

Simple TICO-19: A Dataset for Joint Translation and Simplification of COVID-19 Texts

Matthew Shardlow, Fernando Alva-Manchego • Manchester Metropolitan University, Cardiff University

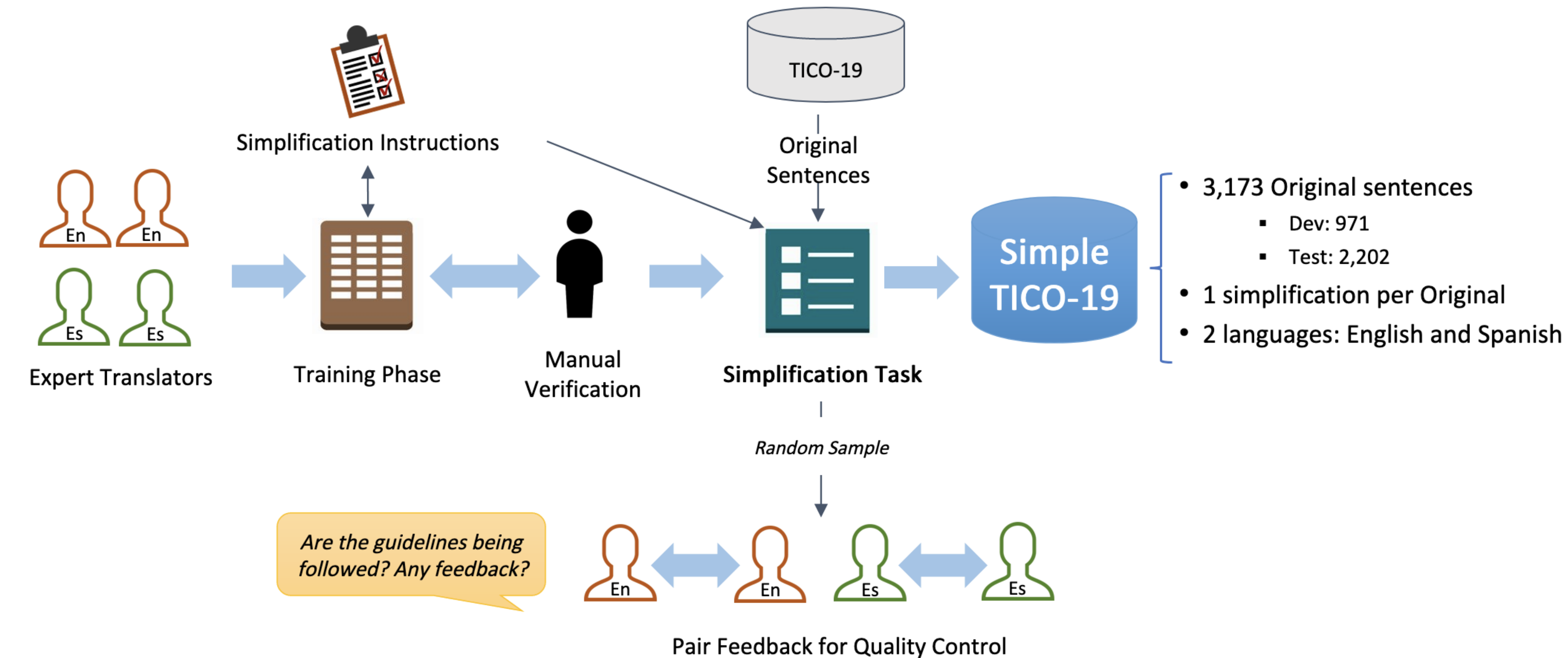
Abstract

- 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

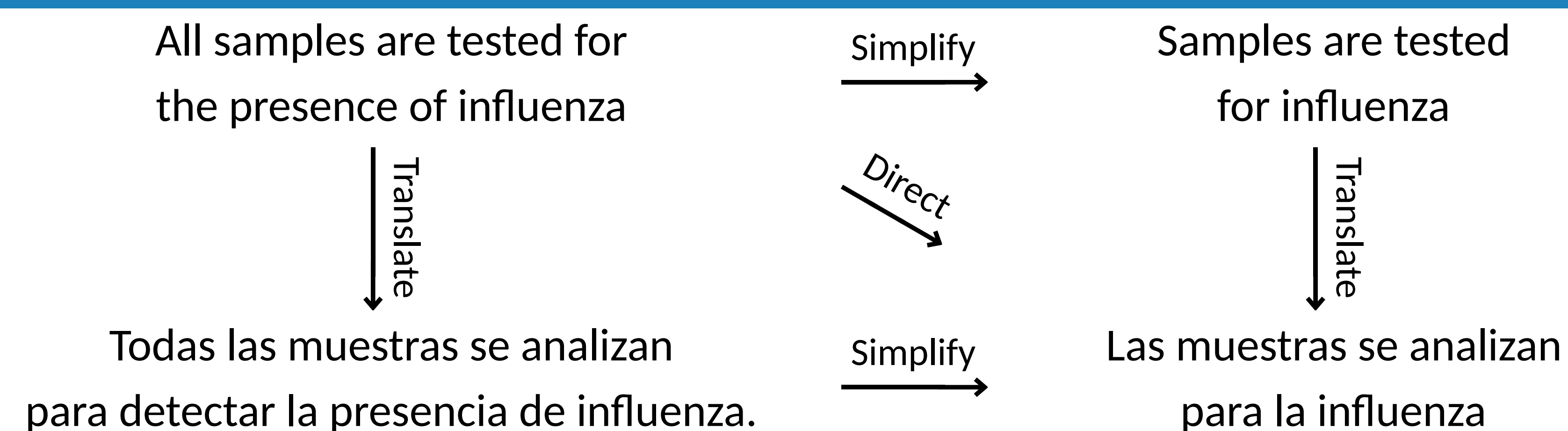


Baseline Results

Data Source	Original-En →				Original-Es →			
	Original-Es	Simplified-Es	Original-En	Simplified-En	Original-En	Simplified-En	Original-Es	Simplified-Es
CMU	33.51	0.678	17.05	0.581	30.82	0.734	29.41	0.683
PubMed	51.63	0.819	42.69	0.757	50.22	0.802	38.32	0.723
Wikinews	55.41	0.826	40.22	0.732	44.43	0.807	32.09	0.717
Wikipedia	52.16	0.875	44.83	0.836	48.91	0.878	38.64	0.833
Wikisource	39.98	0.715	31.85	0.647	42.16	0.775	31.70	0.702
All	51.42	0.841	43.15	0.788	48.66	0.842	38.02	0.783

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.

Routes to joint Translation and Simplification



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

- [1] A. Anastasopoulos, A. Cattelán, Z.-Y. Dou, M. Federico, C. Federmann, D. Genzel, F. Guzmán, J. Hu, M. Hughes, P. Koehn, R. Lazar, W. Lewis, G. Neubig, M. Niu, A. Öktem, E. Paquin, G. Tang, and S. Tur. TICO-19: the translation initiative for COVID-19. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online, December 2020. ACL.
- [2] F. Alva-Manchego, C. Scarton, and L. Specia. Data-driven sentence simplification: Survey and benchmark. *Computational Linguistics*, 46(1):135–187, 2020.
- [3] K. Marchisio, J. Guo, C.-I. Lai, and P. Koehn. Controlling the reading level of machine translation output. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 193–203, 2019.

Contact:

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

