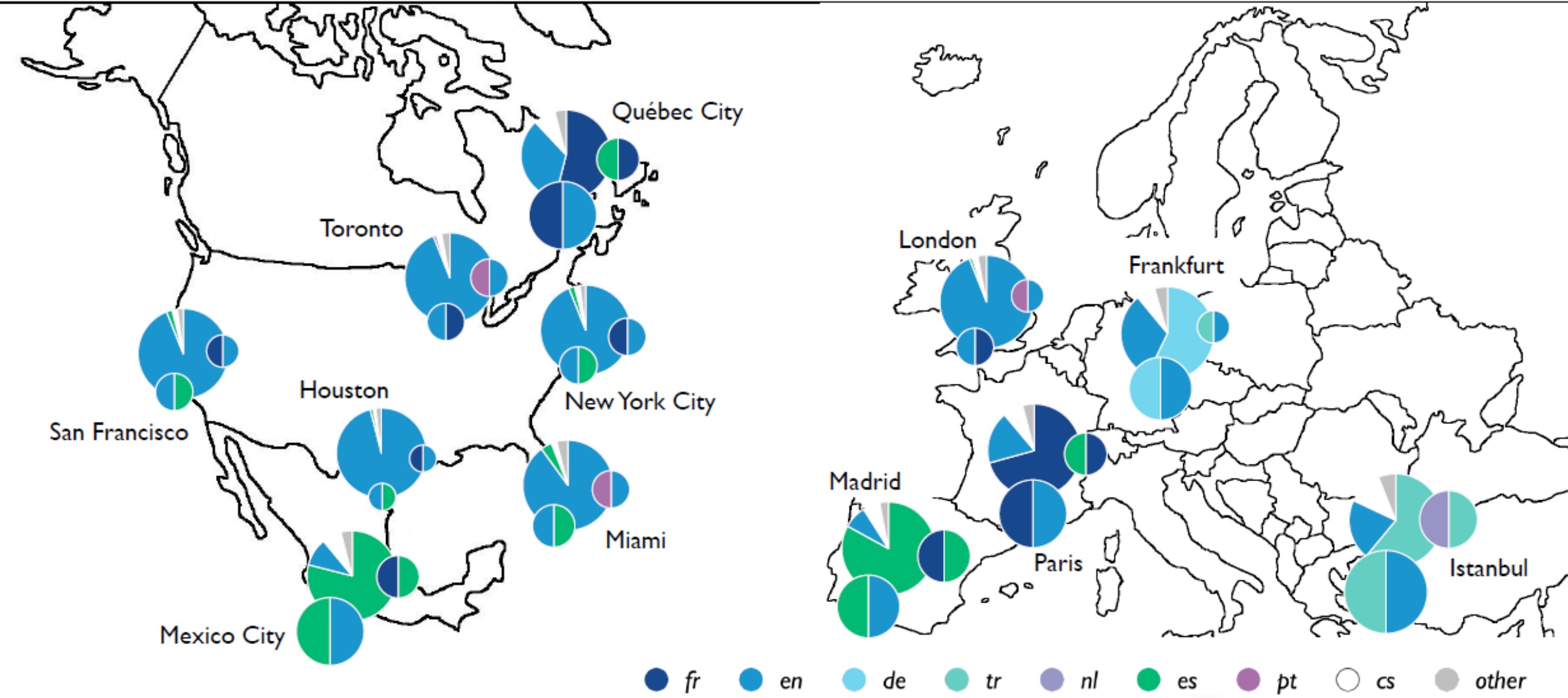
Code-mixing

- Happens in all multilingual societies
- Is predominantly a spoken language phenomenon
- Is generally associated with informal conversations
- Has well-defined socio-pragmatic functions



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

Geographical Distribution of Code-switching on 8M Tweets from 24 cities



NLP Technologies deployed in or for analysis of data from Multilingual societies must be able to process code-mixing

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



Intersteller es
una amazing
movie!

**Social Media
Analytics**

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

Socio-
Pragmatic
Functions of
Code-mixing
*When and
why do
bilinguals
prefer a
certain
language?*

Topic change

Puns

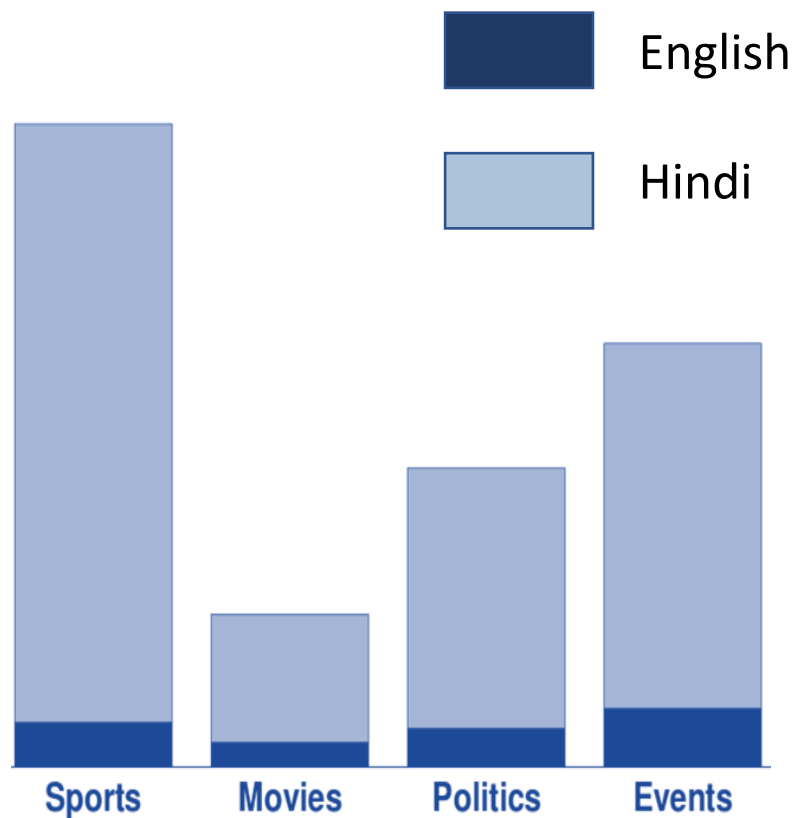
Emphasis

Emotion

Reported Speech

But it's unpredictable!

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



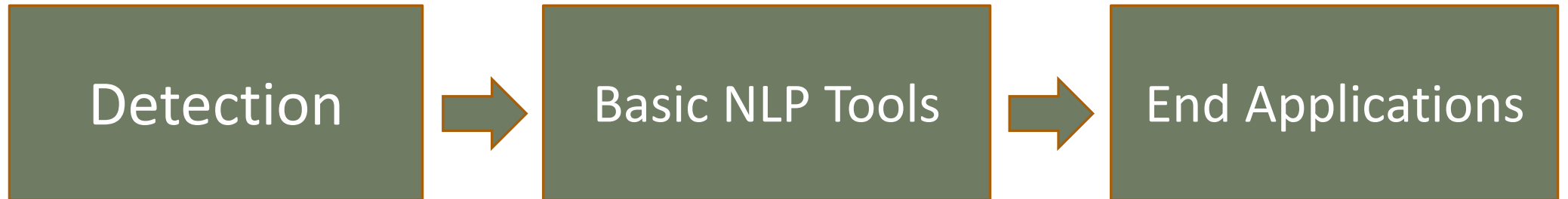
Fraction of tweets with swear words

Study of 830K Tweets from Hi-En bilinguals

1. 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
3. Language change often corresponds with changing sentiment

Inferences drawn from data in a single (usually the majority) language are likely to be misleading for multilingual societies.

Processing Code-mixing



Language Detection

I know when you switched from Inglés a Español

En En En En En En Sp Sp Sp

Pairwise Language Detection

Kalam ke speech se India inspired ho gaya #respect

NE	Hn	En	Hn	NE	En	Hn	Hn	Other
----	----	----	----	----	----	----	----	-------

Can't we just use dictionaries?

Challenges

Dilwale vs. Bajirao Mastani: Even
Super-Films Get the Monday
Blues

Named Entities

What was your favourite moment
at the concert? Was war für euch
der schönste Moment

Ambiguity

Wat n awesum movie it wazzzz!
sabko dekhna chahiye

Spelling Variations

Transliteration

Out-of-Vocabulary Words

Character n-gram based classifiers for each language

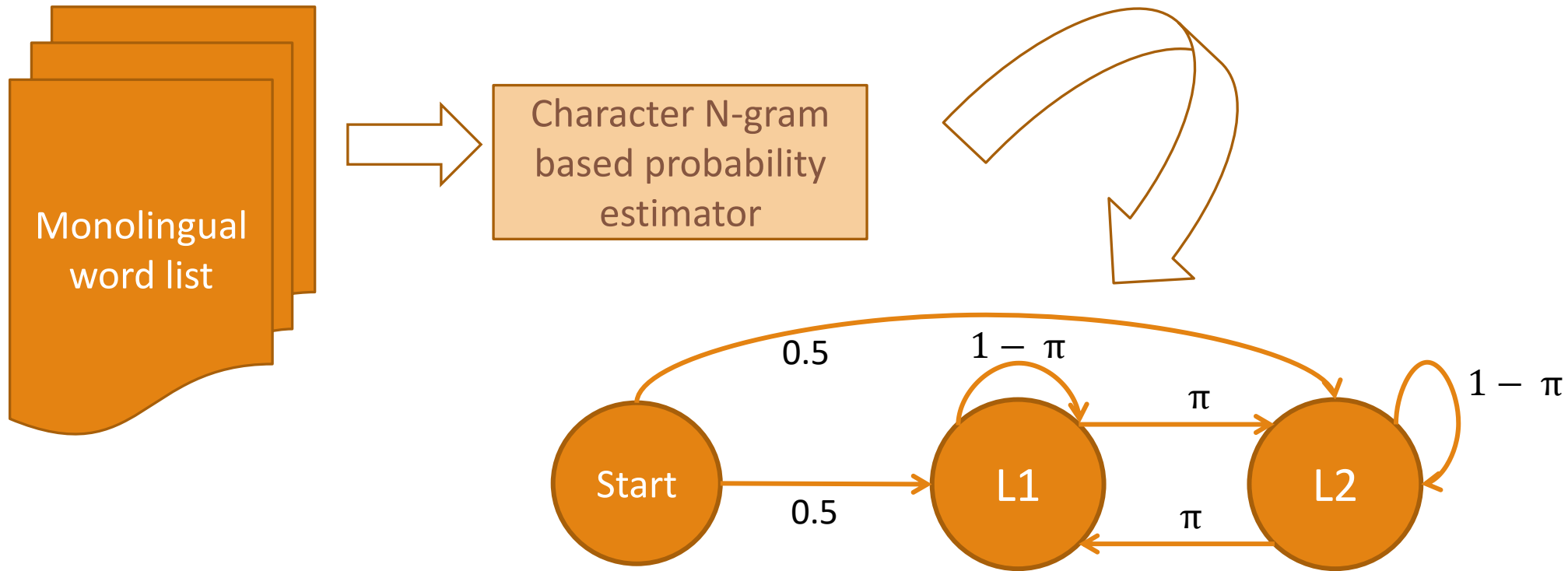
\$INSPIRED\$



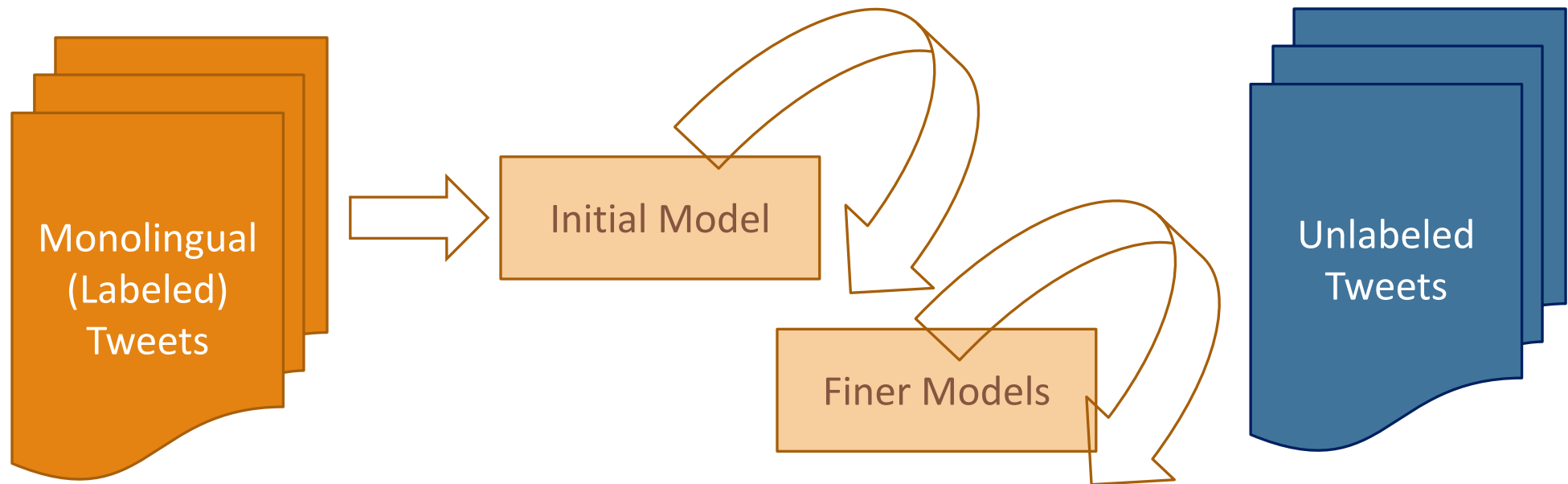
\$IN	INS	NSP	SPI	PIR	IRE	RED	ED\$
------	-----	-----	-----	-----	-----	-----	------

Character Trigrams

1. Simple Pairwise language labeling

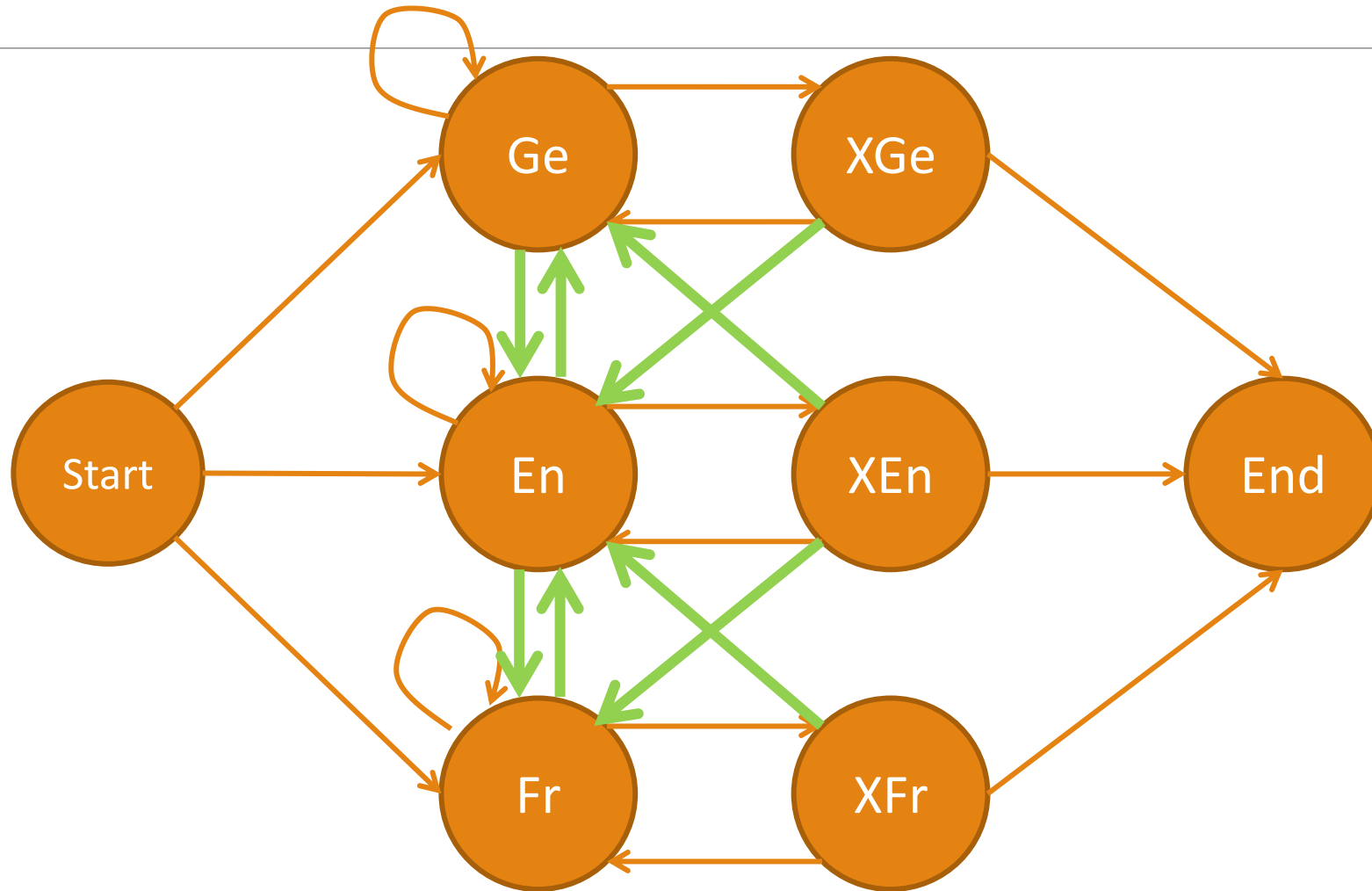


2. Semi-supervised Learning with Weak Labeling



Rijhwani et al. Estimating code-switching on twitter with a novel generalized word-level language detection technique. ACL 2017

Initial Model from Weakly Labeled Data



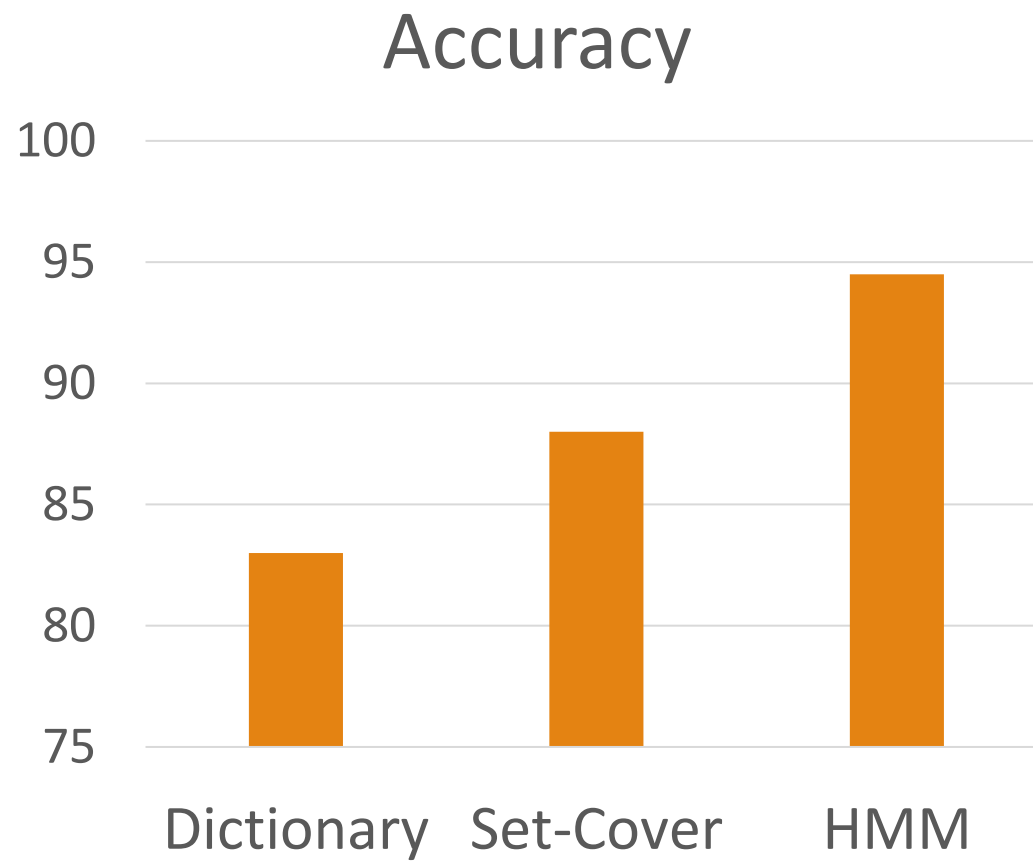
Experiments & Results

Languages (7): Dutch, English, French, German, Portuguese, Spanish, Turkish

Weakly Labeled Data: 85K Tweets for each language

Unlabeled Data: 2M Tweets

Test Set: 5000 Tweets, manually annotated.

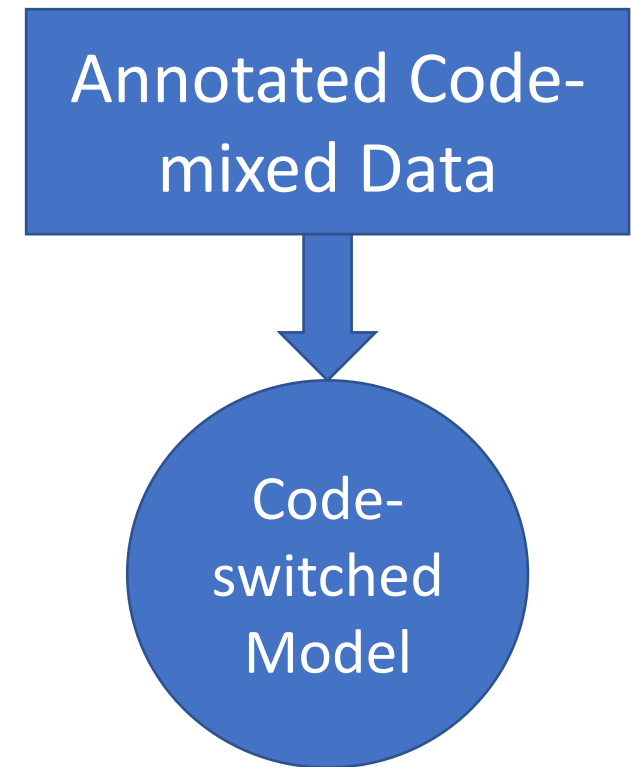


Language Detection

HANDS-ON SESSION

Computational Models of Code-Switching

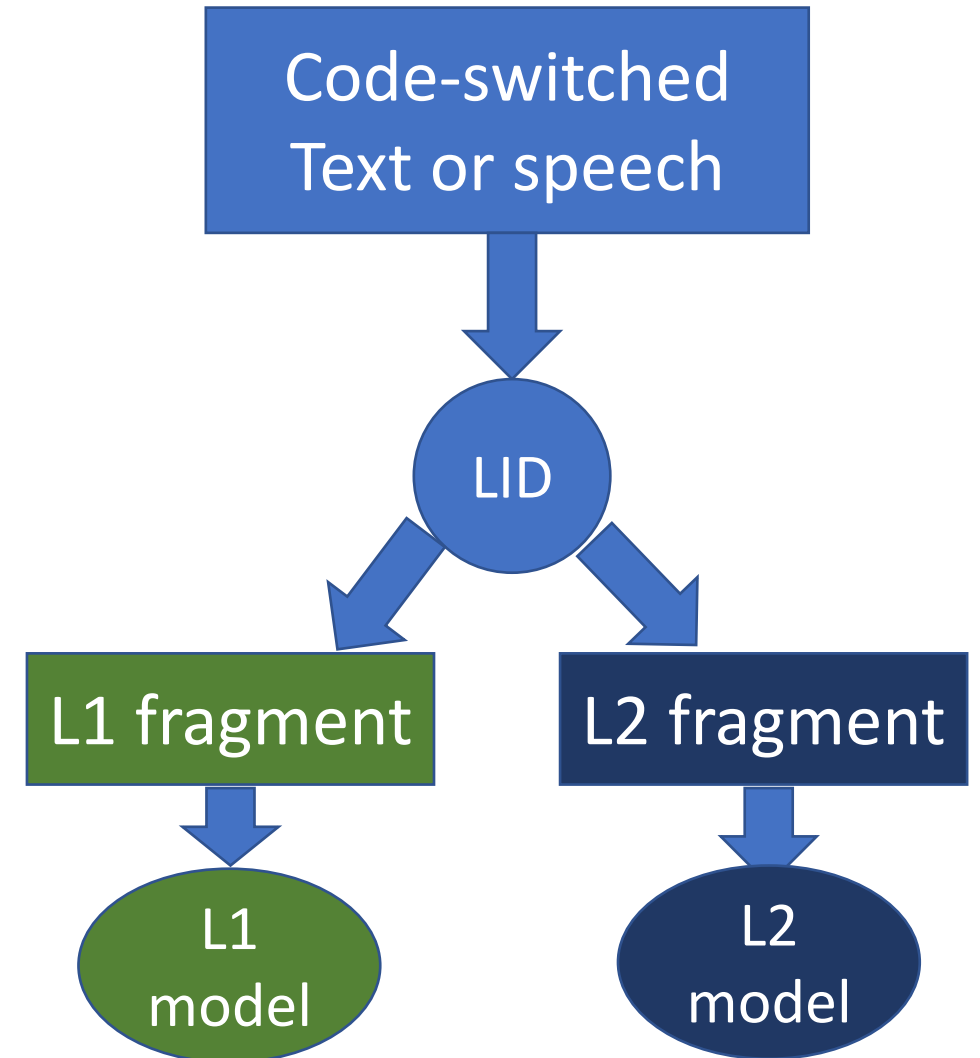
- Supervised *i.e., from scratch*
- Divide & Conquer
- Combining Monolingual Models
- Zero-shot learning



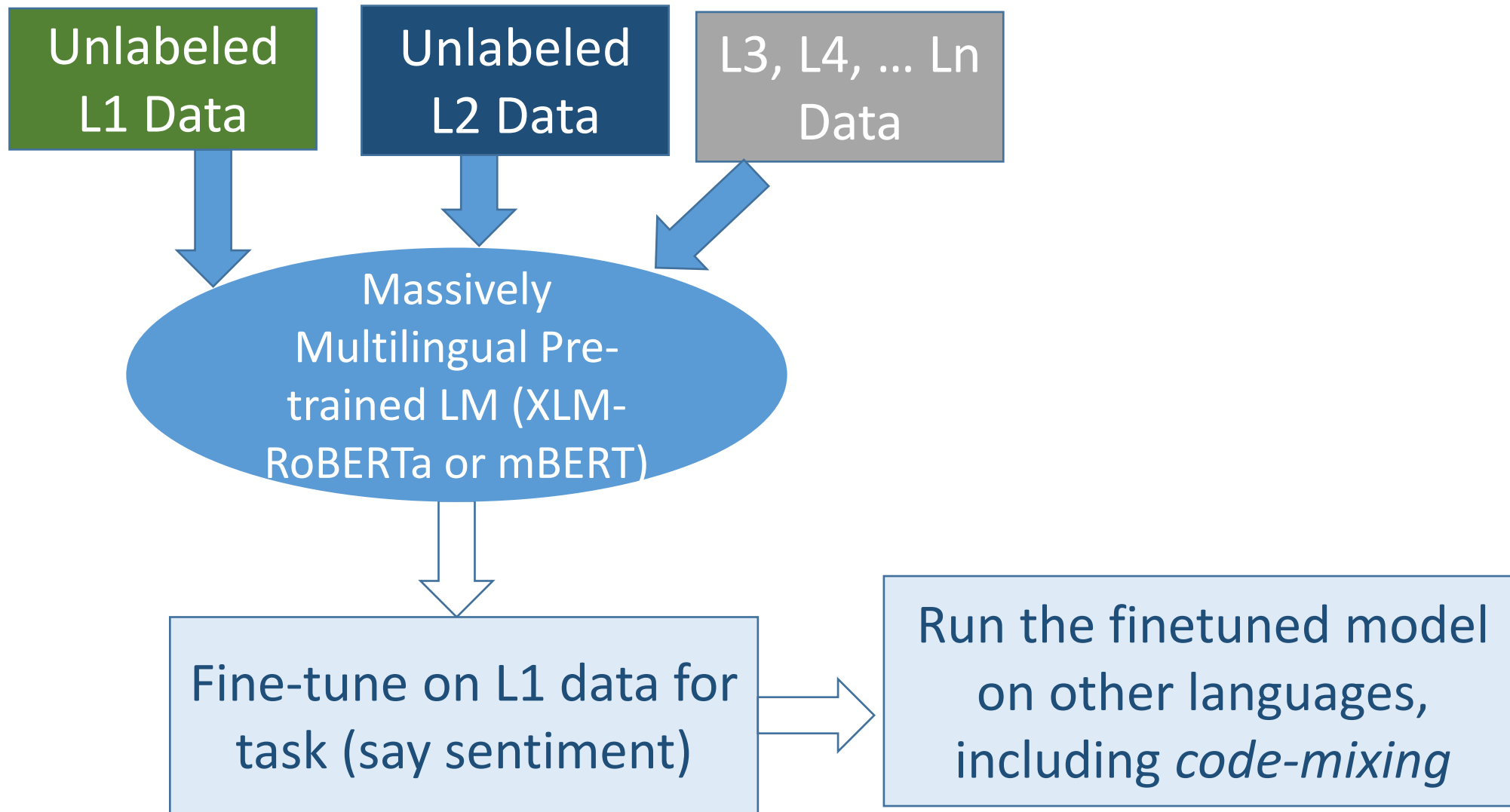
Computational Models of Code-Switching

- Supervised *i.e., from scratch*
- Divide & Conquer
- Combining Monolingual Models
- Zero-shot learning

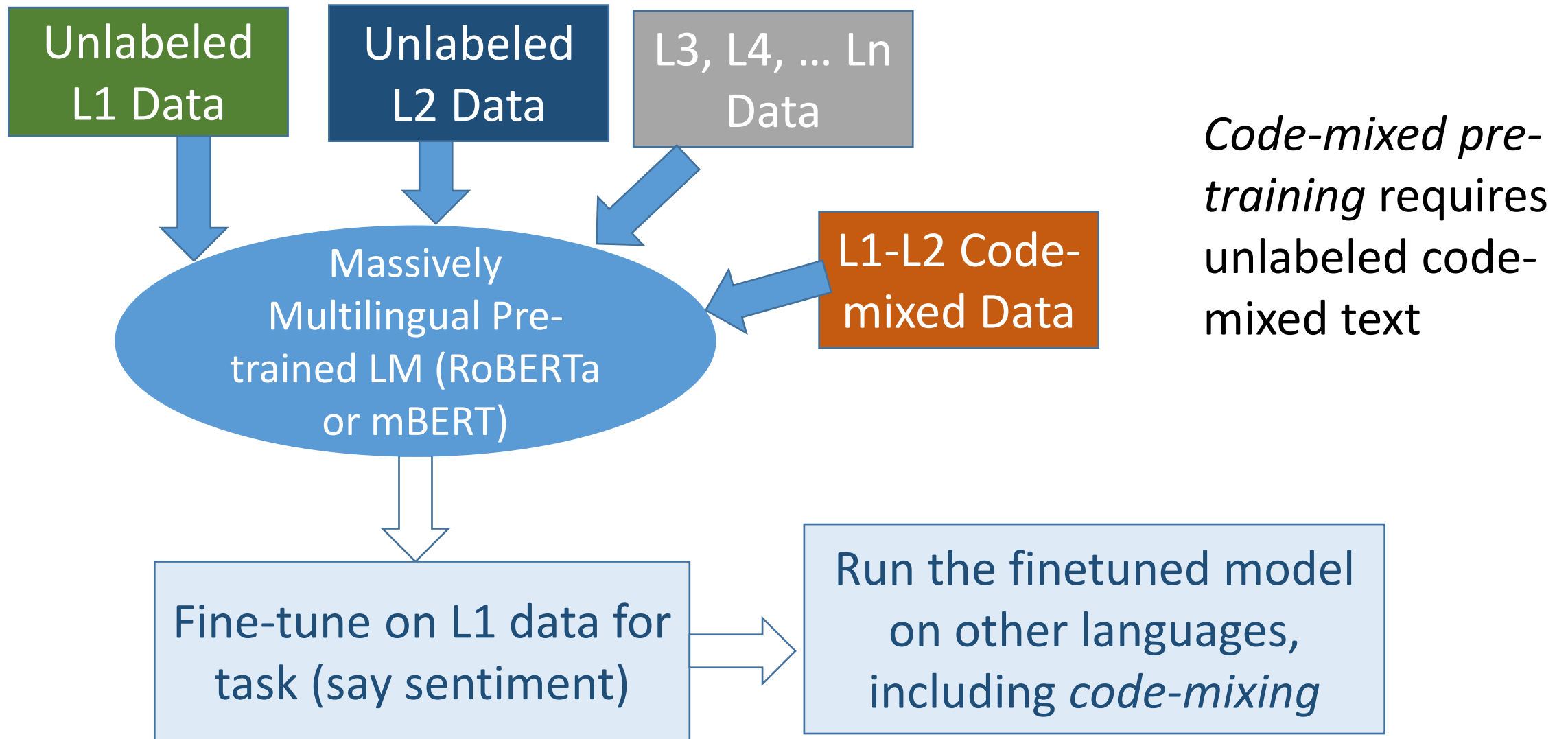
Vyas et al. 2014. En-Hi POS Tagging



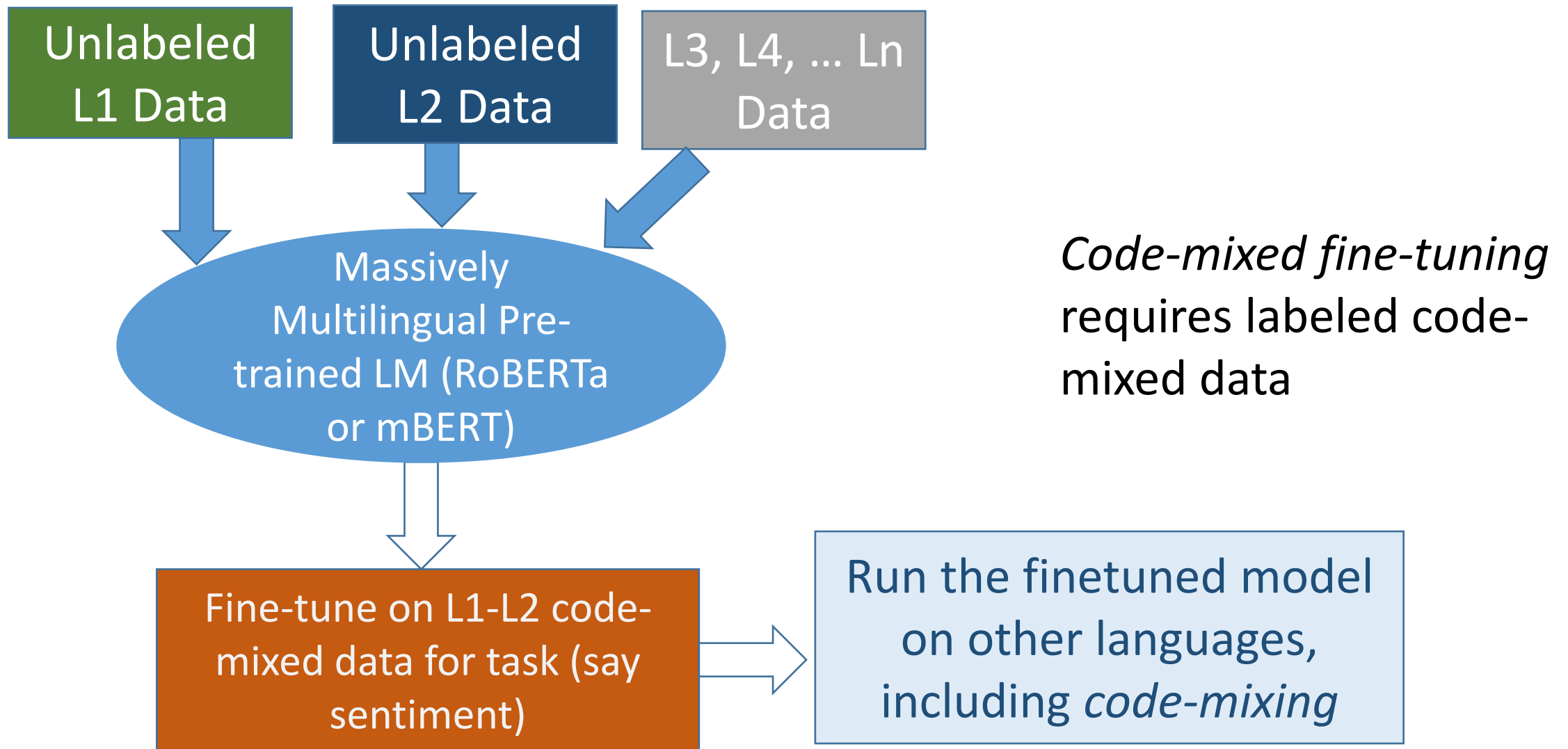
Massively Multilingual Zeroshot Transfer



Massively Multilingual Zeroshot Transfer



Massively Multilingual Zeroshot Transfer



Evaluation Test-benches

GLUECoS

[GitHub - microsoft/GLUECoS: A benchmark for code-switched NLP, ACL 2020](#)

LinCE

[LinCE Benchmark \(uh.edu\)](#)

Resources

- Sitaram et al. (2019) A Survey of Code-switched Speech and Language Processing. Arxiv.
<https://arxiv.org/abs/1904.00784>
- <https://github.com/gentaiscool/code-switching-papers>
- Project Melange: <https://www.microsoft.com/en-us/research/project/melange>
- [EMNLP 2019 Tutorial by Monojit Choudhury et al.](#) [slides]
- [EMNLP-IJCNLP2019: Tutorial \[T2\] Processing and Understanding Mixed Language Data \(Part 1/2\) on Vimeo](#) [video]



© Maureen Nchekwube Nwachi

Understanding code-mixing is not a luxury but a necessity for building NLP systems for multilingual societies.

monojitc@microsoft.com