INSTITUTO TECNOLOGICO DE COSTA RICA
ESCUELA DE CIENCIAS DEL LENGUAJE

DOCUMENTO 1:

INFORME FINAL DE PROYECTO DE INVESTIGACION

NOMBRE DEL PROYECTO:

Generación de un modelo de simplificación automática de textos en español mediante inteligencia artificial

Código del Proyecto: 1470014

INVESTIGADORES:

Nelson Pérez Rojas, nperez@itcr.ac.cr
Saúl Calderón Ramírez, sacalderon@itcr.ac.cr
Mario Chacón Rivas, machacon@itcr.ac.cr
Martín Solís Salazar, marsolis@itcr.ac.cr

Fecha de inicio: 01/07/2022
Fecha de finalización: 30/06/2024

1. CUMPLIMIENTO DE OBJETIVOS

Objetivo general: Desarrollar un modelo para la simplificación automática de textos de temáticas financieras en español mediante inteligencia artificial			
Objetivo específico	**Productos**	**% de logro**	**Comentarios**
Construir un conjunto de datos genérico para el entrenamiento de modelos de simplificación de textos en español	Producto 1: conjunto de datos sintético	100%	
Construir un conjunto de datos compuesto de oraciones simples y oraciones homólogas complejas de textos financieros, para el entrenamiento y evaluación de los modelos de simplificación automática	Producto 2: Conjunto de datos que puede ser usado por la comunidad científica para la creación de modelos que permitan simplificar textos automáticamente Producto 8: artículo 5 Producto 9: artículo 6 Producto 10: artículo 7 Producto 12: Tesis de Maestría en Ciencias de la Computación (estudiante Mario Romero).	100%	

Evaluar modelos de aprendizaje no supervisado para la simplificación automática	Producto 3: artículo 1 Producto 7: artículo 4 Producto 11: reporte de resultados de evaluación de modelos a simplificación de textos	100%	
Evaluar modelos de aprendizaje semi supervisado/ supervisado para la simplificación automática	Producto 4: artículo 2 Producto 5: algoritmo de simplificación automática de textos Producto 6: artículo 3	100%	

2. ACEPTACION DE ARTICULO

Anexo 1 (product 3).
Paper scopus publicado en IEEE:
https://ieeexplore.ieee.org/abstract/document/10379347

Using GPT-3 as a Text Data Augmentator for a Complex Text Detector

Abstract—In this work, we explore the problem of complex text detection. This problem is a frequent challenge when implementing text simplification pipelines. Identifying complex text segments can trigger text simplification models, making a better resource usage as state of the art Large Language Models are expensive to use. We focus in Spanish, as it is an under-represented language, given the lack of simple/complex paired datasets. We use a novel paired dataset in Spanish of financial educational texts to train and test our methods. To improve the performance of the classifier, we propose the usage of text simplifications generated with GPT-3 (data augmenter) to alleviate the need to label a large number of text segments as simple or complex. We use the BERT pre-trained model on Spanish data known as Spanish BERT (BETO) and explore the effect of augmenting target data in the model performance.

Anexo 2 (product 4).
Paper scopus publicado en IEEE:
https://ieeexplore.ieee.org/abstract/document/10379278

Uncertainty Estimation for Complex Text Detection in Spanish

Abstract—Text simplifcation refers to the transformation of a source text aiming to increase its readiblity and understandability for a specific target population. This task is an important step towards improving inclusivity of such target populations (i.e., low scholarity or visually/hearing impaired groups). The recent advancements in the field brought by Large Language Models improve the performance of machine based text simplification approaches. However, using Language Models to simplify large text segments can be resource demanding. A more simple model to classify whether the text segment is worth to simplify or not can improve resource efficiency, in order to avoid unnecessary text prompts to the Large Language Models. Furthermore, text simplicity categorization can also be used for other purposes, such as text complexity measurement. The discrimination of text segments into simple and complex categories might lead to a number of false positives or negatives for a not well-tuned model. A way to control the acceptance threshold, is the implementation of an uncertainty score for each prediction. In this work we explore two simple uncertainty estimation approaches for complex text identification: a Monte Carlo Dropout and an Deep Ensemble Based approach. We use an in-house dataset in the financial education domain for our tests. We calibrated the two implemented methods to find out which performs better, using a Jensen-Shannon based distance between the correct and incorrect outputs of the discriminator. Our tests showed an important advantage of the Monte Carlo Dropout over the Deep Ensemble Based method.

Anexo 3 (product 6).
Paper publicado Revista *Tecnología en Marcha:*

Exploration and selection of LLM models for financial text simplification

Abstract-This research is dedicated to the simplification of Spanish-language financial texts to enhance accessibility for screen readers. We present a qualitative and quantitative analysis of the text simplification process, employing a set of Spanish simplification rules and metrics. Our **study** evaluates the outcomes resulting from the application of three distinct financial datasets to four pre-trained models. The primary objective is to identify the most effective models for text simplification and determine those warranting further investment through fine-tuning and training. This study contributes to improving the accessibility and comprehensibility of financial documents for individuals with visual impairments.

Anexo 4 (producto 7).
Paper scopus publicado en IEEE

https://ieeexplore.ieee.org/document/10032482

Towards Text Simplification in Spanish: A Brief Overview of Deep Learning Approaches for Text Simplification

Abstract—Text simplification refers to the transformation of a specific source text into a target text aiming to increase understanding and readability for one or more specific audiences. This task demands large human efforts and specialized knowledge, which makes the usage of automated or semi-automated computational approaches appealing. The rise of deep learning as an unifying paradigm between seemingly different fields as image analysis, sound processing and natural language processing has considerably influenced the current state of the art approaches for automatic text simplification. Therefore, in this work, we focus on the study of deep learning based state of the art methods for automatic text simplification in the Spanish language. For this end, we first disentangle the different tasks which can be addressed in order to yield a simplified text in general. Later we review the latest deep learning-based approaches, along with the main datasets and performance metrics used in the field. We also describe approaches to deal with small datasets and technical words. Finally, we describe some lessons to build accurate automatic text simplification systems in Spanish, as in this language there is a noticeable shortage of work for text simplification

Anexo 5 (producto 8).
Paper generado
An Extensible Massively Multilingual Lexical Simplification Pipeline Dataset using the MultiLS Framework
https://aclanthology.org/2024.readi-1.4/

Abstract - We present preliminary findings on the MultiLS dataset, developed in support of the 2024 Multilingual Lexical Simplification Pipeline (MLSP) Shared Task. This dataset currently comprises of 300 instances of lexical complexity prediction and lexical simplification across 10 languages. In this paper, we (1) describe the annotation protocol in support of the contribution of future datasets and (2) present summary statistics on the existing data that we have gathered. Multilingual lexical simplification can be used to support low-ability readers to engage with otherwise difficult texts in their native, often low-resourced, languages.

Anexo 6 (producto 9).

Paper generado aún no publicado, en revisión en *IEE ACCESS*:
A Novel Spanish Dataset for Financial Education Text Simplification Targeting Visually Impaired Individuals

ABSTRACT Automatic text simplification is a crucial task in natural language processing, aimed at making texts more comprehensible, particularly for specific groups such as individuals with visual impairments. One of the primary challenges in developing models for automatic Text Simplification (TS) is the scarcity of data, especially in Spanish. This manuscript introduces a novel dataset tailored for Spanish speakers with visual impairments, consisting of 5,314 pairs of original and simplified sentences created using established simplification rules. Additionally, we evaluate the feasibility of augmenting this dataset using large language models such as Generative Pre-training Transformer (GPT)-3, Tuner, and Multilingual T5 (MT5). We compare the simplifications generated by these models with our dataset to assess their effectiveness in data augmentation. The characteristics of our dataset and the findings from these comparisons are discussed in detail. The dataset is publicly available on Hugging Face at https://huggingface.co/datasets/saul1917/FEINA.

Anexo 7 (producto 10).

Paper generado aún no publicado, en revisión en TSAR 2024 (como parte de *The 2024 Conference on Empirical Methods in Natural Language Processing*)

Lexical Complexity Prediction and Lexical Simplification for Catalan and Spanish: Resource Creation, Quality Assessment, and Ethical Considerations

Abstract-Automatic lexical simplification is a task to substitute lexical items that may be unfamiliar and difficult to understand with easier and more common words. This paper presents the description and analysis of two novel datasets for lexical simplification in Spanish and Catalan. This dataset represents the first of its kind in Catalan and a substantial addition to the sparse data on automatic lexical simplification which is available for Spanish. Specifically, it is the first dataset for Spanish which includes scalar ratings of the understanding difficulty of lexical items. In addition, we present a detailed analysis aiming at assessing the appropriateness and ethical dimensions of the data for the lexical simplification task.