# Code-switched Language Models Using Dual RNNs and Same-Source Pretraining

**Saurabh Garg*, Tanmay Parekh*, Preethi Jyothi**

Department of Computer Science and Engineering, Indian Institute of Technology Bombay

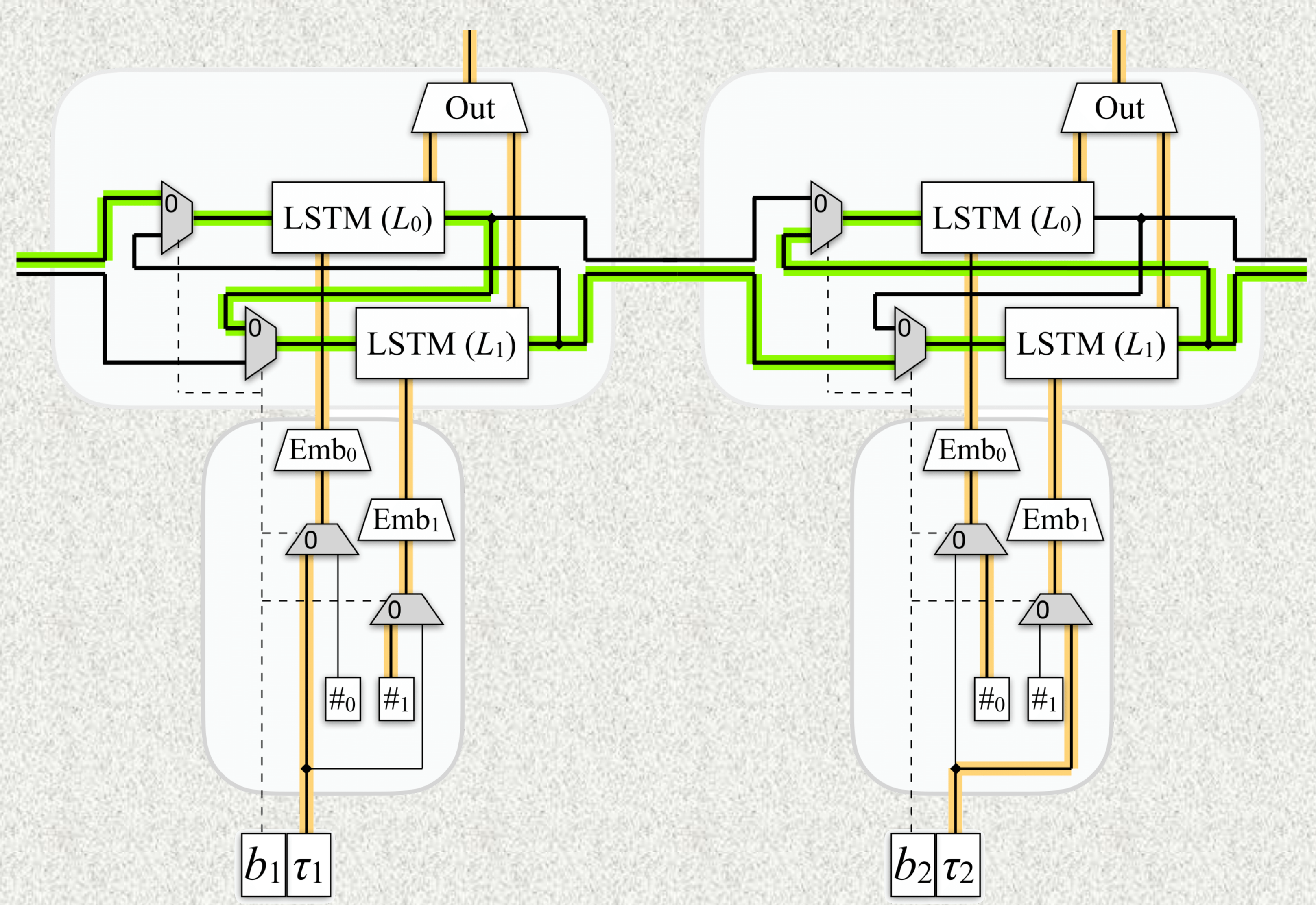
* Joint first authors

## INTRODUCTION

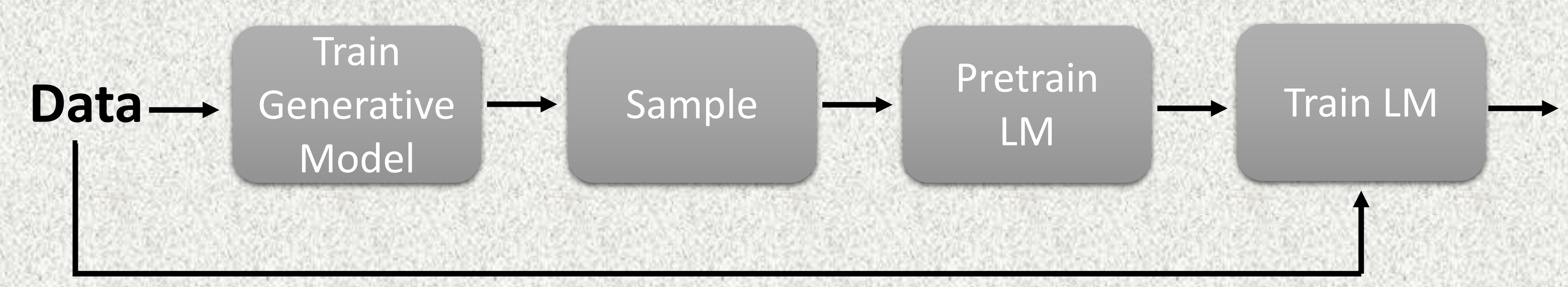
- ❑ Code-switching is a common phenomenon in multilingual societies: Switching between multiple languages within a single utterance.
- ❑ Limited availability of data poses challenges for building language models (LMs) for code-switched text.
- ❑ **Objective**: To build better LMs to handle code-switching, exploiting monolingual text data.

## APPROACH

- ❑ **Dual RNN language models:** Two LSTM cells operating in an upstream-downstream fashion to handle two different languages.
- ❑ Cross-lingual context captured by passing hidden state between these cells.




- ❑ **Same-Source Pretraining:**




- ❑ Generative model that worked best: SeqGAN [1]
  - Uses a reward function determined by a discriminator

## EXPERIMENTAL DATA

- ❑ SEAME corpus [2] of Mandarin-English text

| | Train | Dev | Test |
|---|---|---|---|
| # Utterances | 74,927 | 9,301 | 9,552 |
| # Tokens | 977,751 | 131,230 | 114,546 |
| # English Tokens | 316,726 | 30,154 | 50,537 |
| # Mandarin Tokens | 661,025 | 101,076 | 64,009 |

- ❑ Monolingual text for pre-training
- ❑ Syntactic features as additional input (POS Tags, Brown word clusters and language ID)

## EXPERIMENTAL RESULTS

Perplexities without syntactic features

| | w/o mono data | | with mono data | |
|---|---|---|---|---|
| | Dev | Test | Dev | Test |
| RNNLM | 89.60 | 74.87 | 74.06 | 61.66 |
| D-RNNLM | 88.68 | 72.29 | 72.41 | 60.73 |
| With RNNLM SeqGAN | 79.16 | 65.96 | 72.51 | 60.56 |
| With D-RNNLM SeqGAN | **78.63** | **65.41** | **72.33** | **60.30** |

Perplexities with syntactic features

| | w/o mono data | | with mono data | |
|---|---|---|---|---|
| | Dev | Test | Dev | Test |
| RNNLM | 81.87 | 68.23 | 71.04 | 59.00 |
| D-RNNLM | 81.01 | 66.26 | 70.83 | 59.04 |
| With RNNLM SeqGAN | 77.30 | 63.75 | 68.43 | 55.71 |
| With D-RNNLM SeqGAN | **77.19** | **63.63** | **67.79** | **55.60** |

Decomposed Perplexities

| | Eng-Eng | Eng-Man | Man-Eng | Man-Man |
|---|---|---|---|---|
| RNNLM | 133.18 | 157.18 | 2617.28 | 34.98 |
| D-RNNLM | 140.37 | 151.38 | 2452.16 | 32.89 |
| Mono RNNLM | 101.61 | 181.28 | 2510.48 | 30.00 |
| Mono D-RNNLM | 101.66 | 156.44 | 2442.81 | 29.64 |
| RNNLM SeqGAN | 120.28 | 154.44 | 2739.85 | 30.40 |
| D-RNNLM SeqGAN | 120.26 | 149.68 | 2450.85 | 30.60 |

Percentage increase in unique n-grams using SeqGAN models

| | RNNLM SeqGAN | D-RNNLM SeqGAN |
|---|---|---|
| Bigram | 25.57 | 31.33 |
| Trigram | 75.88 | 83.86 |
| Quadgram | 137.98 | 145.71 |

## REFERENCES

[1] Yu, Lantao, et al. "SeqGAN: Sequence Generative Adversarial Nets with Policy Gradient." *AAAI* 2017

[2] Lyu, Dau-Cheng, et al. "An analysis of a Mandarin-English code-switching speech corpus: SEAME." *Age 2010*

## FUTURE WORK

- ❑ Using same-source pretraining beyond code-switching
- ❑ Dual RNN LMs for speaker diarization