Clear Language Generation: An initial approach to the Spanish Language Simplification Task

Anonymous, Instituto Tecnológico Autónomo de México.

***Index Terms*—Text Simplification, Clear Language, Natural Language Processing, Lexical Simplification, Machine Translation, Generative AI, Large Language Models, Parameter-Efficient Fine-tuning, Quantization.**

I. Introduction

Language Simplification has long been a research topic in the Natural Language Processing community[1][2][3][4][5]. The idea and motivation of simplifying language is, in our consideration, one of the more noble goals that there are. Even though the previous statement is a subjective evaluation of this topic, this opinion is backed by a quick review of most of the specialized applications of language simplification[6]. Tools for assisting autistic people[7] and for improving web accessibility[8] are the type of implementations that are at the core of this task. More technical domain applications are mostly in the medical[9] and legal[10] text domains.

Another area of possible application for language simplification is in the governmental documentation sphere. As the ISO Clear Language Guidelines[11] state, badly written text that use technical or bureaucratic jargon endangers the readers power by creating barriers. This is such an important topic for information democracy, that is, the right that every person has to understand what actions their government is taking, that most countries have their own Clear Language manuals[12].

Text Simplification (TS) is a subset of Language Simplification as it only pertains written formats. In the age of Artificial Intelligence and more recently Language Models, this task has been previously tackled with several approaches with different results as can be seen in one of the most recent surveys[5] on the topic.

TS has proven to be difficult for a number of reasons including dataset availability and model architecture development. These challenges remain and increase for more specific or even niche languages. As language variants may not be in the interest of the majority, investigations on this task are not usually included in State of the Art publications. The models and propositions made by current investigations often use a great amount of computational resources during experimentation. This also poses a computational resource barrier.

Previously mentioned issues pose areas of opportunity. There is a lack of recent investigation on the development of a structure for generating Text Simplification using Clear Language as a guideline. We may think of Clear Language as a subset of Natural Language as it is a variant with specific rules for the improvement of language understanding. It becomes an even more specific area as most Text Simplification models and investigations have been done for the English Language. There is an unpopulated space for Spanish Text Simplification.

This work implements Quantized LoRA for fine-tuning several SOTA Large Language Models. The models used are Llama-2-7B from Meta, Gemma-7B from Google. The task of text simplification is for Spanish text. Its larger aim is to train a single model on the different tasks required for text simplification: lexical simplification, syntactical simplification, stylistic simplification, and ”discourse” simplification.

As previously stated, one of the bigger issues with the generation of a solution for the language simplification task is the lack of datasets, specifically for the Spanish language. At the moment, the dataset that has been created covers Lexical Simplification in Spanish. It was generated from publicly available corpora such as the the Complex Word Identification Shared Task Corpus[13] and the EASIER Corpus[14]. Even tough our overarching goal with this line of investigation is to create a solution for language simplification that is derived form the Clear Language guidelines, the dataset created does not take into consideration the specific requirements for it. We have nonetheless created a set of rules derived from different Clear Language non governmental[11] and governmental[12] documents. They will form a base for future investigations.

Thus, our contributions can be listed as follows:

- Creation of a Dataset for Spanish Lexical Simplification based on the EASIER Corpus.
- Development of Rules derived from Clear Language Documentation. They are proposed to be used for Dataset enhancement and simplification result evaluation. This applications are not covered in this paper but will be investigated in future works.
- Exploration and evaluation of State of the Art fine-tuning methods of causal large language models in the task of lexical simplification via text generation. QLoRA is implemented on the Llama-2-7B model, as well as the Gemma-7B model. Compute is done via Google Colab with many open source libraries, making this work easily replicable.

The structure of this article will start by describing the related contemporary works. They are divided by lexical simplification articles and clear language applications for the Spanish language. Next is the Methodology, where we will describe the dataset and it’s creation, the ruleset for language simplification that was created from Clear Language documentation. The Evaluation of the proposed models will include a comparison of the different models that where used in the investigation with different metrics. We will then discuss the

performance of the proposed models in the Results Analysis. The las section will be the Conclusion.

II. Related Work

There are different models, implementations and applications that seek to simplify text at different scales. This section will be divided into a review of the Text Simplification related works, an exploration of most recent investigations for the specific task of Lexical Simplification, and a description of the articles for Spanish Text Simplification.

A. Text Simplification

A recent review of automatic text simplification tools[5] provides insight regarding tool approaches to text simplification and the sub-tasks that compose this endeavour.

Text simplification can be divided by the following language phenomena:

- Lexical simplification: Identification of complex words, interchanging it with synonyms or simpler definitions appropriate for the context of the sentence.
- Syntactic Simplification: Reduction of sentence structure complexity by reducing sentence length, elimination of unnecessary words, usage of adjectives vs verbs, etc.
- Style Simplification: Textual layout transformation such as creating lists, reducing paragraph length, etc.
- "Discourse" Maintenance: Verification that no information has been lost during the previous simplification steps (pronouns and conjunctions).

The review from Espinosa-Zaragoza, Abreu-Salas, Lloret, Moreda, and Palomar [5] notes that there are no tools or models that encompass all of these different phenomena to create a single input single output process for Language Simplification from a more complex set of sentences. Most of the tools that are currently in this arena are either used for Lexical Simplification. Some articles such as CamemBERT[15] do have Discourse Maintenance, specifically for the french language.

With this review we can get a starting point for the opportunities and difficulties that arise with the language simplification task. Most of the issues come from a lack of data. Some datasets for word complexity, language summarizing and language simplification exist, but they either only work for one task or are not in Spanish, which is our current language of interest for simplification.

B. Lexical Simplification

Recent Lexical Simplifications investigations provide great insights in the possible architectures and design choises that can be made. Truic a, Stan, and Apostol[16] propose SimpLex as a lexical text simplification architecture. Their system uses either word embeddings or transformers for the replacement of words. For the word embedding based approach they use cosine similarity for synonym selection and then they compute the perplexity score of the different sentences created to select the best simplified sentence. For the transformer based approach they select synonims based on the transformer embeddings an then use cosine similarity for best sentence selection.

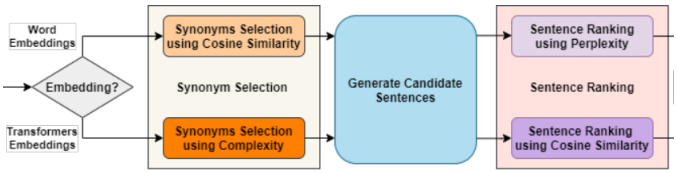

Fig. 1. SimpLex architecture from Embedding bifurcation to Sentence Ranking: From [16]

Even though they use embeddings as sources for synonym generation, they do use a dataset for complex word identification[17]. This is the firs step of their model and is heavily dependant on a large human annotated corpus. Also, the corpus only provides a complex word set, which still only works for lexical simplification.

As lack of data has been a recurrent issue for most investigations, there are some articles that deal with this problem.

Qiang, Jipeng, Zhang, Li, Yuan, Zhu, and Wu[18] propose an unsupervised statistical text simplification method that uses BERT for initialization. In this context, initialization refers to the creation of phrase tables with synonyms of high-similar words. This tables are used as datasets for the monolingual machine translation approach of addressing text simplification.

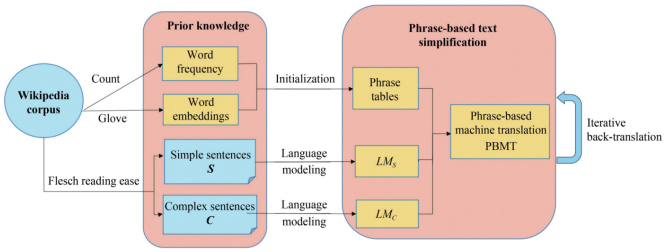

Fig. 2. Architecture of UnsupPBMT method: From [18]

The benefit of using an embedding model for the creation of datasets is the extraction of intrinsic semantic similarities that are thought to be learned by the embedding space, also known as the semantic similarity space. This approach reports improved results over previous lexical simplification methods.

The proposed architecture also uses a phrase based machine translation approach derived from an open source toolkit for statistical machine translation[19], which requires additional compute efforts.

For the production of a specific set of simplified sentences that must match the Clear Language guidelines, this approach was not used in our current investigation. Instead, we create our own set of complex-simple sentence pairs from EASIER, which is a human annotated corpus. This will be further explained in the Methodology section.

Another current approach that uses Large Language Models is presented by Tan, Keren, Luo, Lan, Yuan, and Shu[20], where they propose am LLM-enhanced adversarial editing system for lexical simplification (LEA-LS). With this approach the authors look to create a system that uses LLMs such as

ChatGPT for an instruction-based complex word identification for their removal. After, their self titled Difficulty-aware Filling module replaces the masked position with a simple word.

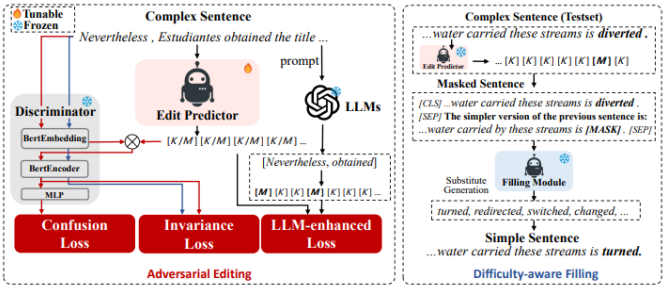

Fig. 3. LLM-enhanced Adversarial Editing System: From [20]

The dual module implementation described above serves as an example of the main dual tasks of Lexical Simplification via a Sequence-to-Sequence model. Encoder-Decoder architectures such as BERT, used in both papers, is a classic example of a model that is used for the masked word replacement task. The difference between uses of BERT is that UnsupPBMT uses its embedding for synonym generation that is consumed by a PBMT; LEA-LS uses it for the direct replacement of the words masked by the causal LLM.

Both of these approaches have several positive contributions but they still need large amounts of data to either train the machine-translator or need a specific set of instructions to several modules.

In regards to the specific creation of multi-task datasets, a most recent framework, MultiLS[21], provides another perspective on the generation of datasets for the Lexical Simplification task. The proposed paper is a very recent publication and it is not implemented in this paper. It can be a subject of investigation in future works.

C. Spanish Text Simplification

Lexical Simplification for the Spanish Language poses its own set of complications. There are currently several works for this task in the Spanish Language. A multilingual approach was taken in [22] as mTLS, a multilingual controllable Transformer-based Lexical Simplification. It uses a fined-tuned version of the T5 model. They also the TSAR-2022 shared dataset[23] for Spanish complex word identification.

Specific implementations for Spanish Lexical Simplifications have been done in the creation of the EASIER system[8][24] by using the CWI and SS/SG datasets. Their investigation provides much inspiration as the implementation takes into consideration some Plain Language aspects such as the need for word simplification.

III. Methodology

Our methodology is composed of the creation of a dataset needed for the fine tuning of Causal LLMs in the task of text generation. We will first give information on the dataset used and then we will delve into the training of the models using parameter efficient fine tuning techiques.

A. Dataset Generation

The starting dataset used is the EASIER CORPUS SSGS dataset which is a Spanish Language dataset of human annotated Suggested Synonyms for detected Complex Words. The SG/SS dataset is composed of 5154 entries with a structure as follows:

- The first column shows the ID of the document.
- The second column shows the ID of the target word.
- The third column shows the target word.
- The fourth column shows the sentence.
- The fifth column shows the suggested synonyms for the target word separated by a comma.

From this dataset, a new dataset was generated for its consumption by a Causal LLM. The SGSS SentencePairs dataset is composed of the original sentences of SG/SS and new sentences formed by replacing the suggested synonyms with the target word. As some sentences of the original dataset have more than one suggestion per word, this resulted in an increase in dataset size to 7894 entries. An example of the original sentences and their new reproductions are presented in Table I.

Original Sentence	New Sentence
La importancia de leer bien el etiquetado antes de comprar un alimento	La importancia de leer bien el letrero antes de comprar un alimento
Por eso es importante saber manejarlas	Por eso es importante saber utilizarlas
Y todo ello redunda en una mejor salud	Y todo ello repercute en una mejor salud

TABLE I
SENTENCE PAIRS SAMPLES.

The simple dataset creation code is available in the project's repository. It also includes the SGSS SentencePair Dataset.

B. Model Training

Low Rank Adaptation is a parameter efficient fine tuning technique that was developed by Hu, Shen, Wallis, Allen-Zhu, Li, Wang, Wang, and Chen[25]. This fine tuning technique freezes the pretrained model weights and injects trainable rank decomposition matrices into each layer of the Transformer architecture, greatly reducing the number of trainable parameters for downstream tasks.

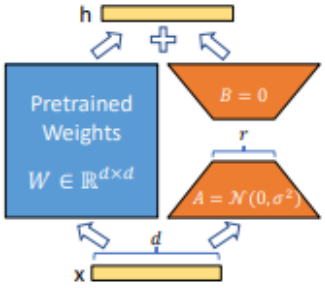

Fig. 4. LoRA reparametrization (only A and B are trained): From [25]

The key advantages of LoRA are improved training efficiency with hardware barrier lowering. For multiple task

module creation it also allows for the usage of the same base pre-trained model by only training the small LoRA modules for each task.

The Q in QLoRA[26] stands for Quantization. This addition to the previous mention work allows for an even more efficient fine-tuning. QLORA backpropagates gradients through a frozen, 4-bit quantized pretrained language model into Low Rank Adapters.

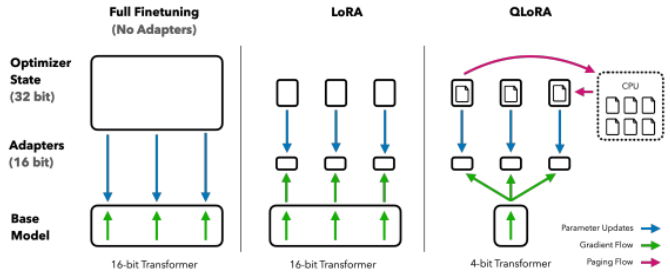

Fig. 5. QLoRA in contrast with other finetuning methods: From [26]

The innovations that allow for the saving of memory without sacrificing performance are a new data type by the name 4-bit NormalFloat that is information theoretically optimal for normally distributed weights, as well as double quantization to reduce the average memory footprint by quantizing the quantization constants.

These advancements in finetuning implementations allow for the training of LLMs in environments such as Google Colab, which is the environment used for this investigation.

Using the HuggingFace Libraries for Tranformer Model Calling, Paremeter Efficient Finetuning and Accelerate for training two models where trained for this task. The BitsandBites library was used for k-bit quantization for training optimization.

For this investigation we chose the open-sourced models LLama 2-7B[27] and Gemma-7B[28]. They where chosen for their proven performance in several tasks versus other State of the Art models. It is vital to mention that these models do have Spanish capabilities but as they where mostly trained on English texts, we expect them to under-perform when compared to tasks in English.

We propose the utilization of these models for the investigation of their performance in this task. We are aware that their Masked LM competitors should be better due to their architecture at bidirectional context consideration for prediction. Our desire is to corroborate and evaluate these newly released models versus previous implementations.

As it was shown in the Related Work section, most implementations of the Lexical Simplification task use Sequence-to-Sequence models for the insertion of synonyms on the masked space left from the removal of complex words. The models used in this investigation are Causal Language Models, which means that they generate text from a learned distribution of possible next tokens. As such, the treatment of the training data for the model varies.

The creation of a formatted instruction is the first step as part of the prompt creation for input on the model. The prompt given to the model should be of the form:

```
"### Instruction:\n{x['sentence']}\n\n### Response:\n{x['new_sentence']}"
```

After the data was transformed to a correct prompt format, an investigation on the tokenized prompt lenght was done. With this we verify if shortening our prompts is needed. If our tokenized prompt length is larger than the context length of the model, training will not be succesfull as the full context will not be taken into account.

For our dataset, the median value of tokenized prompt lenght is around 150 with most outliers at 350 tokens. This is much lower than the threshold for Llama-2-7B (4,096 tokens) and Gemma-7B (8,192 tokens). Also, our rule-set derived from the Clear Language Guidelines proposes a maximum sentence length of 25 words and paragraphs of no more than 8 sentences. Considering an approximate relation of 3 tokens per word, as we use different tokenizers per model, we get a maximum desired size of 600 tokens which are well within the context length.

Future works will include the training of Mixtral-8x7B-v0.1. from Mistral AI, as well as a Sequence-to-Sequence model from the T5 family and nllb-200-distilled-600M from Meta for comparison.

The LS Llama and Gemma QLoRA Colab Notebook provides the complete implementation of the finetuning of the model.

IV. Evaluation

At the moment we have no evaluation of the results of our training. We propose Cosine Similarity, as well as Perplexity and METEOR.[29]

V. Discussion

Results at plain sight after training show that improvements can be made regarding the dataset used. Easy enhancments can be made to improve the training performance. Also, as the models used, Llama2 and Gemma, are both moslty trained in English text, it is expected to see a low performance even after training.

The positive aspects we can extract from the experiments is that QLoRA works as a parameter efficient finetuning technique as it did train the trainable parameters in a Colab environment without the need for specialized hardware.

VI. Conclusions and Future Work

Due to the simple replacement technique used for the dataset creation, article matching to replaced word was not done. Also, more sentence pairs could be created by substituting all synonym words in one sentence. Most sentences of the original dataset have more than one target word and this could increase greatly the dataset for a better implementation of the lexical simplification task.

The rules created from the Clear Language Documentation where not used for the verification of the sentence pairs created. Future work will include this rules for verification

and creation of a fully compliant dataset for all the Language Simplification tasks.

In this investigation we use one LoRA module for one task but future works could use this same structure for training several modules for the completion of each of the Language Simplification tasks.

Future works will include the training of Mixtral-8x7B-v0.1. from Mistral AI, as well as a Sequence-to-Sequence model from the T5 family and nllb-200-distilled-600M from Meta for comparison. With this addition we should get better results as they are multi-language pre-trained models that include Spanish for their out of the box use.

REFERENCES

[1] Temnikova, Irina. "Text complexity and text simplification." (2012).

[2] Shardlow, Matthew. "A survey of automated text simplification." International Journal of Advanced Computer Science and Applications 4, no. 1 (2014): 58-70.

[3] Paetzold, Gustavo H., and Lucia Specia. "A survey on lexical simplification." Journal of Artificial Intelligence Research 60 (2017): 549-593.

[4] Al-Thanyyan, Suha S., and Aqil M. Azmi. "Automated text simplification: a survey." ACM Computing Surveys (CSUR) 54, no. 2 (2021): 1-36.

[5] Espinosa-Zaragoza, Isabel, José Abreu-Salas, Elena Lloret, Paloma Moreda Pozo, and Manuel Palomar. "A review of research-based automatic text simplification tools." In Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing, pp. 321-330. 2023.

[6] Štajner, Sanja. "Automatic text simplification for social good: Progress and challenges." Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021 (2021): 2637-2652.

[7] Barbu, Eduard, M. Teresa Martín-Valdivia, Eugenio Martínez-Cámara, and L. Alfonso Urena-López. "Language technologies applied to document simplification for helping autistic people." Expert Systems with Applications 42, no. 12 (2015): 5076-5086.

[8] Alarcon, Rodrigo, Lourdes Moreno, and Paloma Martínez. "Lexical simplification system to improve web accessibility." IEEE Access 9 (2021): 58755-58767.

[9] Peng, Yifan, Catalina O. Tudor, Manabu Torii, Cathy H. Wu, and K. Vijay-Shanker. "iSimp: A sentence simplification system for biomedicail text." In 2012 IEEE International Conference on Bioinformatics and Biomedicine, pp. 1-6. IEEE, 2012.

[10] Cemri, Mert, Tolga Çukur, and Aykut Koç. "Unsupervised simplification of legal texts." arXiv preprint arXiv:2209.00557 (2022).

[11] ISO 24495-1:2023, Plain language — Part 1: Governing principles and guidelines (SGML)

[12] Dirección General de Simplificación Regulatoria. (2007) Lenguaje Claro. ISBN 970-653-085-1

[13] Yimam, Seid Muhie, Chris Biemann, Shervin Malmasi, Gustavo H. Paetzold, Lucia Specia, Sanja Štajner, Anaïs Tack, and Marcos Zampieri. "A report on the complex word identification shared task 2018." arXiv preprint arXiv:1804.09132 (2018).

[14] Alarcon, Rodrigo, Lourdes Moreno, and Paloma Martínez. "EASIER corpus: A lexical simplification resource for people with cognitive impairments." Plos one 18, no. 4 (2023): e0283622.

[15] Martin, Louis, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de La Clergerie, Djamé Seddah, and Benoît Sagot. "CamemBERT: a tasty French language model." arXiv preprint arXiv:1911.03894 (2019).

[16] Truică, Ciprian-Octavian, Andrei-Ionuţ Stan, and Elena-Simona Apostol. "SimpLex: a lexical text simplification architecture." Neural Computing and Applications 35, no. 8 (2023): 6265-6280.

[17] Maddela, Mounica, and Wei Xu. "A word-complexity lexicon and a neural readability ranking model for lexical simplification." arXiv preprint arXiv:1810.05754 (2018).

[18] Qiang, Jipeng, Feng Zhang, Yun Li, Yunhao Yuan, Yi Zhu, and Xindong Wu. "Unsupervised statistical text simplification using pre-trained language modeling for initialization." Frontiers of Computer Science 17, no. 1 (2023): 171303.

[19] Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan et al. "Moses: Open source toolkit for statistical machine translation." In Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions, pp. 177-180. Association for Computational Linguistics, 2007.

[20] Tan, Keren, Kangyang Luo, Yunshi Lan, Zheng Yuan, and Jinlong Shu. "An LLM-Enhanced Adversarial Editing System for Lexical Simplification." arXiv preprint arXiv:2402.14704 (2024).

[21] North, Kai, Tharindu Ranasinghe, Matthew Shardlow, and Marcos Zampieri. "MultiLS: A Multi-task Lexical Simplification Framework." arXiv preprint arXiv:2402.14972 (2024).

[22] Sheang, Kim Cheng, and Horacio Saggion. "Multilingual controllable transformer-based lexical simplification." arXiv preprint arXiv:2307.02120 (2023).

[23] Saggion, Horacio, Sanja Štajner, Daniel Ferrés, Kim Cheng Sheang, Matthew Shardlow, Kai North, and Marcos Zampieri. "Findings of the tsar-2022 shared task on multilingual lexical simplification." arXiv preprint arXiv:2302.02888 (2023).

[24] Moreno, Lourdes, Helen Petrie, Paloma Martínez, and Rodrigo Alarcon. "Designing user interfaces for content simplification aimed at people with cognitive impairments." Universal Access in the Information Society 23, no. 1 (2024): 99-117.

[25] Hu, Edward J., Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. "Lora: Low-rank adaptation of large language models." arXiv preprint arXiv:2106.09685 (2021).

[26] Dettmers, Tim, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. "Qlora: Efficient finetuning of quantized llms." Advances in Neural Information Processing Systems 36 (2024).

[27] Touvron, Hugo, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov et al. "Llama 2: Open foundation and fine-tuned chat models." arXiv preprint arXiv:2307.09288 (2023).

[28] Team, Gemma, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre et al. "Gemma: Open models based on gemini research and technology." arXiv preprint arXiv:2403.08295 (2024).

[29] Hu, Taojun, and Xiao-Hua Zhou. "Unveiling LLM Evaluation Focused on Metrics: Challenges and Solutions." arXiv preprint arXiv:2404.09135 (2024).