Procesamiento del Lenguaje Natural, Revista nº 70, marzo de 2023, pp. 111-122 recibido 15-12-2022 revisado 09-02-2023 aceptado 15-02-2023

Tuning BART models to simplify Spanish health-related content

Ajuste de modelos BART para simplificación de textos sobre salud en español

Rodrigo Alarcón, Paloma Martínez, Lourdes Moreno
Human Language and Accessibility Technologies group (HULAT)
Universidad Carlos III de Madrid
Leganés, Madrid, Spain
{ralarcon, pmf, lmoreno}@inf.uc3m.es

Abstract: Health literacy has become an increasingly important skill for citizens to make health-relevant decisions in modern societies. Technology to support text accessibility is needed to help people understand information about their health conditions. This paper presents a transfer learning approach implemented with BART (Bidirectional AutoRegressive Transformers), a sequence-to-sequence technique that is trained as a denoising autoencoder. To accomplish this task, pre-trained models have been fine-tuned to simplify Spanish texts. Since fine tuning of language models requires sample data to adapt it to a new task, the process of creating of a synthetic parallel dataset of Spanish health-related texts is also introduced in this paper. The results on the test set of the fine-tuned models reached SARI values of 59.7 in a multilingual BART (mBART) model and 29.74 in a pre-trained mBART model for the Spanish summary generation task. They also achieved improved readability of the original texts according to the Inflesz scale.

Keywords: lexical simplification, Spanish, language models, Spanish, multilingual BART.

Resumen: La alfabetización sanitaria se ha convertido en una habilidad cada vez más importante para que los ciudadanos tomen decisiones sobre su salud en las sociedades modernas. Para ayudar a las personas a comprender la información sobre su estado de salud, es necesaria una tecnología que facilite la accesibilidad de los textos. Este artículo presenta un enfoque de transfer learning implementado con BART (Bidirectional AutoRegressive Transformers), una técnica sequence-to-sequence que se entrena como un autoencoder de eliminación de ruido. Para llevar a cabo esta tarea, se han ajustado modelos preentrenados para simplificar textos en español. Dado que el ajuste de los modelos lingüísticos requiere datos de muestra para adaptarlos a una nueva tarea, en este artículo también se presenta el proceso de creación de un conjunto de datos paralelos sintéticos de textos en español relacionados con la salud. Los resultados en el conjunto de prueba de los modelos afinados alcanzaron valores SARI de 59,7 en un modelo multilingual BART (mBART) y 29,74 en un modelo mBART pre-entrenado para la tarea de generación de resumenes en español. Además lograron mejorar la legibilidad de los textos originales según la escala de Inflesz.

Palabras clave: Simplificación léxica, modelos del lenguaje, Español, BART multilingüe.

1 Introduction

Nowadays, the average citizen has access to much more information through the Internet than at any other time in history with a high impact on most people's daily lives. However, this information may have been written in a form that makes the content hard to understand (Saggion et al., 2015). Difficulty with texts on the Internet can affect a wide range of people such as deaf people, illiterate

ISSN 1135-5948 DOI 10.26342/2023-70-9

people, second language learners, and people with intellectual disabilities, among others (Moreno, Alarcon, and Martínez, 2020).

One of the most difficult content to understand is health-related content because of the excessive use of abbreviations, incomplete sentences, and specific terminology. Poor health literacy is a limiting factor that prevents patients from making well-informed health decisions, which can result in high costs for both healthcare institutions and the patient (Kauchak, Apricio, and Leroy, 2022).

Following this issue, there are standards and guidelines (UNE, 2018; Plainlanguage, 2017b; Plainlanguage, 2017a) that provide accessibility requirements and criteria to make the textual content more cognitively accessible, through the application of easy-to-read and plain language guidelines. For example, these requirements indicate that a text must be written in an active voice, use everyday words and/or use short sentences as much as possible. All these requirements and criteria are defined to provide a familiar and simple vocabulary used in texts in Plain Language. Nonetheless, this issue is difficult to address.

There are ways to follow these directives and manually deal with this problem. For instance, some websites offer simplified versions of their original content oriented to their target users [1][2]. However, this is a time consuming task. Therefore, over the years, different proposals to provide an automatic solution to this problem have emerged, the most prominent of which is Natural Language Processing (NLP) techniques (Alarcon, Moreno, and Martínez, 2021; Alarcón García, 2022).

This article proposes a transfer learning method to simplify Spanish texts with medical content. To achieve this, a state-of-the-art approach is presented, by fine-tuning multilingual BART models (Tang et al., 2020) with parallel data to lexical simplification of Spanish health-related content. This strategy was chosen because it has achieved state-of-the-art results in a diverse set of generation tasks (Martin et al., 2020) and outperforms Text-to-Text transfer transformers (T5) models of comparable size (Lewis et al., 2019).

The contributions of this paper can be outlined as follows:

- Creation of a Spanish synthetic parallel resource for the training and validation of simplification methods in the health domain. This resource contains pairs of original sentences related to simplified ones.
- Proposal of fine-tuning two mBART models for text-to-text generation, with the aim of simplifying Spanish health-related texts.

2 Related Work

Text simplification is the process of lexically and/or syntactically modifying a text to produce a simple version of the original text (Al-Thanyyan and Azmi, 2021), preserving its original meaning. Text simplification could benefit a wide range of people, to mention a few, may include second language learners (Paetzold and Specia, 2016b) or people with some type of disability, such as autism (Barbu et al., 2015), dyslexia (Wilkens, Oberle, and Todirascu, 2020) or some type of intellectual disability (Saggion et al., 2015; Alarcon, Moreno, and Martínez, 2021).

Over the years, resources to support training and/or evaluation of automatic text simplification algorithms have been shared. These resources belong either in a general domain such as resources with content from Wikipedia articles (Yimam et al., 2018; Ferrés and Saggion, 2022a) or other resources with a specific domain, such as resources with a medical vocabulary (Campillos Llanos et al., 2022). Additionally, there have been evaluation campaigns aimed at providing a solution to this task in a modular way (Truică, Stan, and Apostol, 2022), such as workshops that aimed to foster research on the detection of unusual words in a given text (Paetzold and Specia, 2016a; Yimam et al., 2017), others that focused in ranking words according to their complexity (Shardlow et al., 2021) and competitions that aimed to propose replacements for unusual words or phrases (McCarthy and Navigli, 2007). Other works presented strategies using parallel resources, as in the work of (Zhu, Bernhard, and Gurevych, 2010) who proposed a complex word identification trans-

[1] https://plenainclusionmadrid.org/blog/etapa-educativa-inclusion/

[2] https://plenainclusionmadrid.org/blog/reclutador-discapacidad-intelectual/

	EASIER	
	Sentence	**Substitutes**
EASIER	El tabaquismo constituye el principal problema de salud pública prevenible en los países desarrollados siendo un factor determinante de numerosas **patologías**. (Smoking is the main preventable public health problem in developed countries and is a determining factor in numerous **pathologies**.)	enfermedades (diseases), dolencias (afflictions), trastornos (disorders)
	Sentence	**Simple version**
Paralell Instance	El tabaquismo constituye el principal problema de salud pública prevenible en los países desarrollados siendo un factor determinante de numerosas **patologías**. (Smoking is the main preventable public health problem in developed countries and is a determining factor in numerous **pathologies**.)	El tabaquismo constituye el principal problema de salud pública prevenible en los países desarrollados siendo un factor determinante de numerosas **enfermedades**. (Smoking is the main preventable public health problem in developed countries and is a determining factor in numerous **diseases**.)
Paralell Instance	El tabaquismo constituye el principal problema de salud pública prevenible en los países desarrollados siendo un factor determinante de numerosas **patologías**. (Smoking is the main preventable public health problem in developed countries and is a determining factor in numerous **pathologies**.)	El tabaquismo constituye el principal problema de salud pública prevenible en los países desarrollados siendo un factor determinante de numerosas **dolencias**. (Smoking is the main preventable public health problem in developed countries and is a determining factor in numerous **afflictions**.)
Paralell Instance	El tabaquismo constituye el principal problema de salud pública prevenible en los países desarrollados siendo un factor determinante de numerosas **patologías**. (Smoking is the main preventable public health problem in developed countries and is a determining factor in numerous **pathologies**.)	El tabaquismo constituye el principal problema de salud pública prevenible en los países desarrollados siendo un factor determinante de numerosas **transtornos**. (Smoking is the main preventable public health problem in developed countries and is a determining factor in numerous **disorders**.)
	EASY-DPL	
	Sentence	**Target Word - Substitutes**
Easy-DPL	En pacientes con esquizofrenia la incidencia de **acatisia** fue de 6,2% para aripiprazol y de 3,0% para placebo. (In patients with schizophrenia, the incidence of **akathisia** was 6.2% for aripiprazole and 3.0% for placebo.)	acatisia (akathisia) - incapacidad de quedarse quieto (inability to remain still)
Easy-DPL	lteraciones gastrointestinales: Frecuente (1% y <10%): dolor abdominal, diarrea, **dispepsia**, náuseas y vómitos. (Gastrointestinal alterations: Frequent (1% and <10%): abdominal pain, diarrhea, **dyspepsia**, nausea and vomiting.)	dispepsia (dyspepsia) - enfermedades del estómago (diseases of the stomach)
	Sentence	**Simple version**
Paralell Instance	En pacientes con esquizofrenia la incidencia de **acatisia** fue de 6,2% para aripiprazol y de 3,0% para placebo. (In patients with schizophrenia, the incidence of **akathisia** was 6.2% for aripiprazole and 3.0% for placebo.)	En pacientes con esquizofrenia la incidencia de **incapacidad de quedarse quieto** fue de 6,2% para aripiprazol y de 3,0% para placebo. (In patients with schizophrenia, the incidence of **inability to stay still** was 6.2% for aripiprazole and 3.0% for placebo.)
Paralell Instance	Alteraciones gastrointestinales: Frecuente (1% y <10%): dolor abdominal, diarrea, **dispepsia**, náuseas y vómitos. (Gastrointestinal alterations: Frequent (1% and <10%): abdominal pain, diarrhea, **dyspepsia**, nausea and vomiting.)	..Alteraciones gastrointestinales: Frecuente (1% y <10%): dolor abdominal, diarrea, **enfermedades del estómago**, náuseas y vómitos. (Gastrointestinal alterations: Frequent (1% and <10%): abdominal pain, diarrhea, **diseases of the stomach**, nausea and vomiting.)

Table 1: EASIER and EASY- DPL corpora substitutes dataset examples.

lation method with a tree-based simplification model trained on a parallel Wikipedia and simple Wikipedia dataset. A prominent project for the Spanish language is the Simplext project (Saggion et al., 2015), where a parallel resource was generated in Spanish to reduce the syntactic complexity of texts.

A recent competition is CLEF-2022, where three tasks focused on the automatic simplification of scientific texts were proposed. Of these tasks we can highlight tasks 2 and 3, where the teams with the best results were based on the use of language models in their strategies. Task 2 consisted in the detection of terms in a text that requires an explanation for the whole text to be understood (Ermakova et al., 2022a). Participants obtained a train set of 453 examples annotated with difficulty scales and a test set of 116763 sentences, where each participant had to determine a score for the difficulty of the term in the target text. Approaches based on IDF (Inverse Document Frequency) term weighting (Mostert et al., 2022), approaches based on semantic similarity complemented by different lexical and syntactic features (Huang and Mao, 2022), and finally, methods based on transfer learning (Talec-Bernard, 2022) were presented. Task 3 aimed at generating simplified versions of scientific texts (Ermakova et al., 2022b). Participants were given 648 parallel sentences to develop their architectures, and to validate them, they obtained a test set of 116724 instances to be evaluated by the organizers. Transfer learning-based approaches were presented, where models were fine-tuned with the task data and other existing English language corpora (Monteiro, Aguiar, and Araújo, 2022). A similar approach to the one proposed in this paper was described in (Rubio and Martínez, 2022) by fine-tuning a BART model to simplify sentences. This method was highlighted by the task organizers as it showed that tasks 2 and 3 of the competition are largely related.

In addition, in the Shared Task on Lexical Simplification (TSAR 2022) (Saggion et al., 2023) for English, Portuguese and Spanish languages, given a sentence/context and a complex target word, participants had to generate up to 10 possible substitutes ordered by simplicity. To perform this task, the organizers shared different resources for the training and/or validation of systems. ALEXSIS (Ferrés and Saggion, 2022a) dataset, which contains open domain terms, was used in the case of Spanish language. Prior to the publication of the task, the authors of this work experimented with this resource to rank substitutes for target words, achieving an accuracy score of 0.51 (Alarcón García, 2022). However, since the objective of this work is to simplify medical terms, a specific medical domain resource is proposed to train/validate the methods described in this work.

Research with BART out of competitions has been also published recently, as (Cumbicus-Pineda et al., 2022) outperforms other approaches in three different English datasets using several language models, trained with complex sentences to predict simple sentences and others trained with simple sentences to predict complex sentences, achieving higher values in the SARI metric than other similar approaches. (Chamovitz and Abend, 2022) described a BART-based method that also defines a series of simplification operations based on cognitive simplification guidelines, improving the performance compared to a baseline model in a dataset for the English language. Some of these operations consisted of ambiguity reduction, rephrasing, summarizing, reordering or deleting paragraphs. The work of (Štajner, Sheang, and Saggion, 2022) presented a sentence simplification approach by experimenting with transformer models for text simplification such as BART and T5 combined with control mechanisms, achieving results comparable to other previous systems.

This paper is based on metrics and methods from BART's previously described work and presents a text-to-text generation approach by fine-tuning two mBART models for the task of text simplification. To accomplish this task, this paper also describes the process of creating a synthetic Spanish resource containing lexical modifications to original sentences.

3 Datasets

This Section briefly describes the data used to fine-tune the BART language models. These data are obtained from the EASIER[3] and EASY-DPL[4] (Segura-Bedmar and Martínez, 2017) corpora.

3.1 EASIER

The EASIER corpus was created to support Complex Word Identification (CWI) and Substitute Generation/Selection (SG/SS) tasks, two important processes in lexical simplification, targeting an audience with intellectual disabilities. With this objective, linguistic experts in easy-to-read and simple language guidelines have annotated 260 news documents on various topics, including health news. Currently, this resource has gathered 8155 complex words and 7894 proposed substitutes.

For the purpose of text simplification, data from the SG/SS dataset were used (Alarcon, 2021). EASIER corpus contains simple alternative substitutes to existing complex words. To create the instances of the tuning process, parallel versions are created by taking the original sentences, the target complex word, and the proposed substitutes. As a result, 7894 instances were obtained where for each instance there is a code, original sentence, and the same original sentence where one or more words have been replaced. Table 1 shows examples of the original content of the EASIER corpus dataset and the content of the generated parallel versions. The datasets of this resource are available in csv formats.

3.2 EasyDPL

The remaining data used for the experiments in this article come from the Easy-DPL corpus (easy drug leaflets). This corpus was annotated by three annotators trained for their task, where they annotated the adverse effects section of 306 medical leaflets, resulting in 1400 adverse reactions detected along with their simplest synonym. Table 1 shows examples of the original content and the generated parallel versions. This resource is available in XML and BRAT formats.

[3] Easier Corpus Repository github.com/LURMORENO/EASIER_CORPUS

[4] https://github.com/isegura/EasyDPL

3.3 EASIER-EasyDPL dataset

A Spacy model[5] in Spanish was used to generate the parallel dataset to eliminate duplicate instances, tokenizing, and sentence splitting, among other operations. For this version of the resource, possible errors in grammatical forms were ignored when substituting a target word in the original sentence. Table 2 shows some statistics between the resources described above.

	Number of instances	**%**
EASIER	7894	86.5
Easy-DPL	1230	13.5
Total	9124	100

Table 2: Number of instances for the EASIER and Easy-DLP resources.

4 Methods and system description

This Section describes the proposal, which is based on fine-tuning two pre-trained multilingual BART models from HuggingFace. The first model (MBART-50)[6] is 12 layers multilingual sequence-to-sequence model trained on 50 different languages, while the second model (MBART-ESP)[7] is a 12 layer Spanish language fine-tuned version of the first model (Tang et al., 2020) with the wiki_lingua dataset[8] for the summarization task. The hypothesis behind the choice of these models is to determine whether the model fine-tuned to the Spanish language is better at the simplification task than the base model because it was trained to better understand the Spanish language.

BART (Bidirectional AutoRegressive Transformers), (Lewis et al., 2019), is a sequence-to-sequence strategy trained as a denoising autoencoder. This technique resembles BERT and GPT as it uses a standard sequence-to-sequence Neural Machine Translation architecture (transformer) with a bidirectional encoder (Devlin et al., 2018) and a left-to-right decoder (Radford et al., 2018). This model could be fine-tuned to the simplification problem by taking a text sequence as input and producing a text sequence as output. Given a complex text 'x' and its references 'y', a model in inference time is used to select the simplification that maximizes this probability (e.g. *$argmax_y p(y|x)$)*. To train a BART model, a bidirectional encoder similar to BERT is used, where spaces are masked from the input text (adding ”noise”). Also, autoregressive decoder such as GPT is used, which reconstructs the original input, using the output of the encoder and the previous unmasked tokens.

For the experimentation of this work, the training data set described in Section 3 was used to fine-tune the models. The inputs to the process are the source sentence and the simplified sentence. Each model tokenizes each sentence and obtains the embeddings of the inputs for the transformers. With a transformer encoder, it is not necessary to pass each word individually through the input embedding, all words in the sentence are passed simultaneously and the word embeddings are simultaneously determined.

5 Experiments and results

Different experiments were performed with the data described in Section 3. These data were randomly divided into three sets with the help of the sklearn library, a training set (80%), a dev set (10%), and a test set (10%). The experiments and resources described in this article can be found in a public repository[9]. The objective of fine-tuning with this data is to create models capable of generating simplifications as close as possible to those provided by taking into account the lexical modifications of the synthetic parallel versions.

The evaluation metrics are the following:

- SARI: Measures the goodness of words that are added, deleted, and kept by the predictions. This metric was widely used in lexical simplification tasks (Xu et al., 2016).
- ROUGE: Measures the number of matching n-grams between the model-generated text and the dataset´s references. Because of using generative models in this work ROUGE is proposed as an evaluation metric. Although mBART models were fine-tuned with

[5] https://spacy.io/models/es

[6] https://huggingface.co/facebook/mbart-large-50

[7] https://huggingface.co/eslamxm/MBART-finetuned-Spanish

[8] https://huggingface.co/datasets/wiki_lingua

[9] https://github.com/ralarcong/BART_for_simplification

Perspicuity	**Inflesz**
0-40	Very difficult
40-55	Somewhat difficult
55-65	Normal
65-80	Easy enough
80-100	Very easy

Table 3: Interpretation of Inflesz Scale.

parallel data with lexical changes, they sometimes seek to reorder the content of a sentence, especially the MBART-ESP model, which was previously fine-tuned for the task of text summarization (Lin, 2004).

- Inflesz Scale: It was chosen to measure readability levels of the original texts, the target texts, and those predicted by the models. This metric, adapted to today's average Spanish reader, measures perspicuity, which refers to the level of clarity and comprehensibility of a text. Formula 1 shows the calculation of this metric where S represents the number of syllables, P the number of words and F the number of sentences. This metric can be used for any text domain, although it has initially been used in the healthcare domain to assess the readability of informed consent, package leaflets, and health education materials (Barrio-Cantalejo et al., 2008). Table 5 describes the interpretation for every range of values.

$$I = 206.835 - \frac{62.3S}{P} - \frac{P}{F} \qquad (1)$$

To train each model, different values of hyperparameters had to be explored. Fortunately, the Fast.ai library[10] helped by choosing a learning rate appropriate to the configuration set in each minibatch (Smith, 2018). By defining a ”learner” object, the library is able to test between different learning rate values and plot the loss values. Figure 1 shows an example of this, where the learning rate was chosen before it diverges.

Table 5 shows the experimentation with the other hyperparameters. It was observed that the optimal number of epochs for this experiment was 4 since with more epochs the model started to overfit the data to the training data. Figure 2 shows an example of

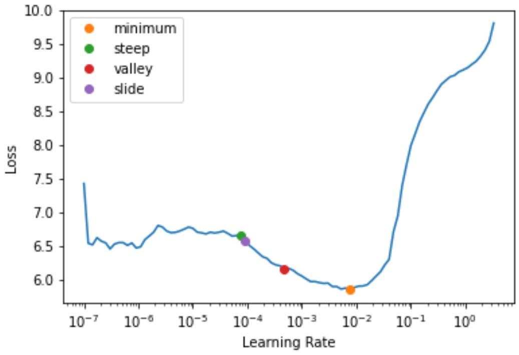

Figure 1: Loss value vs learning rate.

the MBART-ESP model, showing the loss in training and validation at 7 epochs, were at a higher epoch than the optimum the loss in training is reduced but the loss in validation is increased.

Hyperparam.	**Value**	**Best**
# epochs	[1,2,3,4,5,6,7]	[4]
Batch size	[1,2,3]	[1]
Max length, Min length	[(10,30),(10,40), (15,30),(10,50)]	(10,50)
# beams	[3][4][5]	[4]

Table 4: List of tested hyperparameters along with the best choice for the experiment.

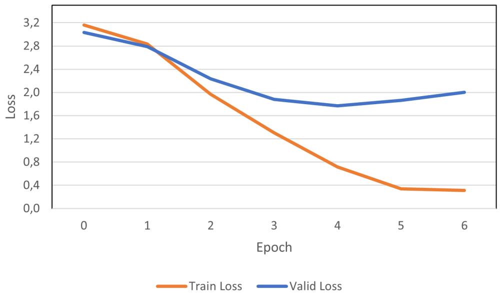

Figure 2: Loss value vs epochs.

When testing batch sizes, it was found that the best results were achieved with the length of 1, by reducing the potential noise that arises with increasing length. Also, increasing the batch length demanded more memory space, so it was decided to give preference to memory usage. On the other hand, when experimenting with the minimum and maximum output lengths, there was a noticeable change in the results when reducing the lengths, so the decision was made to keep the maximum optimal length at 50 words. Finally, when experimenting with the num-

[10] https://docs.fast.ai/

ber of beams, it was decided to keep the default value of 4, since increasing the number of beams dramatically increased the training time without obtaining better results.

Once the optimal configuration was explored, the results shown in Table 5 were obtained with the train set data. The MBART-ESP model reached Rouge 1, Rouge 2, and Rouge L scores of 0.622, 0.477, and 0.573 respectively, and a SARI score of 43.68 points. While the MBART-50 model reached Rouge 1, Rouge 2, and Rouge L scores of 0.859, 0.82, and 0.858 respectively, and a SARI score of 67.3 points.

Additionally, these models were validated with the other two sets. Table 5 shows the results where it can be seen that in the dev set the MBART-ESP model reached scores of 0.682, 0.548, 0.635 in the Rouge 1, Rouge 2, and Rouge L metrics respectively, and a score of 29.175 points in the SARI metric. While the MBART-50 model reached scores of 0.928, 0.883, and 0.928 in the Rouge 1, Rouge 2, and Rouge L metrics respectively, and a score of 58.555 points in the SARI metric. In the test set the MBART-ESP model reached scores of 0.675, 0.535, and 0.627 in the Rouge 1, Rouge 2, and Rouge L metrics respectively, and a score of 29.749 points in the SARI metric. While the MBART-50 model reached scores of 0.926, 0.881, and 0.925 in the Rouge 1, Rouge 2, and Rouge L metrics respectively, and a score of 59.777 points in the SARI metric.

The results on these datasets suggest that because the MBART-ESP model was previously trained for the summarization task, in addition to attempting to perform lexical substitutions it attempts to summarize the content, thus scoring lower on the Rouge and SARI metrics than the MBART-50 model, which has been trained only for the text simplification task. Although ROUGE is not the most appropriate metric for this simplification task, it was used since it allowed the detection of the difference in the predictions of both models, being MBART-50 the one that better performed the necessary lexical replacements, as could be seen with the SARI metric.

An important feature of these fine-tuned models is that they perform the lexical substitutions for which they were trained. Furthermore, it additionally takes into account substitutions from other instances and attempts to modify complex words in the entire target sentence. Appendix A, Table A shows examples of the models input, target, and prediction. In the first example the target word to be replaced is expectación (expectation), however, both models predict a sentence where the target word and the word suscitas (suscitas) are replaced by a simpler substitute.

Different scenarios occur with the second example. The MBART-50 model performs the desired lexical substitutions as in example 1, but in some instances the MBART-ESP model attempts to summarize the content (example 2.1), as it was the model was previously trained for the task of summarization. Therefore, it is concluded that for this specific experimentation, the MBART-50 model is more appropriate, since it focuses on the simplification task to which it was trained (example 2.2).

Finally, to evaluate the readability of the predictions, in each dataset, the Inflesz metric was calculated along the original sentences (Source), the simplified sentences (Reference), and the predictions of the models (Prediction). Table 5 shows this score on every set, where it can be seen that both models improved the readability levels of the original sentences (Source), and in some cases surpassed the readability level of the simplified parallel sentences (Reference), such is the case of the predictions of the MBART-ESP model in the Train and Test sets, obtaining a score of 41.41 and 44.3 respectively. This is due to the fact that this model tends to summarize, and the predictions are shorter, thus obtaining a better score than the MBART-50 model that only performs lexical modifications.

Since the EASIER-EasyDPL dataset is introduced in this paper, there is no direct way of comparison with other approaches. However, Table 5 shows a comparison of our best result with other works for the English language with the SARI metric in task 3 of the SimpleText@CLEF-2022 workshop. As can be seen, the results are comparable to those present in the state of the art, as in (Monteiro, Aguiar, and Araújo, 2022) where they used a T5 model to perform the text simplification task reaching 31.26 SARI values on the workshop's dataset. Another approach to this competition presented the tuning of a BART model for the English text simplifica-

MBART-ESP						
Epoch	**Train Loss**	**Valid Loss**	**Rouge 1**	**Rouge 2**	**Rouge L**	**SARI**
0	3.687307	3.640873	0.309830	0.121615	0.244606	-
1	2.434027	2.700396	0.416706	0.219104	0.345381	-
2	1.119833	1.905820	0.571923	0.406836	0.510689	-
3	0.650932	1.837321	0.622281	0.477601	0.573849	43.6808
MBART-50						
Epoch	**Train Loss**	**Valid Loss**	**Rouge 1**	**Rouge 2**	**Rouge L**	**SARI**
0	1.021594	0.891353	0.679494	0.595874	0.667983	-
1	0.573852	0.563872	0.805591	0.754661	0.802736	-
2	0.371475	0.387605	0.853029	0.812000	0.852026	-
3	0.212017	0.366373	0.859541	0.820185	0.858498	67.3065

Table 5: Train dataset results (4 epochs with optimal configuration).

	Dev			
Fine-tuned model	**Rouge 1**	**Rouge 2**	**Rouge L**	**SARI**
MBART-ESP	0.6821	0.5484	0.6354	29.175
MBART-50	0.9287	0.8837	0.9281	58.555
	Test			
Fine-tuned model	**Rouge 1**	**Rouge 2**	**Rouge L**	**SARI**
MBART-ESP	0.6756	0.5358	0.6276	29.749
MBART-50	0.9261	0.8816	0.9251	59.777

Table 6: Dev and Test datasets results (model trained with optimal configuration).

	Src	**Ref**	**Pred M-ESP**	**Pred M-50**
Train	38.75	40.24	41.41	39.82
Dev	39.21	40.90	39.05	39.73
Test	38.63	40.81	44.30	39.75

Table 7: Inflesz scale results across the datasets.

System	**SARI**
Our approach	**59.7**
HULAT@CLEF	47.8
PortLinguE@CLEF	38.1
CLARA-HD@CLEF	37.4

Table 8: SARI values for the English dataset in the SimpleText workshop.

tion task reaching SARI values of 47.83 (Rubio and Martínez, 2022). In the same competition, the approach of (Menta and Garcia-Serrano, 2022) presented a transfer learning method where they combined control tokens such as word length, paraphrasing or syntactic complexity to help in the predictions of the COVID-SciBERT model, reaching SARI values of 37.4 in the workshop's dataset.

6 Conclusions

This paper presented the process of fine-tuning two mBART pre-trained models for text simplification for the Spanish language. Because this technique requires sample data for its execution, a new synthetic resource that includes data from two corpora oriented to the simplification of Spanish texts containing health-related terminology is also introduced. This resource was divided into three subsets for training, adjustment, and validation of the different fine-tuned models. In the training and fine-tuning phase, different configurations were experimented with in order to capture the best similarity to the target sentences of the sets.

The results in the training dataset shown the great difference between each pre-trained model. Similarly, the results of these models in the dev and test sets showed a great difference. Therefore, the predictions of both fine-tuned models were analyzed, where it was observed that both models lexically modified the target words in a sentence and also modified the learned words in other examples, optimizing the simplification task. But also the pre-trained model for the summarization task in some cases tended to reduce the sentence length instead of performing the lexical modifications, resulting in lower ROUGE and SARI scores, but improving on the Inflesz readability metric. In addition, these fine-tuned models showed comparable results in the SARI metric to approaches in a similar task for the English language.

As future work, it is planned to incorporate new resources to the training/fine-tuning/validation sets containing substitutes to target words with health-oriented content, such as the IULA resource (sentences of clinical cases in Spanish)(Marimon, Vivaldi, and Bel Rafecas, 2017) and also to extend the domain of the models with news resources such as the ALEXSIS dataset (Ferrés and Saggion, 2022b). More resources with plain an easy-to-read texts written by experts are also necessary to obtain models with better performance.

Moreover, as shown in this research, the tuning process was only performed on two embedding models, so it would be interesting to experiment with other multilingual models of different sizes and/or fine-tuned for other tasks.

Acknowledgments

This work is part of the R&D&i ACCESS2MEET (PID2020-116527RB-I0) project financed by MCIN AEI/10.13039/501100011033/, and the ”Intelligent and interactive home care system for the mitigation of the COVID-19 pandemic” project (PRTR-REACT UE) awarded by CAM. CONSEJERÍA DE EDUCACIÓN E INVESTIGACION.

References

Al-Thanyyan, S. S. and A. M. Azmi. 2021. Automated text simplification: A survey. *ACM Computing Surveys (CSUR)*, 54(2):1–36.

Alarcon, R. 2021. Dataset of sentences annotated with complex words and their synonyms to support lexical simplification, March.

Alarcon, R., L. Moreno, and P. Martínez. 2021. Lexical simplification system to improve web accessibility. *IEEE Access*, 9:58755–58767.

Alarcón García, R. 2022. Lexical simplification for the systematic support of cognitive accessibility guidelines. *https://doi.org/10.1145/3471391.3471400.*

Barbu, E., M. T. Martín-Valdivia, E. Martínez-Cámara, and L. A. Urena-López. 2015. Language technologies applied to document simplification for helping autistic people. *Expert Systems with Applications*, 42(12):5076–5086.

Barrio-Cantalejo, I. M., P. Simón-Lorda, M. Melguizo, I. Escalona, M. I. Marijuán, and P. Hernando. 2008. Validación de la escala inflesz para evaluar la legibilidad de los textos dirigidos a pacientes. In *Anales del Sistema Sanitario de Navarra*, volume 31, pages 135–152. SciELO Espana.

Campillos Llanos, L., A. R. Terroba Reinares, S. Zakhir Puig, A. Valverde, and A. Capllonch-Carrión. 2022. Building a comparable corpus and a benchmark for spanish medical text simplification.

Chamovitz, E. and O. Abend. 2022. Cognitive simplification operations improve text simplification.

Cumbicus-Pineda, O. M., I. Gutiérrez-Fandiño, I. Gonzalez-Dios, and A. Soroa. 2022. Noisy channel for automatic text simplification.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

Ermakova, L., I. Ovchinnikov, J. Kamps, D. Nurbakova, S. Araújo, and R. Hannachi. 2022a. Overview of the clef 2022 simpletext task 2: Complexity spotting in scientific abstracts.

Ermakova, L., I. Ovchinnikov, J. Kamps, D. Nurbakova, S. Araújo, and R. Hannachi. 2022b. Overview of the clef 2022 simpletext task 3: Query biased simplification of scientific texts.

Ferrés, D. and H. Saggion. 2022a. Alexsis: a dataset for lexical simplification in spanish. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3582–3594.

Ferrés, D. and H. Saggion. 2022b. ALEXSIS: A dataset for lexical simplification in Spanish. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3582–3594, Marseille, France, June. European Language Resources Association.

Huang, J. and J. Mao. 2022. Assembly models for simpletext task 2: Results from wuhan university research group.

Kauchak, D., J. Apricio, and G. Leroy. 2022. Improving the quality of suggestions for medical text simplification tools. In *AMIA Annual Symposium Proceedings*, volume 2022, page 284. American Medical Informatics Association.

Lewis, M., Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461.*

Lin, C.-Y. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July. Association for Computational Linguistics.

Marimon, M., J. Vivaldi, and N. Bel Rafecas. 2017. Annotation of negation in the iula spanish clinical record corpus. *Blanco E, Morante R, Saurí R, editors. SemBEaR 2017. Computational Semantics Beyond Events and Roles; 2017 Apr 4; Valencia, Spain. Stroudsburg (PA): ACL; 2017. p. 43-52.*

Martin, L., A. Fan, É. de la Clergerie, A. Bordes, and B. Sagot. 2020. Muss: multilingual unsupervised sentence simplification by mining paraphrases. *arXiv preprint arXiv:2005.00352.*

McCarthy, D. and R. Navigli. 2007. Semeval-2007 task 10: English lexical substitution task. In *Proceedings of the fourth international workshop on semantic evaluations (SemEval-2007)*, pages 48–53.

Menta, A. and A. Garcia-Serrano. 2022. Controllable sentence simplification using transfer learning. *Proceedings of the Working Notes of CLEF.*

Monteiro, J., M. Aguiar, and S. Araújo. 2022. Using a pre-trained simplet5 model for text simplification in a limited corpus. *Proceedings of the Working Notes of CLEF.*

Moreno, L., R. Alarcon, and P. Martínez. 2020. Easier system. language resources for cognitive accessibility. In *The 22nd International ACM SIGACCESS Conference on Computers and Accessibility*, pages 1–3.

Mostert, F., A. Sampatsing, M. Spronk, and J. Kamps. 2022. University of amsterdam at the clef 2022 simpletext track. *Proceedings of the Working Notes of CLEF.*

Paetzold, G. and L. Specia. 2016a. Semeval 2016 task 11: Complex word identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 560–569.

Paetzold, G. and L. Specia. 2016b. Unsupervised lexical simplification for non-native speakers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

Plainlanguage. 2017a. Plain english- free guides (co.uk).

Plainlanguage. 2017b. Plain language action and information network (plain).

Radford, A., K. Narasimhan, T. Salimans, I. Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Rubio, A. and P. Martínez. 2022. Hulat-uc3m at simpletext@ clef-2022: Scientific text simplification using bart. *Proceedings of the Working Notes of CLEF.*

Saggion, H., S. Štajner, S. Bott, S. Mille, L. Rello, and B. Drndarevic. 2015. Making it simplext: Implementation and evaluation of a text simplification system for spanish. *ACM Transactions on Accessible Computing (TACCESS)*, 6(4):1–36.

Saggion, H., S. Štajner, D. Ferrés, K. C. Sheang, M. Shardlow, K. North, and M. Zampieri. 2023. Findings of the tsar-2022 shared task on multilingual lexical simplification. *arXiv preprint arXiv:2302.02888.*

Segura-Bedmar, I. and P. Martínez. 2017. Simplifying drug package leaflets written in spanish by using word embedding. *Journal of biomedical semantics*, 8(1):1–9.

Shardlow, M., R. Evans, G. H. Paetzold, and M. Zampieri. 2021. SemEval-2021 task 1:

Lexical complexity prediction. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 1–16, Online, August. Association for Computational Linguistics.

Smith, L. N. 2018. A disciplined approach to neural network hyper-parameters: Part 1–learning rate, batch size, momentum, and weight decay. *arXiv preprint arXiv:1803.09820*.

Talec-Bernard, T. 2022. Is using an ai to simplify a scientific text really worth it. *Proceedings of the Working Notes of CLEF*.

Tang, Y., C. Tran, X. Li, P.-J. Chen, N. Goyal, V. Chaudhary, J. Gu, and A. Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning.

Truică, C.-O., A.-I. Stan, and E.-S. Apostol. 2022. Simplex: a lexical text simplification architecture. *Neural Computing and Applications*, pages 1–16.

UNE. 2018. Une 153101:2018 ex easy to read. guidelines and recommendations for the elaboration of documents.

Wilkens, R., B. Oberle, and A. Todirascu. 2020. Coreference-based text simplification. In *Proceedings of the 1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI)*, pages 93–100.

Xu, W., C. Napoles, E. Pavlick, Q. Chen, and C. Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Yimam, S. M., C. Biemann, S. Malmasi, G. H. Paetzold, L. Specia, S. Štajner, A. Tack, and M. Zampieri. 2018. A report on the complex word identification shared task 2018. *arXiv preprint arXiv:1804.09132*.

Yimam, S. M., S. Stajner, M. Riedl, and C. Biemann. 2017. Multilingual and cross-lingual complex word identification. In *RANLP*, pages 813–822.

Zhu, Z., D. Bernhard, and I. Gurevych. 2010. A monolingual tree-based translation model for sentence simplification. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1353–1361.

Štajner, S., K. C. Sheang, and H. Saggion. 2022. Sentence simplification capabilities of transfer-based models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(11):12172–12180, Jun.

A Prediction examples

Original	**Target**
(1) En todos ellos suscitas gran expectación en la comunidad científica, con la que colaboras en el departamento de medicina deportiva del Instituto Nacional de Educación Física. (In all of them you arouse great expectation in the scientific community, with which you collaborate with the sports medicine department of the National Institute of Physical Education.)	(1) En todos ellos suscitas gran curiosidad en la comunidad científica, con la que colaboras en el departamento de medicina deportiva del Instituto Nacional de Educación Física. (In all of them you arouse great curiosity in the scientific community, with whom you collaborate in the sports medicine department of the National Institute of Physical Education.)
(2) Además de todo lo anterior y como curiosidad, un 3% reconoció que quería conocer a sus nietos que habían nacido durante el confinamiento y a los que todavía no habían podido ver, y un 2% quería celebrar la boda de sus hijos o la Comunión de sus nietos que había sido suspendida. (In addition to all of the above and as a curiosity, 3% recognized that they wanted to meet their grandchildren who had been born during confinement and whom they had not yet been able to see, and 2% wanted to celebrate the wedding of their children or the Communion of their grandchildren that had been suspended.)	(2) Además de todo lo anterior y como curiosidad, un 3% reconoció que quería conocer a sus nietos que habían nacido durante el confinamiento y a los que todavía no habían podido ver, y un 2% quería celebrar la boda de sus hijos o la Comunión de sus nietos que había sido suspendida. (In addition to all of the above and as a curiosity, 3% recognized that they wanted to meet their grandchildren who had been born during isolation and whom they had not yet been able to see, and 2% wanted to celebrate the wedding of their children or the Communion of their grandchildren that had been suspended.)
Prediction	
(1) En todos ellos provocan gran interés en la comunidad científica, con la que colaboras en el departamento de medicina deportiva del Instituto Nacional de Educación Física. (In all of them, they provoke great interest in the scientific community, with which you collaborate in the sports medicine department of the National Institute of Physical Education.)	
(2.1) Aunque su madre falleció con el fin de conseguirlo han sido todos los nietos y mostrar los nietos que nos permitan a los niños y niñas, o incluso con quien lo sufre. (Although his mother passed away in order to get it have been all the grandchildren and show grandchildren that allow us to children, or even with the one who suffers it.) (2.2) Además de todo lo anterior y como curiosidad, un 3% reconoció que quería conocer a sus nietos que habían nacido durante el confinamiento y a los que todavía no habían podido ver, y un 2% quería celebrar la boda de sus hijos o la Comunión de sus nietos que había sido suspendida. (In addition to all of the above and as a curiosity, 3% recognized that they wanted to meet their grandchildren who had been born during isolation and whom they had not yet been able to see, and 2% wanted to celebrate the wedding of their children or the Communion of their grandchildren that had been suspended.)	

Table 9: Examples of model's predictions.