Towards Readability-Controlled Machine Translation of COVID-19 Texts

Fernando Alva-Manchego
Cardiff University
alvamanchegof@cardiff.ac.uk

Matthew Shardlow
Manchester Metropolitan University
m.shardlow@mmu.ac.uk

Abstract

This project investigates the capabilities of machine translation (MT) models for generating translations at varying levels of readability, focusing on texts about COVID-19. Funded by the European Association for Machine Translation and by the Centre for Advanced Computational Sciences at Manchester Metropolitan University, we collected manual simplifications for English and Spanish texts in the TICO-19 dataset, and assessed the performance of neural MT models in this new benchmark. Future work will implement models that jointly translate and simplify, and develop suitable evaluation metrics.

1 Introduction

"Multilingual Translation with Readability-Controlled Output Generation" is a project that received funding from the European Association for Machine Translation (under its programme "2021 Sponsorship of Activities") and from the Centre for Advanced Computational Sciences at Manchester Metropolitan University. We aim to develop machine translation (MT) models that generate translations that can be understood by non-expert readers, focusing on texts with medical information. This is pertinent in the context of the COVID-19 pandemic, where there is a disparity in the availability of health-related content produced in English, compared to other languages.

The project has the following objectives: (1) to collect a dataset with simplified versions of parallel texts in English and Spanish about COVID-19; (2) to assess how well existing state-of-the-art MT models perform on our new benchmark; and (3) to investigate additional model architectures and/or resources that are needed to generate and evaluate simplified in-domain translations.

Lang.	Complexity	W/S	Sy/W	FRE↑	S-P↑
English	Original	23.015	6.444	45.69	–
	Simplified	21.838	6.308	52.70	–
Spanish	Original	27.623	6.287	–	75.17
	Simplified	24.749	6.271	–	79.49

Table 1: Statistics of Simple TICO-19: average number of words per sentence (W/S), average number of syllables per word (Sy/W), and estimated readability with Flesch Reading Ease (FRE) for English and Szigriszt-Pazos (S-P) for Spanish.

The first two goals of the project were carried out from January 2021 to December 2021, and resulted in the release of the Simple TICO-19 dataset (Shardlow and Alva-Manchego, 2022).[1] We continue to work with the new dataset to further investigate the nature of readability-controlled output generation in the MT context. We hope to apply for further funding at a national and European level as a result of this work.

2 The Simple TICO-19 Dataset

We leveraged the TICO-19 benchmark (Anastasopoulos et al., 2020), which contains 3,000 sentences related to the COVID-19 pandemic, translated from English into 36 languages and from several sources (e.g. academic publications, speech corpora, news articles, etc.). For our project, we collected manual simplifications for the English and Spanish subsets, resulting in the Simple TICO-19 dataset, where each sentence has either a simplified version of itself, or a decision has been taken that the sentence is already sufficiently simple. Table 1 shows some high level statistics of the resulting corpus, including readability indices such as Flesch Reading Ease (FRE) (Flesch, 1948) for

[1] https://github.com/MMU-TDMLab/SimpleTICO19

English, and Szigriszt-Pazos (S-P) (Szigriszt Pazos, 2001) for Spanish. These indices, in particular, showcase the improvements in readability from the original sentences in the dataset to their simplified versions, for both languages.

3 Machine Translation Baselines

To obtain baseline results, we leveraged models pre-trained on `opus-mt-en-es` with MarianMT as architecture. Table 2 reports BLEU (Papineni et al., 2002) and BERTScore (Zhang et al., 2020) as evaluation metrics on all the test set and per data source therein, considering original–en as source and two targets: original–es and simplified–es. The highest scores are obtained when original–es is the target, showing that standard neural MT models cannot generate simplified texts by default. Also, performance varies depending on the data source, indicating the effect of the style of text.

	orig–en → orig–es		orig–en → simp–es	
Data Source	BLEU	BERTScore	BLEU	BERTScore
CMU	33.51	0.678	17.05	0.581
PubMed	51.63	0.819	42.69	0.757
Wikinews	55.41	0.826	40.22	0.732
Wikipedia	52.16	0.875	44.83	0.836
Wikisource	39.98	0.715	31.85	0.647
All	51.42	0.841	43.15	0.788

Table 2: Results per data source of our baseline models on the test set of Simple TICO-19.

4 Future Work

Translation and Simplification. In order to incorporate simplification capabilities into MT models, we will first experiment with pipeline systems that translate and then simplify (and vice-versa) leveraging state-of-the-art models for each task. We will then work on models that perform both tasks jointly, exploring multi-task architectures.

Controllable Translation. We will study how to train models that generate outputs at diverse readability levels. We will explore varying the proportion of translation and simplification training instances to control the readability of the translations. We will rank target-side simple sentences according to the proportion of complex words and syntactic complexity, and use this ranked list to create different readability levels that allow training models for multiple degrees of complexity.

Evaluation. We will develop novel metrics suitable for the joint translation and simplification task, specifically for the medical domain. For instance, we will combine traditional similarity-based metrics, such as BLEU and BERTScore, with readability indices. While the latter are more suitable for analysing documents, we plan to adapt them for sentence-level assessment using complex word identification approaches and heuristics. We will then measure the correlation of our new metrics with human judgements on adequacy and simplicity of automatic translations.

Acknowledgements

This project was funded by the European Association for Machine Translation (EAMT) under its programme "2021 Sponsorship of Activities", and by the Centre for Advanced Computational Sciences at Manchester Metropolitan University.

References

Anastasopoulos, Antonios, Alessandro Cattelan, Zi-Yi Dou, Marcello Federico, Christian Federmann, Dmitriy Genzel, Franscisco Guzmán, Junjie Hu, Macduff Hughes, Philipp Koehn, Rosie Lazar, Will Lewis, Graham Neubig, Mengmeng Niu, Alp Öktem, Eric Paquin, Grace Tang, and Sylwia Tur. 2020. 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. Association for Computational Linguistics.

Flesch, Rudolph. 1948. A new readability yardstick. *Journal of Applied Psychology*, 32(3):221.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July. ACL.

Shardlow, Matthew and Fernando Alva-Manchego. 2022. Simple TICO-19: A dataset for joint translation and simplification of covid-19 texts. In *Proceedings of the 13th Language Resources and Evaluation Conference*, Marseille, France, June. European Language Resources Association.

Szigriszt Pazos, Francisco. 2001. *Sistemas predictivos de legilibilidad del mensaje escrito: fórmula de perspicuidad*. Universidad Complutense de Madrid, Servicio de Publicaciones.

Zhang, Tianyi, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.