Towards a Knowledge-Aware AI
A. Dimou et al. (Eds.)

doi:10.3233/SSW220020

esT5s: A Spanish Model for Text Summarization

Adrian VOGEL-FERNANDEZ, Pablo CALLEJA and Mariano RICO [1]
Ontology Engineering Group, Universidad Politécnica de Madrid (UPM), Spain

Abstract. Deep Learning models based on the Transformer architecture have revolutionized the state of the art of NLP tasks. As English is the language in which most significant advances are made, languages like Spanish require specific training, but this training has a computational cost so high that only big corporations with servers and GPUs are capable of generating them. This work has explored how to create a model for the Spanish language from a big multilingual model. Specifically, a model aimed at creating text summarization, a very common task in NLP. The results, concerning the quality of the summarization (ROUGE score), point out that these small models, for a specific language, achieve similar results than much bigger models, with a reasonable training in terms of time required and computational power, and are significantly faster at inference.

Keywords. Deep learning, T5, Spanish, Text summarization

1. Introduction

Summarization is a Natural Language Processing task that consists of condensing the most relevant information from a document. This task can be divided into two categories: extractive summarization and abstractive summarization. Extractive summarization consists of identifying and copying the most relevant and useful information pieces (typically sentences) from the original content. In contrast, abstractive summarization requires a deeper understanding of the language to summarize the most relevant content, paraphrasing the original sentences, combining and using synonyms or new words, without losing information and preserving cohesion and coherence [1]. Thus, abstractive summarization is a difficult task in natural language processing.

Currently, the state-of-the-art of language models based on transformers [2] have reached a high level of language comprehension. However, all research is mainly focused on the English language and then applied to other languages. Even important languages such as Spanish, which is the second language spoken in the world, has an enormous

[1]Corresponding Author: Mariano Rico; E-mail:mariano.rico@upm.es.
The authors gratefully acknowledge the computer resources at Artemisa, funded by the European Union ERDF and Comunitat Valenciana as well as the technical support provided by the Instituto de Física Corpuscular, IFIC (CSIC-UV). Also we acknowledge the Universidad Politécnica de Madrid for providing computing resources on Magerit Supercomputer. This work was funded partially by the project Knowledge Spaces (PID2020-118274RB-I00), funded by MCIN/AEI/ 10.13039/501100011033; and project HCommonK (RTC2019-007134-7, funded by MCIN/AEI/ 10.13039/501100011033).

gap in their language models compared to English [3]. For example, the T5 model [4] is one of the best language models, which exploits the features of text-to-text transfer learning and it is usually used for the text summarization task, it is only trained for English language, and there is no Spanish version yet.

Despite this lack of models for non-English languages, there are multilingual models (also including English). This is the case of the multilingual T5 (mT5) [5] which is trained in 101 languages, including English and Spanish among them. These multilingual approaches outperform monolingual models because similar languages have positive transfer between them [6]. However, the effort required to create (and also execute) these multilingual models is very high. Our approach takes advantage of these multilingual models to create a single-language model, much more efficient in training and execution.

Despite there are several summarization models for Spanish publicly available (specifically in HuggingFace), after testing them and, to the best of our knowledge, there is only one model supported by a peer reviewed publication, the so named `NASES` model. However, this model is intended for small texts (up to 512 tokens), achieving low quality values in the tests carried out.

This work presents the creation of a Spanish model for text summarization based on the T5 model, specifically designed for Spanish for text summarization affordable in terms of computational requirements. We want to emphasize that this method is generic and can be applied to **any other** language, not only Spanish.

Our method is capable of creating a summarization model for the Spanish language in less than 1 hour of computing time, using a single GPU and a the T5 multilingual language model. In our case, a NVIDIA V100 16 GB, that many organizations or individuals can afford. As an alternative, Google Colaboratory [2] allows you to train for free neural models using a similar GPU for a maximum of 1 hour for small model versions. The Pro version of Colaboratory allows you to train bigger models (longer execution time) for a small fee (10 USD/month) with even better GPUs.

The simplest version of the resulting models of this work was trained in 49 minutes, and it is capable of summarizing a Spanish text with a reasonable quality: the summary does not have typos and always produces sentences lexically and grammatically correct. The quality of this model is close to the performance of the state-of-the-art model in terms of the ROUGE score [7], the standard quality metric in this field.

Also, we have developed other models with different configurations that obtain a better performance, but with training time above 1 hour. The training of these slightly bigger models is affordable by institutions or individuals since the computation time is below 24 hours (specifically 17 hours) with a single GPU.

Concerning the time required to summarize a Spanish text, in general terms, the bigger the model is, the longer execution time it requires. All resulting models presented in this work are faster than the models in the state of the art. The experiments show summarization time faster (between x6.56 and x1.43).

For the sake of reproducibility, the source code is available [3]. The models will be publicly available at our page in HuggingFace `https://huggingface.co/oeg`.

[2] `https://colab.research.google.com/`

[3] `https://github.com/oeg-upm/t5-spanish-news-summarization`

2. State of the Art

The state of the art in abstractive summarization is the PEGASUS model [8]. It introduces Gap Sentence Prediction (GSP) as a pre-training mechanism. This model achieves the highest ROUGE scores, but it is only for the English language, and it is not feasible to reproduce it for other languages given its computational cost.

The state of the art for Spanish text abstractive summarization is the model published by Hasan et al. [9] based on the mT5 model (this model will be referred as `mT5Hasan`) and `NASES` [10], both reporting the highest ROUGE scores. It is important to mention that `mT5Hasan` ROUGE scores are calculated with the XL-Sum dataset, while `NASES` scores are calculated with a dataset that is not public and authors do not point out any indication on the dataset used. In order to test the model and compare it to the proposed models, it is needed to calculate the ROUGE scores in the test partition of the XL-Sum dataset. Moreover, it is important to notice that the training of the `NASES` model has a maximum input size of 512 tokens, but `mT5Hasan` and the proposed models can handle up to 768 tokens. This makes the proposed models much better to deploy in real life applications since they can handle longer input texts.

The `NASES` model creates a generic Spanish language model (from a Spanish corpus) trained using the GSG (Gap Sentences Generation) [8] technique (used by the Google PEGASUS model). This technique is specifically designed to improve the performance of the summarization task. However, the proposed model does not train any generic model but specializes a reduced version of the multilingual generic model T5. The NASES model, to our best knowledge, is the only Spanish summarization model trained from scratch, without a pretrained language model.

`mT5Hasan` [9] is a multilingual T5 summarization model based on mT5 and trained on all 44 languages of the XL-Sum dataset, which makes the model performance better for each of the individual languages since different languages can have a positive transfer between them [11]. This model is currently the best performing Spanish summarization model, but it is important to notice that this model was trained using 1 million examples of text and summary pairs of 44 languages during 4 days using an 8 GPU cluster. Both state of the art models have been used to compare results with the resulting models of this work.

The evaluation in this work has been performed with the three ROUGE score metrics: ROUGE-1 measures the number of 1-grams that are equal in the reference and the model summary, ROUGE-2 with 2-grams, and ROUGE-L with the longest common subsequence between the model summary and the reference. Although ROUGE has its limitations [12,13].

For the Spanish language, there are not so many publicly available datasets for text summarization. The best datasets available for Spanish, with an evaluated quality, are multilingual. The main multilingual datasets are XL-Sum [9] and MLSUM [14]. XL-Sum contains summaries obtained from BBC news, but its Spanish portion only has 44,413 examples and MLSUM has 290,645 examples obtained from *El País*, a Spanish newspaper. However, only XL-Sum dataset has been evaluated by humans following an answering template [9]. Thus, we consider XL-Sum a higher quality dataset and it will be used for the fine-tuning process.

3. Methodology

This section presents the process of converting the multilingual T5 model into a single language model (pruning) and then, specialize (fine-tune) the model to perform abstractive summarization. Figure 1 shows an overview of the whole process.

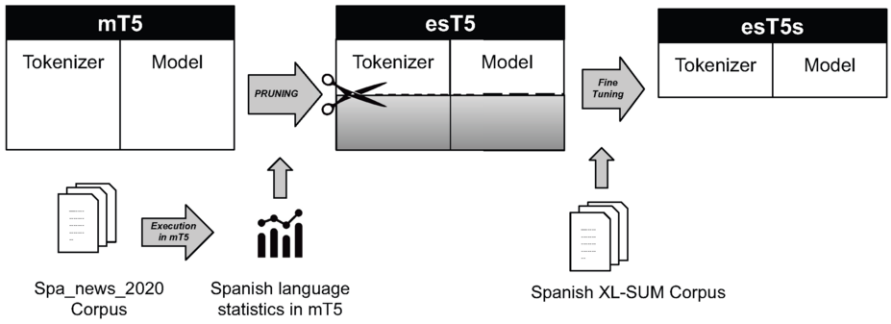

Figure 1. Overview of the method: pruning of mT5 + fine tunning process from the mT5 model.

3.1. Model pruning

The process starts taking the original mT5, a big trained multilingual model for 101 languages (Spanish among them). From this model, we have applied the pruning method [15] but specialized for mT5. This method removes the embeddings of other languages to perform a lossless compression over the multilingual model into a single language.

As mT5 is distributed in different sizes, which refers to the maximum length of the input tokens, we have used the mT5-small and the mT5-base models (512 and 768).

The idea of the pruning method is to reduce the number of parameters of embeddings that are not used and reduce the size of the vocabulary. In the mT5-small model, the 85.2% of the model parameters are used for embeddings and in the mT5-base model, the 65.96% of the model parameters are used for embeddings.

The method needs a reference corpus of the target language to analyze the vocabulary used and how is represented in the multilingual model. For the Spanish language, the corpus used has been `spa_news_2020_1M-sentences` which is a sentence corpus from Leipzig corpora collection[4]. The corpus contains more than 19,000 sentences about Spanish news. Then, the tokenizer of the mT5 is used to tokenize the sentences of the corpus. This process finds out that only 25.9% of the vocabulary of the multilingual tokenizer is used and that the 25,000 most frequent Spanish tokens comprise 99% of the Spanish vocabulary of the corpus.

To prune the original tokenizer of the mT5, the following tokens have been selected: the first tokens from the original tokenizer which include some special tokens and subwords or characters and the 25,000 most frequent Spanish tokens detected in the analysis of the corpus. The resulting tokenizer vocabulary has only 10% of the original tokens.

[4] Available at https://corpora.uni-leipzig.de/es?corpusId=spa_news_2020

To reduce the number of model's parameters, the unused embeddings from non-Spanish languages of the encoder, decoder and the language model head are removed. This process is made by adjusting the length of the embeddings to keep only the Spanish language representation. This reduces the mT5-small model size from 1.2GB to 274MB and the mT5-base model size from 2.3GB to 928MB, which represents the 22% and the 40.34% of the original size respectively. The resulting models are named esT5-small and esT5-base.

3.2. Fine-tuning process

The last process is fine-tuning the language models to perform abstractive summarization. The resulting models will be referenced with the name esT5s. This process is performed using the dataset partitions provided by XL-Sum dataset of training and validation corpora.

First, a model has been trained using the small version (esT5-small) for 3 epochs with a batch size of 8, an AdamW optimizer [16] with a learning rate of 0.0001 into a single GPU. This is the simplest model presented in this work which contains the main targets: it is trained in less than one hour and has achieved enough evaluation results. Also, the small model has been trained for different epochs to study future improvements or limitations of the size of the model. The selection strategy was to keep the model with lowest validation loss for each epoch during the experiments.

Moreover, the esT5-base model has been used to train different models with different epochs in order to explore the possible performance obtained in a single GPU. However, the batch size has been reduced to 4 to fit in the GPU memory. The GPU used for these experiments was a NVIDIA V100 with 16GB of memory.

4. Evaluation

Table 1 presents the ROUGE scores obtained for the resulting models and for the referenced models of mT5Hasan and NASES evaluated with XL-Sum dataset. The XL-Sum test set contains 4,763 examples. The simplest version of esT5s model using the small model language esT5 and 49 minutes of training (3 epochs) achieves considerable good results in comparison with the results provided with mT5Hasan, which needed 96 hours of training. Moreover, different experiments have been performed with the small model language to see the improvements over time with 9 and 15 epochs. The results show a continuous improving trend, meaning that better results can be achieved with more training time but, in this case, this work studies methods to produce results with limited resources.

Also, similar experiments have been performed with the base model esT5 in order to see its performance with similar epochs. The results show an improvement of ROUGE values of these models. The last experiment performed with the esT5-base model has been to do a fine-tuning without time restrictions in order to create the best possible model. In this case, the best ROUGE results have been achieved at epoch 17, closer to the results of the mT5 model. Beyond this number of epochs there is no significant improvement in the model.

The evaluation shows also that the size of the model limits the quality of summarization, as shown in the differences between models trained with the small model or

Model	Dataset	Epochs	Time	ROUGE-1	ROUGE-2	ROUGE-l	max length	exec. time
esT5s-small	XL-Sum	3	49'	22.21	5.28	17.44	512	80 min.
esT5s-small	XL-Sum	9	210'	22.54	5.86	17.74	512	80 min.
esT5s-small	XL-Sum	15	267'	23.30	6.48	18.47	512	80 min.
esT5s-base	XL-Sum	3	180'	23.91	7.16	19.09	768	310 min.
esT5s-base	XL-Sum	9	9h	24.26	7.86	19.68	768	310 min.
esT5s-base	XL-Sum	17	17h	**25.30**	**8.21**	**20.29**	768	310 min.
mT5Hasan	XL-Sum	-	96h*	26.21	8.74	21.06	768	460 min.
NASES	DACSA	-	NA	16.63	2.64	13.32	512	-

Table 1. ROUGE metric for different models. Note: *The mT5Hasan model reports a training time of 4 days (96h) in a cluster of 8 GPUs. Our models use a much modest hardware (1 GPU).

with the base one. The ROUGE scores for NASES model are lower that expected. This is because NASES model was trained using the DACSA dataset. This dataset is not public but, looking at some examples in their work, seems that the dataset is comprised of news and summaries shorter than the ones contained in XL-Sum. As a result, the knowledge obtained during their training process is not transferred as much as expected in the summarization process of the XL-Sum texts. This explains the low ROUGE score achieved in the experiments when compared with the results published by their authors.

Also, an evaluation of the execution time that models need to generate summaries have been performed. This evaluation has focused on the generation of the summaries of test set of XL-Sum (4,763 examples) to calculate the different ROUGE values. The models used in the evaluation has been the esT5s-small (15 epochs), the esT5s-base (17 epochs) and the mT5Hasan. The results are shown in column *Execution time* in table 1.

The mean processing time for the esT5s-small model is 1 hour 20 minutes and 5 hours and 10 minutes for the esT5s-base model. As mT5Hasan model time is 7 hours 40 minutes, we conclude that both models are faster not only for training times but also for summary generation, with a significant difference for the esT5s-small model.

5. Conclusions

The computational effort required to create summarization models is beyond the scope of any ordinary company or research laboratory. The proposed methodology creates models archiving a performance similar to big models with less computation cost.

Specifically, the proposed method outperforms the training time and computation power required to generate an abstractive summarization model. For small companies or research labs it is important to reuse and adapt big models that cannot be generated from scratch. Our models outperforms in execution time: they can summarize the test portion of the XL-Sum dataset in 1 hour 20 minutes (for the small model) and 5 hours 10 minutes (for the base model) while the state of the art model takes 7 hours 40 minutes.

Also it is important to take into consideration model sizes: `mT5Hasan` is 2.3GB in size, while the resulting models esT5s is 274MB in size for the small and 928MB in size for the base. Taking into account the faster computation times and the smaller sizes, these models are much easier to deploy in a web or portable applications.

References

[1] Gupta S, Gupta S. Abstractive summarization: An overview of the state of the art. Expert Systems with Applications. 2019;121:49-65.

[2] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Advances in neural information processing systems. 2017;30.

[3] Gutiérrez-Fandiño A, Armengol-Estapé J, Pàmies M, Llop-Palao J, Silveira-Ocampo J, Carrino CP, et al. Spanish language models. arXiv preprint arXiv:210707253. 2021.

[4] Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. arXiv preprint arXiv:191010683. 2019.

[5] Xue L, Constant N, Roberts A, Kale M, Al-Rfou R, Siddhant A, et al. mT5: A massively multilingual pre-trained text-to-text transformer. arXiv preprint arXiv:201011934. 2020.

[6] Lample G, Conneau A. Cross-lingual language model pretraining. arXiv preprint arXiv:190107291. 2019.

[7] Lin CY. Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out; 2004. p. 74-81.

[8] Zhang J, Zhao Y, Saleh M, Liu P. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In: International Conference on Machine Learning. PMLR; 2020. p. 11328-39.

[9] Hasan T, Bhattacharjee A, Islam MS, Samin K, Li YF, Kang YB, et al. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. arXiv preprint arXiv:210613822. 2021.

[10] Ahuir, Vicent and Hurtado, Lluís-F and González, José Ángel and Segarra, Encarna. NASca and NASes: Two Monolingual Pre-Trained Models for Abstractive Summarization in Catalan and Spanish. Applied Sciences. 2021;11(21). Available from: https://www.mdpi.com/2076-3417/11/21/9872.

[11] Conneau A, Khandelwal K, Goyal N, Chaudhary V, Wenzek G, Guzmán F, et al. Unsupervised Cross-lingual Representation Learning at Scale. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics; 2020. p. 8440-51. Available from: https://aclanthology.org/2020.acl-main.747.

[12] Dorr B, Monz C, Schwartz R, Zajic D. A methodology for extrinsic evaluation of text summarization: does ROUGE correlate? In: Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization; 2005. p. 1-8.

[13] Schluter N. The limits of automatic summarisation according to ROUGE. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers. Valencia, Spain: Association for Computational Linguistics; 2017. p. 41-5.

[14] Scialom T, Dray PA, Lamprier S, Piwowarski B, Staiano J. MLSUM: The multilingual summarization corpus. arXiv preprint arXiv:200414900. 2020.

[15] Abdaoui A, Pradel C, Sigel G. Load What You Need: Smaller Versions of Multilingual BERT. CoRR. 2020;abs/2010.05609. Available from: https://arxiv.org/abs/2010.05609.

[16] Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv:14126980. 2014.