Automatic Punctuation Model for Spanish Live Transcriptions

Mario Pérez-Enríquez, Jose Manuel Masiello-Ruiz, Jose Luis López-Cuadrado, Israel González-Carrasco, Paloma Martínez, Belén Ruiz-Mezcua
Computer Science Department
Universidad Carlos III de Madrid
Av. de la Universidad, 30, Leganes, Madrid (Spain)
{maripere,jllopez,igcarras,pmf,brm}@inf.uc3m.es
jmasiell@eco.uc3m.es

Abstract

With the widespread adoption of automatic transcription tools, acquiring speech transcriptions within seconds has become a reality. Nonetheless, many of these tools yield unpunctuated outputs, potentially incurring additional costs. This paper presents a novel approach to integrating punctuation into the transcriptions generated by such automatic tools, specifically focusing on Spanish-speaking contexts. Leveraging the RoBERTa-bne model pre-trained with data from the Spanish National Library, our training proposal is augmented with additional corpora to enhance performance on less common punctuation marks, such as question marks. Also, the proposed model has been trained through fine-tuning pre-trained models, involving adjustments for token classification and using SoftMax to identify the highest probability token. The proposed model obtains promising results when compared with other Spanish reference paper models. Ultimately, this model aims to facilitate punctuation on live transcriptions seamlessly and accurately. The proposed model will be applied to a real-case education project to improve the readability of the transcriptions.

Keywords: Large Language models, Punctuation marks, BERT

1. Introduction

Automatic Speech Recognition (ASR) systems play a crucial role in facilitating technology interaction, enabling effective voice commands and transcription services. However, the challenge that currently confronts these systems, particularly in Spanish speech, is their tendency to produce unpunctuated text as output.

This paper focuses on the vital task of punctuation restoration in the Spanish language as an essential issue to enhance ASR system performance and comprehension. Unpunctuated text, whether transcribed speech, dictated notes, or spontaneous conversations, can be challenging to decipher, especially in lengthy or complex passages (Ninčević and Zanchi, 2012). Punctuation marks, including periods, commas, question marks, and exclamation points, serve as crucial cues for sentence boundaries, intonation, and context, significantly aiding in the comprehension of spoken Spanish.

This endeavour gains substantial importance in scenarios where ASR systems are deployed to transcribe Spanish audio data for diverse applications, including transcription services, voice assistants, and accessibility tools. Accurate punctuation restoration not only enhances the readability and clarity of Spanish transcriptions but also supports downstream natural language processing tasks such as language understanding, sentiment analysis, and summarization (Alam et al., 2015).

In this paper, we delve into the effectiveness of Transformer-based models for the task of punctuation restoration in the Spanish language. Our goal is to develop models capable of inferring appropriate punctuation marks in unpunctuated Spanish text, thereby improving the overall usability and utility of ASR systems for Spanish speakers. Through systematic experimentation and analysis, we aim to uncover the nuances, challenges, and opportunities inherent in this task within the context of the Spanish language.

2. Related Work

Although advances have been made, not all ASR provides automatic punctuation, and those that do, like Google, do not cover all languages, as is the case of Spanish one.

Earlier studies on punctuating speech transcriptions in English restored punctuation using acoustic, lexical, signal processing, or a combination of these features (Gravano et al., 2009). More recent approaches combined pre-trained word vectors with Convolutional Neural Networks (CNN) using transcripts from TED talks (Che et al., 2016). Further studies improved these approaches by ensembling DNN, TBRNN and BLSTM-CRF models to a single DNN student model (Yi et al., 2017) and combining convolutional and BNN models (Vinícius Treviso et al., 2016). (Schweter and Ahmed, 2019) tested LSTM, Bi-LSTM and CNN architectures trained with EuroParl (English and German) and SETimes (several languages not including Spanish) corpus and compared with OpenNLP. They tuned the En-

LREC-COLING 2024, pages 1953–1958

glish hyperparameters and used them for other languages, concluding that these models can be considered language-independent. The authors also tested their proposal on various low NLP resource languages such as Greek and Romanian.

In the last years, several approaches have been based on pre-trained Transformer models, such as in the ASR output of the medical domain using BioBERT, a variation of BERT for medical corpora (Lee et al., 2020). Based on these pre-trained models, multilingual approaches have been tested. For example, winners of the SEPP-NLG Shared Task use a multilingual RoBERTa model for French, German, English, and Italian, obtaining state-of-the-art results for these languages (Guhr et al., 2021). Later this model was adapted to the Dutch language with good results (Vandeghinste and Guhr, 2023). Furthermore, BERT-based models have also been used for low-resource languages such as Bangla, obtaining promising results (Alam et al., 2020).

However, most available resources for training punctuation models are in English. In Spanish (González-Docasal et al., 2021) applies a BERT-based approach based on BETO for Spanish (Cañete et al., 2020), Berteus for Basque (Agerri et al., 2020) and IXAM-Bert for Spanish and Basque (Otegi et al., 2020). Also (Pan et al., 2023) also test several BERT-Based models in Spanish and Portuguese. They obtained better results for the Spanish language with the BETO for labels combining punctuation and capitalization.

This paper shows a Spanish-specific approach based on MarIA (Fandiño et al., 2022), a language model based on a 570GB dataset from the Spanish National Library that it is pre-trained with a huge amount of Spanish texts.

3. Methodology

3.1. Data analysis

In the realm of Natural Language Processing (NLP), the initial phase begins with evaluating and processing training data. In the context of predicting punctuation marks, it's crucial to carefully vet and clean the corpora to prevent the model from learning incorrect punctuation rules.

Multiple unlabeled Spanish corpora were employed to examine the impact of different corpora on punctuation prediction. This analysis helps us understand how corpus selection affects the model's punctuation restoration performance. The corpora used in this work are:

(1) Europarl: (Koehn, 2011) This dataset comprises the Spanish segment of the European Parliament records extracted from its proceedings. It consists of 2,174,141 individual examples. This corpus was sourced from the Opus corpora (Tiedemann, 2012).

(2) SQuAD: (Carrino et al., 2019) The SQuAD corpus has been translated into Spanish, yielding a collection of 98,165 examples. It includes questions and their corresponding answers (Carrino et al., 2019).

(3)Mintzai: (Vicomtech, 2020) This corpus is a Basque-Spanish parallel dataset derived from the parliamentary sessions of the Basque government, encompassing a total of 56,886 examples (Etchegoyhen et al., 2020).

Given the utilization of unlabeled corpora, a preprocessing step was employed to align the training data with the input data used for predictions. ASR output is typically presented in lowercase without punctuation marks. Consequently, the utilized corpora were converted to lowercase and removed all punctuation marks. Furthermore, during the cleaning process, meticulous attention was paid to eradicating artefacts from the corpora, including HTML elements, emails, partial parentheses (beginning or ending, but not both), hidden characters, symbols, numbers, and dates.

The corpora were then treated as a token classification task, where each token was associated with the label, indicating the presence (1 to n labels) or absence (0) of punctuation. For instance, a sentence like "Hello my name is Mario" would be represented as tokens and labels: "[Hello, my, name, is, Mario] [2, 0, 0, 0, 1]," where the number 1 represents a period and number 2 a comma.

Subsequently, statistical analysis was conducted on the corpora to assess their balance and gain insights into prediction scores post-training. An imbalanced corpora can profoundly impact a model's performance, emphasizing the importance of achieving balance, even though it may not be entirely feasible when working with unlabeled raw corpora.

Hence, the datasets created for this study are organized as follows: **Dataset 1** (comprising solely the Europarl corpus), **Dataset 2** (combining the Europarl and SQuAD corpora) and **Dataset 3** (comprising the Mintzai corpus exclusively).

These consolidated datasets were partitioned into training and testing sets with an 80/20 split ratio.

As Table 1 illustrates, the dataset's imbalance is rooted in our languages' inherent usage patterns of punctuation marks. Periods and commas, for instance, are highly prevalent, while colons and semicolons are used more sparingly. To rectify this disparity, we chose to merge different corpora, including Europarl and SQuAD, with the primary objective of augmenting the occurrence of question marks. This consolidation led to an almost threefold increase in the number of instances.

Label	Dataset 1	Dataset 2	Dataset 3
.	1.913.643	1.943.147	172.755
,	2.671.863	2.917.335	348.554
!	7204	7623	3526
¡	6524	7363	3031
?	62.333	199.113	9639
¿	60.745	192.075	9431
:	72.978	78.243	8719
;	39.886	43.948	6475
Total	**4.835.176**	**5.103.286**	**562.130**

Table 1: Comparison between datasets

3.2. Approaches

Our model is trained through fine-tuning pre-trained models, involving adjustments for token classification.

Token classification resembles named entity recognition, categorizing words into specific tokens with labels like location, person, or organization. Therefore, each word receives a token in this task, forming an input array with labels. During prediction, the model returns the token for each word using SoftMax to identify the highest probability token.

We experimented with various pre-trained models for this task using Spanish-only models, particularly:

RoBERTa-bne-base (Fandiño et al., 2022): This variant of RoBERTa (Liu et al., 2019) is a highly optimized model derived from BERT, featuring a nearly identical architecture but with distinct training characteristics. Specifically, it has been trained as a masked language model, where 15% of tokens are masked. This encoder-only model is exclusively trained on Spanish corpora sourced from the National Library of Spain, with training conducted at the Barcelona Supercomputing Center. Given its RoBERTa-base architecture, it boasts a hidden size of 768, an intermediate size of 3072, 12 attention heads, and 12 hidden layers, with a token vocabulary comprising 50,262 tokens.

RoBERTa-bne-large(Fandiño et al., 2022): Similar to RoBERTa-bne-base model, RoBERTa-bne-large is a larger version, featuring a hidden size of 1024, an intermediate size of 4096, 16 attention heads, and 24 hidden layers.

3.3. Evaluation

To train the models, we utilized the following hyperparameters: AdamWeightDecay with a learning rate set at 2.6e-05, a decay rate of 0.01, and no warmup steps. In the case of the base version, a batch size of 16 was employed, while it was reduced to 2 for the large version. These hyperparameters align with state-of-the-art configurations for each model, as per Hugging Face guidelines, while adhering to established standards in other aspects.

To prevent overfitting, a training strategy consisting of two epochs was employed, with the number of training steps set equal to the product of the dataset size and the number of epochs. Throughout the training process, we utilized the internal loss of the models.

The hardware configuration utilized included a 12th Gen Intel 12700k processor with 32GB of RAM, complemented by a dedicated Nvidia 3060 GPU featuring 12GB of video memory. Notably, the dedicated GPU was fully utilized in terms of computation and memory usage during training. Additionally, mixed floating-point 16 precision was applied to expedite training. Despite this high-end hardware configuration, models such as RoBERTa-bne-large, when paired with datasets 1 and 2, required a substantial 30 hours per epoch, while smaller models completed an epoch in 3 hours.

Before we delve into the experimental results, we must highlight that not all punctuation marks are of equal significance. The period is critical in marking sentence boundaries and introducing capitalization for sentence beginnings. Close in importance are the comma and question mark. Conversely, punctuation marks like colons and semicolons bear less weight overall. Therefore, in the context of our results, models that perform better on these pivotal punctuation marks demonstrate superior overall performance.

Table 2 underscores the model's proficiency in accurately predicting instances where punctuation is unnecessary, labelled as "0". Additionally, it highlights the model's ability to predict periods and commas, which are fundamental punctuation marks. Notably, including the SQuAD dataset has significantly improved the model's performance, particularly in predicting question marks. Consequently, the model paired with dataset 2 yields superior overall results.

Upon comparing the base and large versions of RoBERTa, we notice comparable prediction accuracy, as outlined in Table 4. Given the significant training time demanded by the larger model and the similar results obtained, we focus our evaluations solely on the RoBERTa-base model for efficiency in both training and inference. As Table 3 illustrates, we notice remarkably similar overall results by comparing datasets 2 and 3. However, a significant distinction becomes apparent, with an 8-point advantage favouring dataset 2 in the period category, signifying its clear advantage in sentence segmentation. Additionally, the dataset 2 model demon-

Label	DS1 Precision	DS1 Recall	DS1 F1	DS2 Precision	DS2 Recall	DS2 F1
0	0.99	0.99	**0.99**	0.99	0.99	**0.99**
.	0.96	0.97	**0.97**	0.95	0.97	0.96
,	0.86	0.85	**0.86**	0.83	0.82	0.83
!	0.41	0.09	0.15	0.43	0.10	**0.16**
¡	0.51	0.07	**0.13**	0.51	0.07	**0.13**
?	0.63	0.52	0.57	0.95	0.93	**0.94**
¿	0.61	0.51	0.56	0.96	0.92	**0.94**
:	0.71	0.57	0.63	0.71	0.58	**0.64**
;	0.53	0.26	**0.35**	0.51	0.24	0.33

Table 2: Metrics comparison between datasets 1 (DS1) & 2 (DS2) testing RoBERTa base model

Label	DS2 Precision	DS2 Recall	DS2 F1	DS3 Precision	DS3 Recall	DS3 F1
0	0.99	0.99	**0.99**	0.98	0.99	0.98
.	0.95	0.97	**0.96**	0.88	0.89	0.88
,	0.83	0.82	**0.83**	0.82	0.82	0.82
!	0.43	0.10	0.16	0.45	0.25	**0.32**
¡	0.51	0.07	0.13	0.40	0.16	**0.22**
?	0.95	0.93	**0.94**	0.75	0.55	0.63
¿	0.96	0.92	**0.94**	0.72	0.50	0.60
:	0.71	0.58	**0.64**	0.63	0.59	0.61
;	0.51	0.24	0.33	0.67	0.33	**0.44**

Table 3: Comparison between dataset 2 (DS2) and dataset 3 (DS3) testing RoBERTa base

Label	RoBERTa base F1	RoBERTa large F1
0	**0.99**	**0.99**
.	**0.96**	**0.96**
,	**0.83**	0.82
!	**0.16**	0.15
¡	**0.13**	0.11
?	**0.94**	**0.94**
¿	**0.94**	**0.94**
:	**0.64**	**0.64**
;	**0.33**	**0.33**

Table 4: F1 comparison with RoBERTa base & large models on dataset 2

Label	DS2 F1	DS3 F1	AutoPunct BERT+sound	AutoPunct BERT
0	**0.99**	0.98	-	-
.	**0.96**	0.88	0.87	0.86
,	**0.83**	0.82	0.80	0.80
!	0.16	**0.32**	0.03	0.13
¡	0.13	0.22	0.16	**0.24**
?	**0.94**	0.63	0.60	0.56
¿	**0.94**	0.60	0.65	0.64
:	**0.64**	0.61	0.47	0.48
;	0.33	**0.44**	0.31	0.33

Table 5: Comparison between dataset 2 (DS2), dataset 3 (DS3) and AutoPunct system with BERT+sound and only BERT

strates significantly improved results for question marks while exhibiting lower performance in the case of exclamation marks and semicolons. Of particular interest is the Mintzai corpus's superior performance in the prediction of exclamation marks, a notable outcome considering its higher frequency of occurrence for these labels. Finally, in Table 5, we present a comparison between our models trained on dataset 2 and 3, the AutoPunct system utilizing BERT and considering word silences, and the AutoPunct model with only BERT (González-Docasal et al., 2021) . Our models focus solely on punctuation prediction, simplifying the task. AutoPunct employs a combined architecture comprising BRNN, BERT, and word silence distributions for its training process, along with handling capitalization, making its training process more complex and involving more labels to predict. Despite AutoPunct utilizing word silences, which intuitively would improve performance along with BERT, our proposal consistently outperforms it across nearly all punctuation labels, including question marks and sentence segmentation, in both BERT and BERT with word silence settings. This superior performance could be attributed to our streamlined data

processing and training with fewer labels, which allows the model to train more effectively.

4. Conclusions and future work

This paper has presented a novel approach focused on Spanish to integrate punctuation into text transcriptions. The proposed approach is based on the RoBERTa-bne model pre-trained with data from the Spanish National Library. Our experimentation with the transformer-based model for punctuation restoration in Spanish has yielded promising results, improving the performance of reference paper models in Spanish. The results in predicting question marks and sentence segmentation are remarkable.

Addressing the issue of unbalanced corpora, we emphasize the critical need for validated, diverse datasets in future research. Merging datasets has yielded significant improvements in punctuation restoration, but the reliance on curated corpora is paramount for ensuring model reliability across varied linguistic contexts. This necessity becomes particularly evident in the case of the presented model, which excels in formal speech scenarios but, in future work, can be improved with specific training in colloquial and chaotic conversations.

Furthermore, continued efforts in corpus merging hold promise for comprehensively addressing underrepresented punctuation signs and advancing the field. Considering the trained model utilizes the transformer encoder-only approach, investigating the effectiveness of generative large language models is the logical continuation. This analysis could provide valuable insights into the strengths and limitations of different model architectures.

Although our training utilized only text-based corpora, future works could enhance model prediction by integrating audio cues such as pauses and variations in tone in our training pipeline. This approach could capture additional contextual information that may not be evident in written text alone. However, implementing such a strategy would necessitate modifying existing corpora to accommodate the incorporation of auditory features.

Finally, the proposed model will be applied to a real-case education project to improve the readability of the transcriptions. In this scenario, time consumption must be low, and accuracy is very important.

These advancements will further advance ASR technology for Spanish speakers, ultimately enhancing accuracy and usability in speech recognition applications.

5. Acknowledgements

This work is part of the grants TED2021-132182A-I00 and PID2020-116527RB-I0 funded by MCIN/AEI/10.13039/501100011033 and by European Union NextGenerationEU/PRTR.

6. Bibliographical References

Rodrigo Agerri, Iñaki San Vicente, Jon Ander Campos, Ander Barrena, Xabier Saralegi, Aitor Soroa, and Eneko Agirre. 2020. Give your text representation models some love: the case for basque. In *Proceedings of the 12th International Conference on Language Resources and Evaluation*.

Firoj Alam, Bernardo Magnini, and Roberto Zanoli. 2015. Comparing named entity recognition on transcriptions and written texts. *Harmonization and Development of Resources and Tools for Italian Natural Language Processing within the PARLI Project*, pages 71–89.

Tanvirul Alam, Akib Khan, and Firoj Alam. 2020. Punctuation restoration using transformer models for high-and low-resource languages. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, pages 132–142.

Casimiro Pio Carrino, Marta R Costa-Jussà, and José AR Fonollosa. 2019. Automatic spanish translation of the squad dataset for multilingual question answering. *arXiv preprint arXiv:1912.05200*.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020*.

Xiaoyin Che, Cheng Wang, Haojin Yang, and Christoph Meinel. 2016. Punctuation prediction for unsegmented transcript based on word vector. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 654–658.

Thierry Etchegoyhen, Haritz Arzelus, Harritxu Gete, Aitor Alvarez, Inma Hernaez, Eva Navas, Ander González-Docasal, Jaime Osácar, Edson Benites, Igor Ellakuria, et al. 2020. Mintzai: Sistemas de aprendizaje profundo e2e para traducción automática del habla. *Procesamiento del Lenguaje Natural*, 65:97–100.

Asier Gutiérrez Fandiño, Jordi Armengol Estapé, Marc Pàmies, Joan Llop Palao, Joaquin Silveira

Ocampo, Casimiro Pio Carrino, Carme Armentano Oller, Carlos Rodriguez Penagos, Aitor Gonzalez Agirre, and Marta Villegas. 2022. Maria: Spanish language models. *Procesamiento del Lenguaje Natural*, 68.

Ander González-Docasal, Aitor García-Pablos, Haritz Arzelus, and Aitor Álvarez. 2021. Autopunct: A bert-based automatic punctuation and capitalisation system for spanish and basque. *Procesamiento del Lenguaje Natural*, 67:59–68.

Agustin Gravano, Martin Jansche, and Michiel Bacchiani. 2009. Restoring punctuation and capitalization in transcribed speech. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4741–4744. IEEE.

Oliver Guhr, Anne-Kathrin Schumann, Frank Bahrmann, and Hans-Joachim Böhme. 2021. Fullstop: Multilingual deep models for punctuation prediction. In *Swiss Text Analytics Conference*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Siniša Ninčević and Željka Zanchi. 2012. The importance of correct punctuation and capitalisation. *Transactions on Maritime Science*, 1(01):47–57.

Arantxa Otegi, Aitor Agirre, Jon Ander Campos, Aitor Soroa, and Eneko Agirre. 2020. Conversational question answering in low resource scenarios: A dataset and case study for basque. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 436–442.

Ronghao Pan, José Antonio García-Díaz, and Rafael Valencia-García. 2023. Evaluation of transformer-based models for punctuation and capitalization restoration in spanish and portuguese. In *International Conference on Applications of Natural Language to Information Systems*, pages 243–256. Springer.

Stefan Schweter and Sajawel Ahmed. 2019. Deepeos: General-purpose neural networks for sentence boundary detection. In *Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019)*.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Vincent Vandeghinste and Oliver Guhr. 2023. Fullstop: Punctuation and segmentation prediction for dutch with transformers. *ArXiv*, abs/2301.03319.

Marcos Vinícius Treviso, Christopher Shulby, and Sandra Maria Aluísio. 2016. Sentence segmentation in narrative transcripts from neuropsychological tests using recurrent convolutional neural networks. *arXiv e-prints*, pages arXiv–1610.

Jiangyan Yi, Jianhua Tao, Zhengqi Wen, Ya Li, et al. 2017. Distilling knowledge from an ensemble of models for punctuation prediction. In *Interspeech*, pages 2779–2783.

7. Language Resource References

Carrino et al. 2019. *Automatic Spanish Translation of the SQuAD Dataset for Multilingual Question Answering*. distributed via Hugging Face: dataset squad_es.

Philipp Koehn. 2011. *Europarl v7*. OPUS. distributed via OPUS The Open Parallel Corpus, 7.0.

Vicomtech. 2020. *mintzai-sl*. Fundacion Centro de Tecnologias de Interaccion Visual y Comunicaciones Vicomtech. https://github.com/Vicomtech/mintzai-ST.