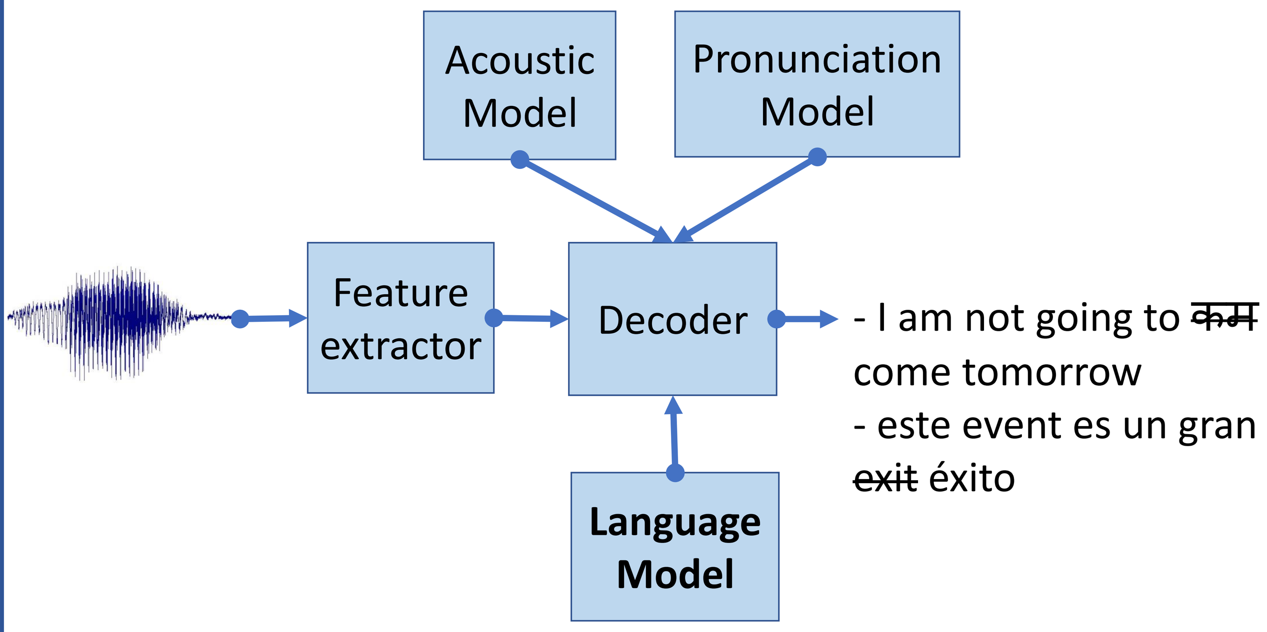
Language Modeling for Code-Mixing: The Role of Linguistic Theory based Synthetic Data

Adithya Pratapa, Gayatri Bhat, Monojit Choudhury, Sunayana Sitaram, Sandipan Dandapat, Kalika Bali

1. Introduction

- Code-Mixing (CM) refers to juxtaposition of linguistic units from two or more languages in a single conversation/utterance
- Language model (LM) has applications in ASR, Machine Translation



Challenges in modeling CM language,

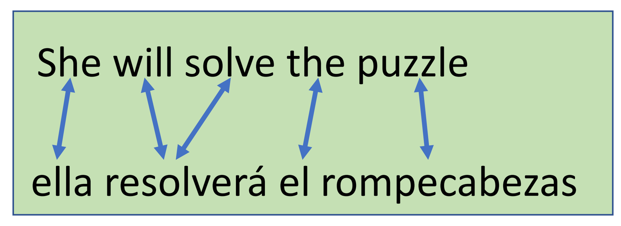
- CM is rare in formal text
- Even in the available CM data, switch points are few (~10%)

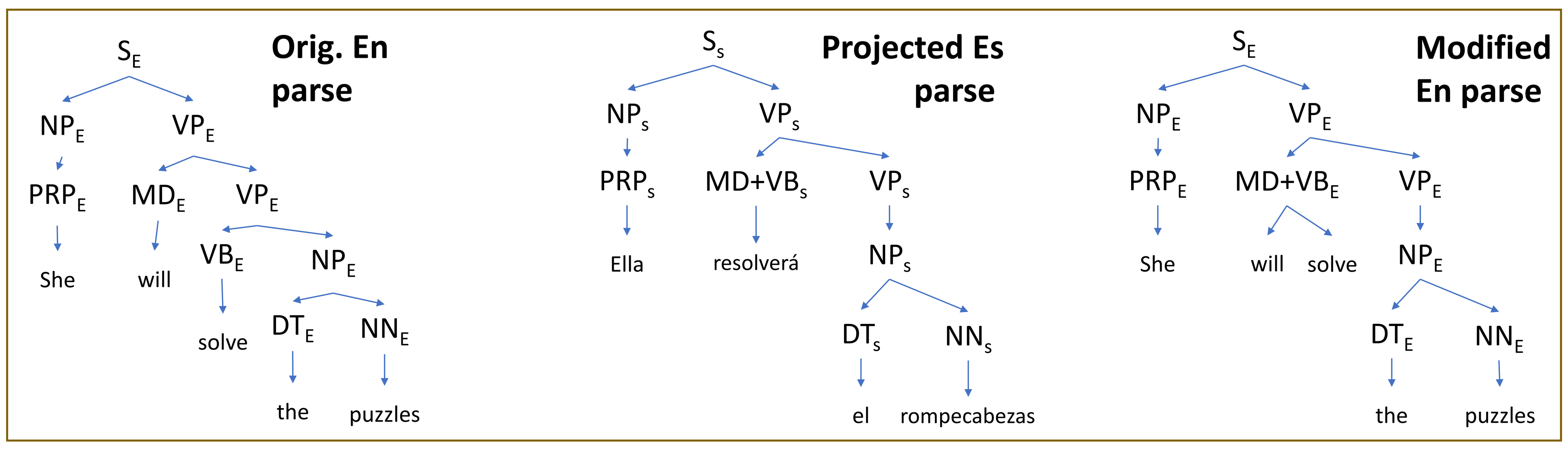
Can we leverage the readily available monolingual data?

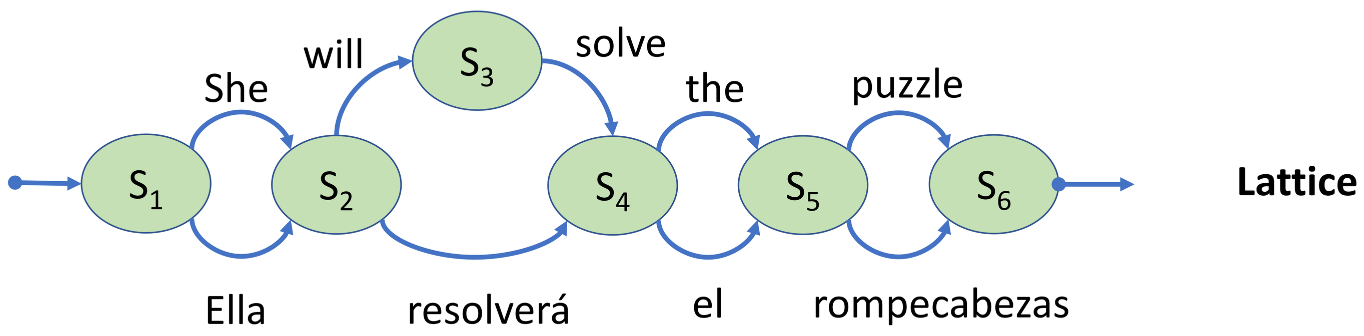
2. Linguistic Models of Code-Mixing

- *Equivalence Constraint Theory*: (*Poplack*, 1980; *Sankoff*, 1998)
  - Any monolingual fragment in CM sentence must occur in one of the monolingual sentences
  - CM sentence shouldn't deviate from both monolingual grammars
  - The two grammars must be equivalent at switch points

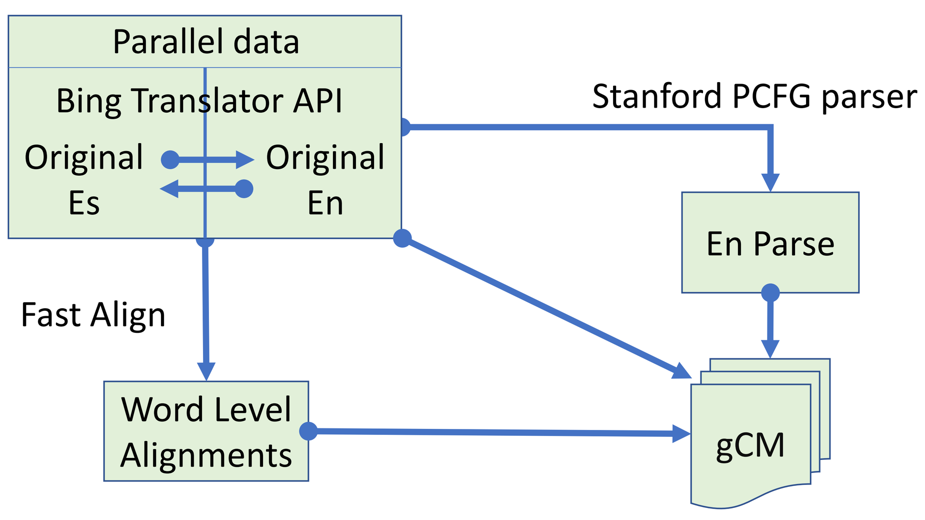
Parallel sentences with alignments







3. Generating Synthetic Data

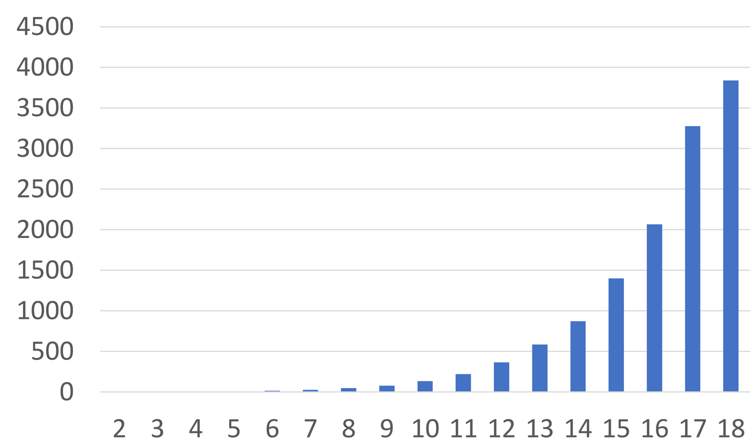


- Used Pseudo Fuzzy-match score to threshold the quality of translations
- En-parse is projected onto the Es sentence using word-level alignments
- rCM train, validation and test-17 (Rijhwani et al. 2017), test-14 (Solorio et al. 2014)

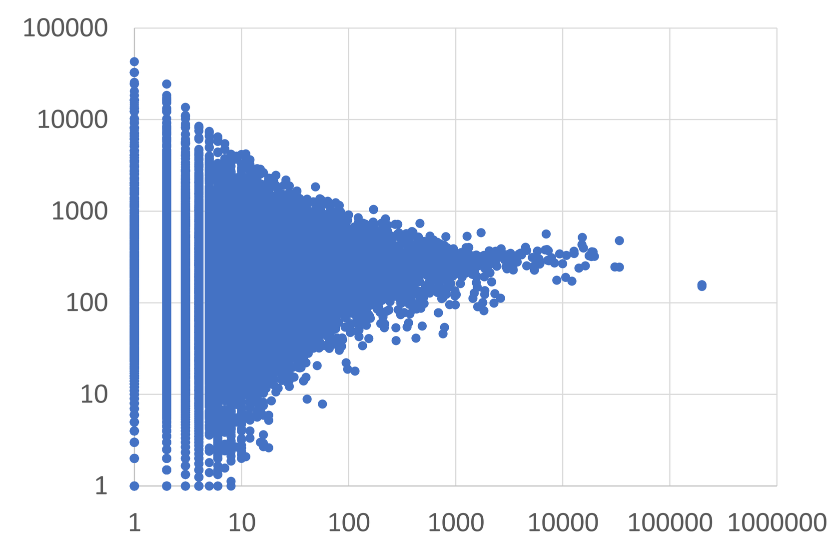
Dataset	# Tweets	# Words	CMI	SPF
English	100K	850K	0	0
Spanish	100K	860K	0	0
Train	100K	1.4M	0.31	0.105
Validation	100K	1.4M	0.31	0.106
Test-17	83K	1.1M	0.31	0.104
Test-14	13K	138K	0.12	0.06
gCM	31M	463M	0.75	0.35

4. Sampling gCM data

- A pair of monolingual sentences can give rise to a large (exponential) number of CM sentences, but only a few are observed in real data
- Even the statistical properties of this gCM data are different from real CM data



\# gCM sentences vs avg. input sentence length

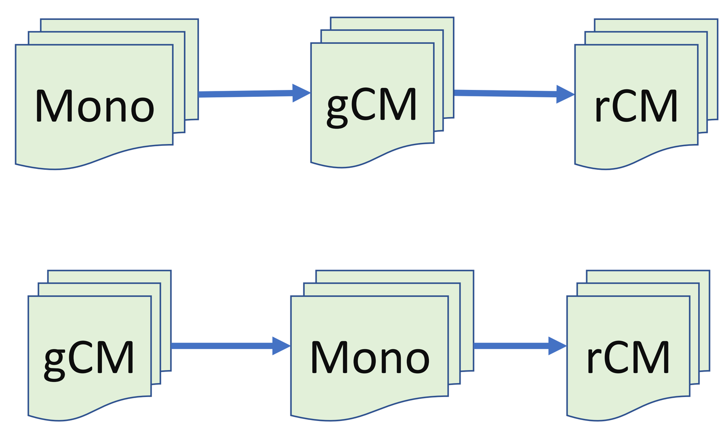


fractional increase in word freq. in gCM vs Original freq. of unigrams.

- Hence the need for sampling, based on,
  - Random (***χ-gCM***)
  - Code mixing index (CMI) (**↑*-gCM*** & **↓*-gCM***)
  - Switch point fraction (SPF) (***ρ-gCM***)

5. Training Curricula

- LM can be trained sequentially on different orderings of Mono, gCM and rCM resulting in various training curricula
- Real CM (rCM) data at the end of training is found to be most effective (*Baheti et al.* 2017)



6. Experiments and Results

Expt. ID	Training Curricula			Overall PPL		Avg. SP PPL	
				Test-17	Test-14	Test-17	Test-14
1	rCM			2018	1822	5670	8864
2	Mono			1607	892	23790	26901
3	Mono	rCM		1041	861	4824	7913
4(a)	**Mono**	**gCM**					
4(a)-χ	Mono	χ-gCM		1771	1119	5869	6065
4(a)-↑	Mono	↑-gCM		1872	1208	9167	8803
4(a)-ρ	Mono	ρ-gCM		1618	1116	6618	7293
4(b)	**gCM**	**Mono**					
4(b)-χ	χ-gCM	Mono		1680	903	21028	20300
4(b)-↓	↓-gCM	Mono		1917	973	28722	25006
4(b)-ρ	ρ-gCM	Mono		1641	871	26710	22557
5(a)	**Mono**	**gCM**	**rCM**				
5(a)-χ	Mono	χ-gCM	rCM	1038	836	4386	5958
5(a)-↑	Mono	↑-gCM	rCM	1058	961	5078	6861
5(a)-ρ	Mono	ρ-gCM	rCM	1011	830	4829	6807
5(b)	**gCM**	**Mono**	**rCM**				
5(b)-χ	χ-gCM	Mono	rCM	1019	790	4987	7018
5(b)-↓	↓-gCM	Mono	rCM	1025	800	5489	7476
5(b)-ρ	ρ-gCM	Mono	rCM	986	772	4912	6547

Rows 1–3: **Baselines**

Row 5(b)-ρ: **Best Model**

- Effect of rCM size**:**
  - As expected, PPL drops with increasing amount of rCM data
  - gCM data still helps even though diminishingly
  - In general, the baseline (Model 3) needs *twice as much amount of rCM data* to perform as good as our Model 5(b)-ρ
- Even though gCM helps, rCM data is indispensable
- SPF based sampling performs the best
- PPL at SPs is much higher than overall PPL, showing the inherent complexity of modeling CM language
- Modeling shorter run lengths is found to be challenging

# rCM / Expt.	0.5K	1K	2.5K	5K	10K	50K
3	1238	1186	1120	1041	991	812
5(b)-ρ	1181	1141	1068	986	951	808

References:

Ashutosh Baheti, Sunayana Sitaram, Monojit Choudhury, and Kalika Bali. 2017. Curriculum design for code-switching: Experiments with language identification and language modeling with deep neural networks. In Proc. of ICON-2017, Kolkata, India, pages 65–74.

Shana Poplack. 1980. Sometimes Ill start a sentence in Spanish y termino en espaol. Linguistics, 18:581– 618.

Shruti Rijhwani, R Sequiera, M Choudhury, K Bali, and C S Maddila. 2017. Estimating code-switching on Twitter with a novel generalized word-level language identification technique. In ACL.

David Sankoff. 1998. A formal production-based explanation of the facts of code-switching. Bilingualism: language and cognition, 1(01):39–50.

Thamar Solorio et al. 2014. Overview for the first shared task on language identification in codeswitched data. In 1st Workshop on Computational Approaches to Code Switching, EMNLP, pages 62– 72.

Microsoft

Contact: monojitc@microsoft.com