Evaluating and Developing Methods of Generating Code-Switched Text

Oral Project Proposal, 07-300

Abhishek Vijayakumar

Mentors/Related Researchers at CMU

Alan Black

Khyathi Chandu

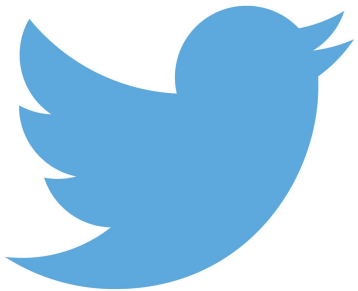
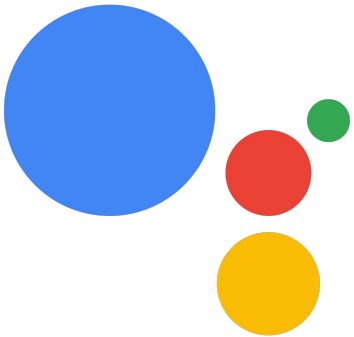
Code-switching is increasingly common in informal digital communication.

ES:	Vi que	tu mirada	ya estaba	llamándome
ES-EN:	Vi que	tu look	ya estaba	calling me
EN:	I saw that	your look	was already	calling me

Code-switching is increasingly common in informal digital communication.

ES:	Vi que	tu mirada	ya estaba	llamándome
ES-EN:	Vi que	your look	ya estaba	calling me
EN:	I saw that	your look	was already	calling me

Example Domains


Training state-of-the-art models for code-switching requires a high volume of code-switched (CS) data.

- SOTA models are neural, requiring large amounts of training data
- Real CS data is harder to collect than monolingual data, and is often noisy

Training state-of-the-art models for code-switching requires a high volume of code-switched (CS) data.

- SOTA models are neural, requiring large amounts of training data
- Real CS data is harder to collect than monolingual data, and is often noisy

Solution: Generate synthetic CS data.

Generating high-quality synthetic CS text will improve the performance of language technologies in an unexplored domain of information and communication.

What defines quality?

- Similarity to real CS text
- **Usefulness in model training**

The first part of the project is the evaluation of existing methods of generating CS text.

Types of Methods

- Non-neural (substitutive) methods
- Neural (generative) methods

Types of Resources

- Monolingual data $
- Translation lexicon $$
- Bilingual parallel data $$
- Translation engine $$$
- Real CS data $$$$$

The second part of the project is improving upon a generative technique.

Neural Method

- Reduce resource cost by reducing
  - **Quantity** of real code-switched data required
  - **Quality** of real code-switched data required
    - e.g. blog posts vs tweets

References

[1] Gayatri Bhat, Monojit Choudhury, and Kalika Bali. Grammatical constraints on intra-sentential code-switching: From theories to working models, 2016.
[2] Simran Khanuja, Sandipan Dandapat, Anirudh Srinivasan, Sunayana Sitaram, and Monojit Choudhury. Gluecos: An evaluation benchmark for code-switched NLP. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, Proceedings of the 58th Annual 14 Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, pages 3575–3585. Association for Computational Linguistics, 2020.
[3] Bidisha Samanta, Sharmila Reddy, Hussain Jagirdar, Niloy Ganguly, and Soumen Chakrabarti. A deep generative model for code switched text. In Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19, pages 5175–5181. International Joint Conferences on Artificial Intelligence Organization, 7 2019.
[4] Ishan Tarunesh, Syamantak Kumar, and Preethi Jyothi. From machine translation to code-switching: Generating high-quality code-switched text. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 3154–3169, Online, August 2021. Association for Computational Linguistics.
[5] https://www.canstockphoto.com/illustration/texting.html
[6] https://static01.nyt.com/images/2014/08/10/magazine/10wmt/10wmt-superJumbo-v4.jpg
[7] https://upload.wikimedia.org/wikipedia/commons/thumb/c/cb/Google_Assistant_logo.svg/1024px-Google_Assistant_logo.svg.png
[8] https://static.wikia.nocookie.net/logopedia/images/5/54/Sphiri_Normal.png/revision/latest/scale-to-width-down/250?cb=20200731195552