Code-Mixing on Sesame Street: Dawn of the Adversarial Polyglots

Samson Tan[§♮] **Shafiq Joty**[§‡]
[§]Salesforce Research Asia [♮]National University of Singapore [‡]Nanyang Technological University
{samson.tan,sjoty}@salesforce.com

Abstract

Multilingual models have demonstrated impressive cross-lingual transfer performance. However, test sets like XNLI are monolingual at the example level. In multilingual communities, it is common for polyglots to code-mix when conversing with each other. Inspired by this phenomenon, we present two strong black-box adversarial attacks (one word-level, one phrase-level) for multilingual models that push their ability to handle code-mixed sentences to the limit. The former (POLYGLOSS) uses bilingual dictionaries to propose perturbations and translations of the clean example for sense disambiguation. The latter (BUMBLEBEE) directly aligns the clean example with its translations before extracting phrases as perturbations. BUMBLEBEE has a success rate of 89.75% against XLM-R$_{large}$, bringing its average accuracy of 79.85 down to 8.18 on XNLI. Finally, we propose an efficient adversarial training scheme, Code-mixed Adversarial Training (CAT), that trains in the same number of steps as the original model. Even after controlling for the extra training data introduced, CAT improves model accuracy when the model is prevented from relying on lexical overlaps (+3.45), with a negligible drop (-0.15 points) in performance on the original XNLI test set. t-SNE visualizations reveal that CAT improves a model's language agnosticity.

Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching, page 141
June 11, 2021. ©2021 Association for Computational Linguistics
https://doi.org/10.26615/978-954-452-056-4_019