Project Proposal

CSCI-GA 2590. Natural Language Processing (Spring 2023)

Deepti Saravanan (ds6812)

Yuxuan Xia (yx2432)

# Text Simplification: Leveraging QA-based lexical simplification and Structural Rule-based Controlled Text Generation

We propose an end-to-end framework for domain-independent sentence-level text simplification. The main objective is to combine text pre-processing for lexical simplification and controlled text generation[1] techniques for structural simplification while also preserving the semantics of the sentence. We plan to use open-domain datasets such as WikiSmall and WikiLarge. For the text pre-processing component, we propose a new approach to lexical simplification not tried in past literature in this context. This was inspired by [5]. Our technique involves question-answering-based POS tagging of tokens which would be our identified tokens of high importance, where we frame questions from different perspectives covering the who, when, and how aspects. Framing the right questions is a challenging task. These identified important tokens would then be masked at random up to a threshold percentage and fed into Large Language Models (LLMs) for filling in using the surrounding context. We expect this approach of pre-processing to be a feasible alternative to techniques used in identifying and replacing complex terms in past literature [4]. The output from this text preprocessing component would be fed into the controlled text generation component for structural simplification. We plan to start with a sequence-to-sequence model unless we land on a better approach later during the development of the framework. The idea is similar to the machine translation technique, which is widely used in previous works[2, 3]. But in our case, we input a complex English sentence along with some rules on structural characteristics such as the relative placement of different POS tags. The output would be a simpler English sentence preserving the overall meaning of the original sentence. The main challenge here is to research sentence structures to come up with these rules.

The proposed framework would be evaluated from multiple aspects. First, we evaluate the simplicity and readability of the generated sentence using Flesch-Kincaid Grade Level and SARI[6]. The second aspect is to test how similar the generated simplified text is to the original input sentence. For this, we plan to employ either statistical similarity testing or LLM-based measure depending on the text generation approach employed. Finally, the applicability of the solution would be tested in both domain-independent and domain-specific environments such as sentences from Legal Case Files and Medical Articles.

**References**

[1] Martin, L., Sagot, B., de la Clergerie, É., & Bordes, A. (2020). Controllable Sentence Simplification.

[2] Vo, V., Wang, W., & Buntine, W. (2022). Unsupervised Sentence Simplification via Dependency Parsing.

[3] Lu, X., Qiang, J., Li, Y., Yuan, Y., & Zhu, Y. (2021). An Unsupervised Method for Building Sentence Simplification Corpora in Multiple Languages.

[4] Cemri, M., Çukur, T., & Koç, A. (2022). Unsupervised Simplification of Legal Texts.

[5] Mehta, S., Rangwala, H., & Ramakrishnan, N. (2022). Improving Zero-Shot Event Extraction via Sentence Simplification.

[6] Xu, W., Napoles, C., Pavlick, E., Chen, Q., & Callison-Burch, C. (2016). Optimizing Statistical Machine Translation for Text Simplification. Transactions of the Association for Computational Linguistics, 4, 401–415. <https://doi.org/10.1162/tacl_a_00107>