Simple TICO-19: A Dataset for Joint Translation and Simplification of COVID-19 Texts
Matthew Shardlow, Fernando Alva-Manchego • Manchester Metropolitan University, Cardiff University

Abstract
- We introduce Simple TICO-19, a new language resource containing manual simplifications of the English and Spanish portions of the TICO-19 corpus.
- Text simplification is the process of automatically reducing the complexity of a text through editing the vocabulary and style.
- We describe the annotation process used to develop our corpus, which entailed designing an annotation manual and employing four annotators.
- Each annotator simplified over 6,000 sentences from the English and Spanish portions of the TICO-19 corpus.
- We report statistics on the new dataset and propose baseline methodologies for automatically generating the joint translation and simplifications contained in our dataset.

Background
TICO-19 was developed to allow researchers to develop machine translation systems for multiple languages in the context of the Covid-19 pandemic [1]. We propose to use automated simplification [2] to also simplify these texts, possibly whilst jointly translating. Prior work has shown it is possible to control the readability of translated outputs [3].

Dataset Collection
- We ran the baseline model on the original texts from TICO-19 and evaluated these results against (a) the original references and (b) our simplified references. The results show higher BLEU and BERT scores when evaluating against the original references. This indicates that complex language in the original texts is also present in the translations.

Model
We use the Marian-NMT implementation found in the SimpleTransformers library. We used models for EN-ES and ES-EN translation from Opus-MT: We set the model parameters to mirror the performance of the original baselines in the Tico-19 paper, setting a beam size of 12 and a maximum output length of 200 tokens. The code and data used to produce our results is in the GitHub repository at the link below.

References

Contact:
m.shardlow@mmu.ac.uk
AlvaManchegoF@cardiff.ac.uk
github.com/MMU-TDMLab/SimpleTICO19