Procesamiento del Lenguaje Natural, Revista nº 70, marzo de 2023, pp. 27-38 recibido 29-11-2022 revisado 31-01-2023 aceptado 10-02-2023

Evaluation of transformer-based models for punctuation and capitalization restoration in Catalan and Galician

Evaluación de modelos basados en Transformers para el sistema de recuperación de puntuación y mayúsculas en Catalán y Gallego

Ronghao Pan[1], **José Antonio García-Díaz**[1],
Pedro José Vivancos-Vicente[2], **Rafael Valencia-García**[1]
[1]Facultad de Informática, Universidad de Murcia, Campus de Espinardo, 30100 Murcia, España
[2]VÓCALI Sistemas Inteligentes S.L., Parque Científico de Murcia,
Carretera de Madrid km 388. Complejo de Espinardo, 30100 Murcia, España
{ronghao.pan, joseantonio.garcia8, valencia}@um.es
pedro.vivancos@vocali.net

Abstract: In recent years, the performance of Automatic Speech Recognition systems (ASR) has increased considerably due to new deep learning methods. However, the raw output of an ASR system consists of a sequence of words without capital letters and punctuation marks. Therefore, a capitalization and punctuation restoration system are one of the most important post-processes of ASR to improve readability and to enable the subsequent use of these results in other NLP models. Most models focus solely on English punctuation resolution, and recently new models of Spanish punctuation restoration have emerged. However, none focus on capitalization and punctuation restoration in Galician and Catalan. In this sense, we propose a system for capitalization and punctuation restoration based on Transformers models for Catalan and Galician. Both models perform very well, with an overall performance of 90.2% for Galician and 90.86% for Catalan, and have the ability to identify proper names, country names, and organizations for uppercase restoration.
Keywords: Automatic Speech Recognition, Transformers, Punctuation Restoration, Capitalization Restoration, Catalan, Galician.

Resumen: En los últimos años, el rendimiento de sistemas de Reconocimiento Automático del habla ha aumentado considerablemente gracias a nuevos métodos de deep learning. Sin embargo, la salida bruta de estos sistemas consiste en secuencias de palabras sin mayúsculas ni signos de puntuación. Recuperar esta información mejora la legibilidad y permite su posterior uso en otros modelos de PLN. La mayoría de las soluciones existentes se centran únicamente en inglés; aunque recientemente han surgido nuevos modelos de restauración de la puntuación en español. Sin embargo, ninguno se centra en gallego y catalán. En este sentido, proponemos un sistema de restauración de mayúsculas y puntuación basado en modelos Transformers para estos idiomas. Ambos modelos tienen un rendimiento muy bueno: 90,2% para el gallego y 90,86% para el catalán. Además, también tienen la capacidad de identificar nombres propios, nombres de países y organizaciones para la restauración de mayúsculas.
Palabras clave: Reconocimiento Automático del Habla, Transformers, Recuperación de puntuación, Recuperación de mayúsculas, Catalán, Gallego.

ISSN 1135-5948 DOI 10.26342/2023-70-2

1 Introduction

In recent years, the performance of Automatic Speech Recognition (ASR) systems has increased significantly due to recent advances in deep learning methods. The improved performance of ASR has enabled the development of a wide range of applications in various fields, such as voice assistants, customer care, and healthcare, making it increasingly important in our daily lives. However, the ASR system often generates a stream of unpunctuated words as output, which noticeably reduces its overall readability and comprehensibility (Jones et al., 2003). Moreover, the most advanced Natural Language Processing (NLP) models are mostly trained with punctuated text, such as Wikipedia texts (Cañete et al., 2020). Thus, unpunctuated texts reduce the possibility of being used in these models (Peitz et al., 2011), because the lack of punctuation would seriously degrade the performance of the language models. For example, in Basili et al. (2015) there is a performance difference of over 10% when the models are trained with newspaper texts and tested with unpunctuated transcripts for the entity recognition system.

Recent developments in transformer-based pre-trained models have proven to be successful in many NLP tasks across different languages, and these models have been explored very litter for the punctuation restoration problem. In this work, we present a model of punctuation and capitalization restoration for Catalan and Galician. Both models are composed of a transformer architecture that uses an adapted pre-trained language model as a starting point for transferring knowledge to a specific task, as in this case, the identification of capital letters and punctuation marks. Currently, for Galician and Catalan there are different monolinguals and multilingual models based on BERT or RoBERTa, with different performances. Thus, this work also analyses the behavior of different pre-trained models for the task of automatic restoration of punctuation and capital letter.

This paper is structured as follows: Section 2 presents an overview of the state of the art of punctuation and capitalization restoration system. In Section 3, materials and methods are presented and described in detail. Section 4 presents the performed experiment and the results obtained by different pre-trained language models. In Section 5, error analysis is conducted with a few representative examples. Finally, in Section 6 the conclusions and future work are discussed.

2 Related work

Nowadays, the task of automatically recovering capitalization and punctuation marks has been extensively studied in many systems. These approaches can be broadly divided into three categories in terms of applied features (Yi et al., 2020): those using prosody features derived from acoustic information, those using lexical features, and the combination of the previous two features-based methods.

In recent years, the problem of punctuation retrieval has been addressed with different approaches, from the use of deep learning algorithms, such as Che et al. (2016), which used pre-trained word embedding to train feedforward deep neural network and Convolutional Neural Network, to architectures based on Recurrent Neural Networks (RNNs) combined with Conditional Random Fields (CRF) and pre-trained vectors (Tilk and Alumäe, 2016). Tilk and Alumäe (2016) used RNNs with an attention mechanism to improve performance over Deep Neural Networks and CNN models. Recent advances in transformer-based pre-trained models have proven to be successful in many NLP tasks, so new transformer-based approaches based on BERT-type architectures have emerged (Courtland, Faulkner, and McElvain, 2020), which have been shown to achieve values of up to 83.9% on the F1-score in the well-known and reference IWSLT 2012 dataset (Federico et al., 2012). Another study (Alam, Khan, and Alam, 2020) explored different transformer-based models for both English and Bangla using different pre-trained models and used bidirectional LSTM (BiLSTM) on top of the pre-trained transformer network. However, most of these models mainly focus on solving the problem of punctuation in the three most common punctuation marks, such as period (.), comma (,), and question (?) in English.

Recently, new models of punctuation restoration in Spanish have emerged, such as punctuation restoration in Spanish customer support transcripts using transfer learning (Zhu et al., 2022), and a BERT-based automatic punctuation and capitalization system for Spanish and Basque (González-Docasal et

al., 2021), but there is no adapted model for Catalan and Galician.

The system presented in this paper addresses 5 different punctuation marks for Catalan and Galician, which are described in Section 3. Both models are composed of a transformer architecture but, unlike prior works that solely studied one architecture (BERT), we experiment with different pre-trained models based on BERT, and RoBERTa (see Section 3.3), thus analyzing the monolingual and multilingual models used. We also propose an augmentation scheme that improves performance. Our augmentation is closed related to the augmentation techniques proposed in (Alam, Khan, and Alam, 2020) where authors consider *unknow* word substitution, random insertion, and random deletion. We propose a different version of it in our approach, which uses the back-translation technique for the substitution task described in Section 3.2.

3 Materials and methods

We frame the restoration of punctuation and capitalization as a sequence token classification problem in which the model predicts the punctuation marks that each word in the input text may have. The main advantage of using this approach is that no dependency information is lost and that, given a word in a sentence or sequence, the model can use which word is on the right or left to predict its punctuation mark.

Instead of covering all possible punctuation marks in Catalan and Galician, we only include 5 types of target punctuation that are commonly used and are important to improve the readability of the transcription: period (.), comma (,), question (?), exclamation (!), and colon (:). More specifically, the model predicts which punctuation mark appears next to a given token. However, our model also has the ability to restore capital letters, so for each type of punctuation two labels are added, e.g. for the comma, we have two labels: '`,u`' indicates that the token is of uppercase type and has the comma, and '`,l`' denotes that the token is of lowercase type. Therefore, there are a total of 12 classes that the model needs to predict: '`l`' (lower case), '`u`' (upper case), '`?u`' (upper case with a question), '`?l`' (lower case with a question), '`!u`' (upper case with an exclamation), '`!l`' (lower case with an exclamation), '`,u`' (upper case with a comma), '`,l`' (lower case with a comma), '`.u`' (upper case with a period), '`.l`' (lower case with a period), '`:u`' (upper case with a colon) and '`:l`' (lower case with a colon).

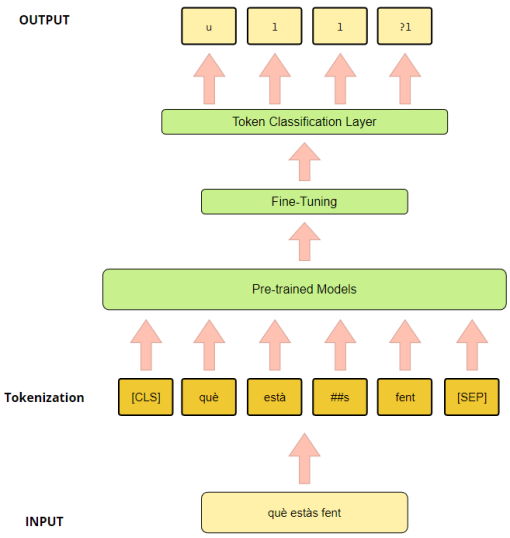


Figure 1: Capitalization and punctuation restoration model structure.

In Figure 1, we report the punctuation restoration model structure. Briefly, it can be described as follows. First, we pre-process the input data through by the tokenization process (see Section 3.4). Second, we use several pre-trained transformer-based models as a starting point, as this reduces computational costs and allows us to use the latest generation models without having to train one from scratch (see Section 3.3). Third, we train the pre-trained models to fit a token classification task, transferring the knowledge from the pre-trained model. This stage is also known as fine-tuning. Finally, the best models from each pre-trained model are evaluated on the test dataset. As can be seen in Figure 1, the input sentence "què estàs fent" does not have any punctuation, and the model predicts that the word "què" is uppercase type, the word "estàs" is lowercase type, and the word "fent" is lowercase and has a question mark after it to produce the output sentence "Què estàs fent?".

We split the dataset into three parts (as shown in Table 2) to evaluate model accuracy: the training set (60%), the validation set (20%), and the test set (20%). The performance of each model is evaluated on the test set after it has been finetuned on vari-

ous combinations of sources and training processes.

3.1 Dataset

We use OpusParaCrawl (Bañón et al., 2020) dataset for Catalan and Galician capitalization and punctuation restoration, which consists of parallel corpora from Web Crawls collected in the ParaCrawl project. These datasets contain 42 languages, and 43 bitexts with a total number of 3.13G sentence fragments by crawling hundreds of thousands of websites, using open-source tools. Usually, parallel corpora are essential for building high-quality machine translation systems and have found uses in many other natural language applications, such as paraphrases learning (Bannard and Callison-Burch, 2005). In this case, the main reason for using this database for capitalization and punctuation restoration is that the texts are already divided into sentences and have all the punctuation marks for each language. After the cleaning, extraction and selection process, a total of 50,000 sentences are selected for Catalan and Galician with 1,136,708 and 1,062,565 words respectively. During the selection process, less common punctuation marks, such as question marks, exclamation marks, and colons, have been preferentially selected to balance the dataset.

3.2 Data augmentation

For this study, we propose an augmentation method inspired by the study of Alam, Khan, and Alam (2020), as discussed above. By training the models with well-trained and correctly punctuated datasets, the trained models lack the knowledge of the typical errors made by an ASR system. Therefore, our augmentation method relies on the type of errors made by the ASR during recognition using the random insertion and deletion technique, and back-translation of sentences to augment the training set.

Currently, most synonym substitution models focus mainly on the English language, so we have chosen the back-translation technique for our synonym substitution task. This technique consists of first translating the text into a given language and then back-translating it into the source language. Thus, when translating the text from one language to another, the translation models usually replace some words with their synonyms or create a new sentence with the same meaning. In this study, the "Helsinki-NLP" models have been used to translate an original text in Catalan or Galician into English and then back-translate it into the source language. In contrast to Alam, Khan, and Alam (2020), we consider all three techniques to have the same prevalence. With this in mind, to process the input text with augmentation, we use three adjustable parameters with the same value (0.33) to control the probability of each of them. Table 1 shows an example of each technique, where Text 1 corresponds to the back-translation technique, Text 2 to the random insertion, and Text 3 to the random detection technique.

3.3 Pre-trained models

Transfer learning and pre-trained transformer-based models have been popular in computer vision and widely adopted for various NLP tasks since the introduction of BERT (Devlin et al., 2019). For Catalan and Galician, available pre-trained resources include multilingual models such as mBERT (Devlin et al., 2019), DistilmBERT (Sanh et al., 2019), and XLM-RoBERTa (Conneau et al., 2019), as well as monolingual such as BERTa for Catalan (Armengol-Estapé et al., 2021), and Bertinho for Galician (David Vilares, 2021). In our experiment, we used such pre-trained language models for capitalization and punctuation restoration tasks. Moreover, we briefly discuss the monolingual language of Catalan and Galician, and the multilingual models used in this study. The following models are used:

- **Bertinho**: It is a robust monolingual model based on the BERT for Galician. It has two versions created with 6 and 12 transformer layers, respectively, and trained with a limited amount of resources (around 45 million words on a single 24GB GPU) (David Vilares, 2021). For our experiment, we have used the 12 transformers layers version.
- **BERTa**: It is a transformer-based masked language model based on the RoBERTa for the Catalan language. It has been trained on a medium size corpus collected from web crawling and public corpora (Armengol-Estapé et al., 2021). The training corpus consists of several corpora gathered from web

	Text	Data augmentation
1	Els nens d' aquesta edat no estan desenvolupats fisicament per carregar gaire pes, i per tant les motxilles son petites i lleugeres.	Els nens d' aquesta edat no es desenvolupen principalment per carregar massa pesats, i per tant les bosses són petites i lleugers.
2	Sessio de formacio per a pares i mares d' adolescents centrada en la promocio d' habits i estils de vida saludables.	Sessio de en per a pares i mares d' adolescents centrada formacio la promocio d' habits i estils de vida saludables.
3	En algun moment despres d' aixo, va atracar l' Illa Shimotsuki per aliments i subministraments, estant a prop d' en Zoro.	d' aixo, atracar l' Illa Shimotsuki aliments subministraments, a prop Zoro.

Table 1: Examples of data augmentation.

Dataset	Total	l	u	?u	?l	!u	!l	,u	,l	.u	.l	:u	:l
Galician													
Train	682,444	522,508	75,880	880	5,501	955	5,027	10,492	32,979	3,732	15,806	2,064	6,620
Train (Augmented)	920,017	704,358	98,404	1,172	7,477	1,250	6,794	15,719	47,814	5,233	14,713	2,728	9,686
Dev	166,551	127,253	18,655	196	1,326	250	1,246	2,428	8,068	943	3,993	583	1,610
Test	213,570	163,697	23,422	260	1,609	286	1,576	3,322	10,449	1,131	4,967	679	2,172
Catalan													
Train	726,486	569,875	73,357	800	5,481	970	5,149	9,709	33,711	3,450	15,710	2,227	6,047
Train (Augmented)	981,345	768,552	96,442	1,080	7,385	1,293	6,966	13,532	48,161	4,827	21,771	2,993	7,385
Dev	181,517	142,034	18,507	231	1,398	271	1,321	2,377	8,284	853	4,178	517	1,546
Test	228,705	179,531	22,794	255	1,717	310	1,621	3,071	10,771	1,102	4,954	650	1,929

Table 2: Distribution of the datasets.

crawling and public corpora: (1) the Catalan part of the DOGC corpus (Tiedemann, 2012), (2) a collection of translated Catalan movie subtitles, (3) the non-shuffled version of the Catalan part of the OSCAR corpus, (4) a web corpus of Catalan called CaWac (Ljubešić and Toral, 2014), and the Catalan Wikipedia articles.

- **RoBERTinha**: It is RoBERTa-like language model trained on Oscar Galician corpus, and based on the approach presented by Ortiz Suárez, Romary, and Sagot (2020).
- **mBERT**: It is a transformer model pretrained on a large multilingual data corpus of about 104 languages with the largest Wikipedia using Masked Language Modeling (MLM) target (Devlin et al., 2019).
- **DistilmBERT**: This model is a distilled version of BERT base multilingual model (mBERT) (Devlin et al., 2019). It has been trained on the concatenation of Wikipedia in 104 languages listed. The model has 6 layers, 768 dimensions and 112 heads, totalizing 134 parameters (Sanh et al., 2019).
- **XLM-RoBERTa**: This model was proposed in Conneau et al. (2019). It is a multilingual version of RoBERTa trained by Facebook AI Research (FAIR). It has been trained on 2.5 TB of filtered CommonCrawl data containing 100 languages and has demonstrated superior performance on task such as text classification and multi-language text generation compared to other existing language models.

3.4 Tokenization

The main feature of transformer networks is their self-attention mechanism, whereby each word in the input can learn what relation it has with the others (Yi and Tao, 2019). As shown in Figure 1, all models are based on different transformers-based models, such as BERT, RoBERTa or XLM-RoBERTa, and all of them need the input data to be preprocessed by the tokenization process, which consists of decomposing a larger entity into smaller components called *tokens*. For tokenization, we use model-specific tokenizers, and Figure 2 shows some examples of each of them. The models used in this study use the tokenization of sub-words with the WordPiece algorithm as BERT or the Byte-Pair Encoding (BPE) algorithm in the RoBERTa and XLM-RoBERTa-based models, so there are words that split into several tokens as in the case of BERT frequent tokens are grouped into one token and less frequent to-

kens are split into frequent tokens (Bostrom and Durrett, 2020). The main differences in the tokenizers used are as follows:

- In BERT, it uses special tokens such as [CLS] and [SEP] to indicate the beginning and end of a sentence.
- In both RoBERTa and XLM-RoBERTa, the first word of the sentence is not prefixed with any special characters and uses $< s >$ and $< /s >$ to indicate the beginning and end of a sentence.
- In RoBERTa, all tokens in the sentence are prefixed with "Ġ", and when a word is split into several sub-words, the first sub-word is prefixed with "Ġ" and the remaining sub-words are not prefixed with any special characters.
- In XLM-RoBERTa all tokens in the phrase are prefixed with "▁", and when a word is split into various sub-words, the first sub-words are prefixed with "▁" and the rest of the sub-words are not prefixed with any special character.

Therefore, it is necessary to adjust the subword labels and treat special tokens so that they are ignored during training. For this purpose, we have applied the following techniques:

- Assign -100 labels to special tokens such as [CLS], [SEP], $< s >$ and $< /s >$ so that they are ignored during training.
- Assign all sub-words the same label as the first sub-word to solve the sub-word tokenization problem.

4 Results and analysis

We evaluated our proposed transfer learning approaches using the dataset described in Section 3.1. As shown in Figure 1, we fine-tune pre-trained models using various data and fine-tuning strategies to demonstrate the performance of each pre-trained model in sequence labeling tasks for capitalization and punctuation restoration. We provide the results obtained using different pre-trained models including both monolingual for Catalan and Galician (Bertinho, and BERTa), and multilingual (mBERT, DistilmBERT and XLM-RoBERTa).

4.1 Galician

In Table 3, we report our experimental results on the Galician models with Macro-f1 and Weighted-f1. All models are evaluated using Macro-f1 over 12 classes to evaluate the individual performance of each punctuation mark and Weighted-f1 to see the overall performance of the models.

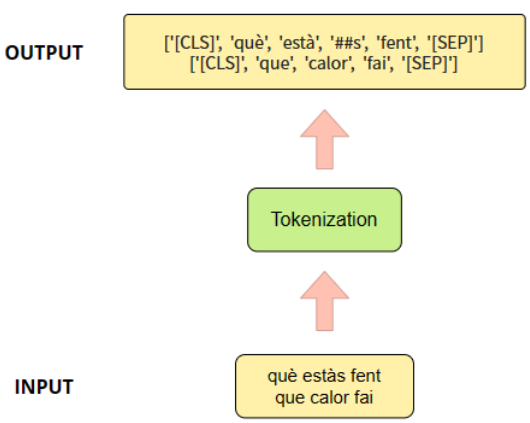
As can be seen in Table 3, all models with augmentation have obtained the best results. Monolingual models, such as Bertinho, perform better than models with a more complex architecture and a larger corpus (DistilmBERT). However, RoBERTinha, which is a monolingual model trained on a reduced corpus, obtained the worst result. XLM-RoBERTa archives a better result than the other models, as it was trained on a large corpus and has a large vocabulary. Our best result is obtained using XLM-RoBERTa with augmentation, and it has a 70,91% accuracy in Macro-f1 and 90.199% overall performance.

In Table 4, the evaluation of each punctuation and capitalization label of the XLM-RoBERTa model with augmentation are displayed. As can be observed, the label that indicates the token is capitalized and has an exclamation mark ('`!u`') or a colon ('`:u`') are the ones that obtain the lowest Macro-f1 because the number of occurrences (see Section 3.1) in this dataset is not sufficient for proper training and evaluation. However, the model predicts capitalized words with question marks ('`?u`') with an accuracy of 67.68%, despite having few occurrences in the training set.

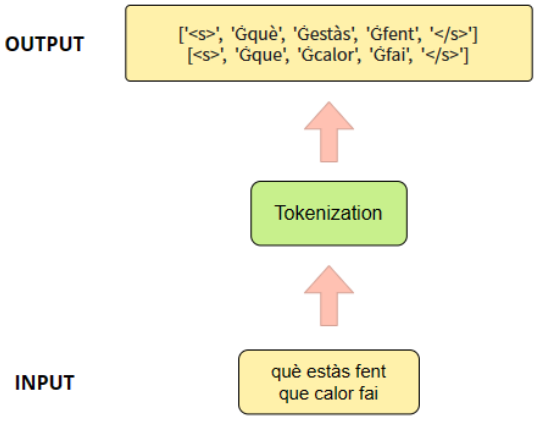
4.2 Catalan

Table 5 shows the macro-averaged *F1* score and weighted-averaged *F1* score for each experiment with the combination of different datasets and the pre-trained model for Catalan. As can be seen, the monolingual models perform better than the multilingual models because the multilingual models (such as mBERT and DistilmBERT) have lower Catalan language content in the training data. BERTa archives a better result than the other models, as it was trained on a large Catalan corpus and has a large vocabulary. Our best result is obtained using BERTa with augmentation, and it has a 69,34% accuracy in Macro-f1 and 90.85% overall performance.

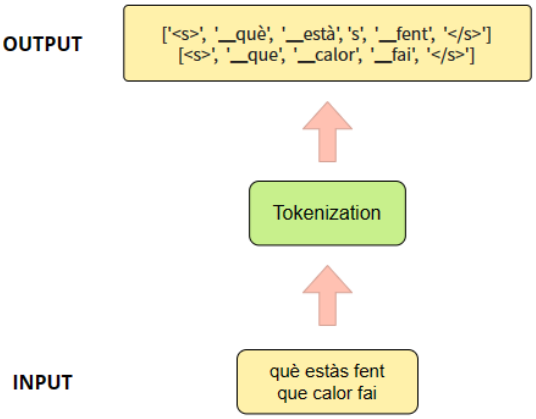
In Table 6, the evaluation of each punctuation and capitalization token of the BERTa



(a) Examples of BERT tokenization



(b) Examples of RoBERTa tokenization



(c) Examples of XLM-RoBERTa tokenization

Figure 2: Examples of tokenization.

	Dataset		Augmented dataset	
Model	Macro-F1 avg	Weighed F1 avg	Macro-F1 avg	Weighted F1 avg
mBERT-cased	66.681	89.297	67.592	89.230
DistilmBERT	62.997	87.843	63.942	87.907
Bertinho	65.687	88.888	66.525	88.906
XLM-RoBERTa	70.448	90.370	**70.941**	**90.199**
RoBERTinha	58.807	87.164	60.252	87.174

Table 3: Results on the dataset and augmented dataset for test sets in Galician.

model with augmentation is shown. As can be seen, the same happens as with the Galician models, that the model is not accurate in classifying the tokens as `!u` and '`:u`' by their number of occurrences in the training set.

Table 4 and 6 illustrate that both models perform well in predicting capitalized words, and with our transfer learning-based sequence labeling approach, the models classify tokens based on the other words. So, they can identify proper names, country names, and organizations well, as shown in Figure 3 and 4.

5 Error analysis

In this section, we analyze the errors of the Catalan and Galician capitalization and punctuation restoration models. For this purpose, we have used the model that has obtained the best result according to Table 5 and 3. To evaluate the performance of these models and to check in which case the models give erroneous predictions, a normalized confusion matrix with truth labels has

	Precision	Recall	F1-score
!l	0.62340	0.62807	0.62573
!u	0.53280	0.50187	0.51688
,l	0.66314	0.69474	0.67857
,u	0.67739	0.68365	0.68050
.l	0.77420	0.73895	0.75616
.u	0.74133	0.75590	0.74854
:l	0.66059	0.62150	0.64045
:u	0.60220	0.56994	0.58563
?l	0.80984	0.78769	0.79861
?u	0.75980	0.61024	0.67686
l	0.95435	0.95795	0.95615
u	0.85912	0.83876	0.84882
Macro avg	0.72151	0.69910	**0.70941**
Weighted avg	0.90205	0.90209	**0.90199**

Table 4: Classification report of the XLM-RoBERTa with data augmentation.

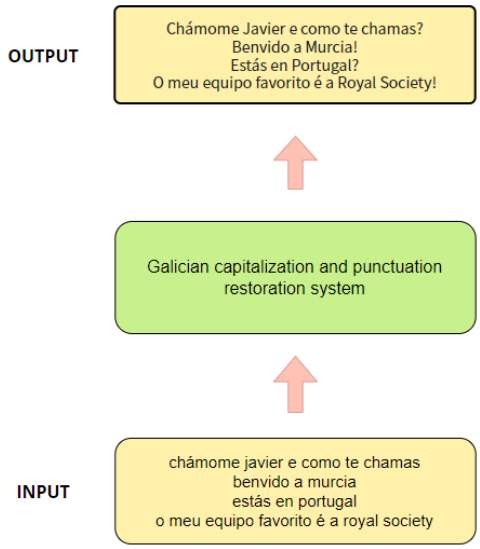


Figure 3: Galician capitalization restoration system examples.

been used, which consists of a table showing the distribution of the predictions of a model compared to the truth label of the data. The confusion matrix of both models is shown in Figure 5.

Concerning the model of capitalization and punctuation retrieval in Catalan, taking into account the confusion matrix (see Figure 5a), it is observed that it does not make many relevant classification errors, such as confusing a comma with a period, and colons with a period marks, which would affect the sentence ending early. Therefore, the focus can be set on the relationship in other punctu-

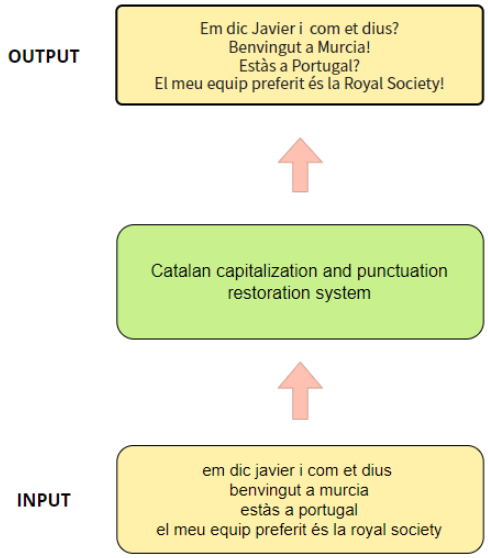


Figure 4: Catalan capitalization restoration system examples.

ations. Through the confusion matrix, it is observed that the model often confuses period marks with the exclamation, and colons with commas. Thus, a set of examples from the test dataset misclassified by the BERTa model trained with the augmented dataset has been analyzed. Table 7 shows the misclassified examples, and it can be seen that it is often very difficult to differentiate between periods and exclamations, as both punctuation marks are placed at the end of the sentence and the only difference is that exclamations are used to show emphasis or an emotional exclamation. Therefore, across texts it is difficult to identify the emotions of a sentence.

Furthermore, in Table 7 we can see that sentences with pronouns such as *what*, *who*, *how*, *where*, *when*, and *which* are ambiguous, as they can be both interrogative and exclamatory pronouns. In our models, when it receives a sentence with only one word and this word is one of the pronouns mentioned above, it always classifies it as an interrogative pronoun (‘`?u`’) instead of an exclamatory pronoun (‘`!u`’). In this case, it is very difficult to solve this problem, as both solutions are valid and the sentence has only one word, so our models cannot use word relation to classify it well.

With respect to Galician capitalization and punctuation restoration model, we have analyzed the confusion matrix (see Figure 5b) and the different misclassified examples of the

	Dataset		Augmented dataset	
Model	Macro-F1 avg	Weighed F1 avg	Macro-F1 avg	Weighted F1 avg
mBERT-cased	65.174	89.341	65.050	89.334
DistilmBERT	60.391	88.042	61.981	88.223
XLM-RoBERTa	69.354	90.753	69.269	90.679
BERTa	68.762	90.735	**69.343**	**90.858**

Table 5: Results on the dataset and augmented dataset for test sets in Catalan.

	Precision	Recall	F1-score
!l	0.63830	0.61224	0.62500
!u	0.52941	0.45652	0.49027
,l	0.68776	0.69414	0.69093
,u	0.65998	0.66643	0.66319
.l	0.77012	0.74903	0.75943
.u	0.69411	0.66018	0.67672
:l	0.59961	0.57685	0.58801
:u	0.57850	0.53147	0.55399
?l	0.79294	0.77802	0.78541
?u	0.69670	0.68764	0.69214
l	0.95894	0.96376	0.96134
u	0.84158	0.82810	0.83479
Macro avg	0.70400	0.68370	**0.69343**
Weighted avg	0.90823	0.90900	**0.90858**

Table 6: Classification report of the BERTa with data augmentation.

XLM-RoBERTa model trained with the augmented dataset. We have seen that the same thing happens as in the Catalan model. The model does not make many relevant classification errors, such as confusing a comma with a period, and colons with period marks, which would affect the sentence ending early. However, it often confuses periods with exclamations and colons with commas, and always classifies pronouns as interrogative.

6 Conclusions and further work

This paper presents two models of capitalization and punctuation restoration, one for Catalan and one for Galician, based on a transfer learning approach through different pre-trained models. The system has been trained for 5 types of punctuation and 2 types of capitalization. In addition, the models are able to identify certain proper names and names of countries and organizations for the capitalization restoration task. Both models have been trained and tested with the OpusParaCrawl dataset. Moreover, we propose an augmentation technique, which improves the performance of the models by up to 1.45% for some models such as RoBERTinha. Our best result is obtained using XLM-RoBERTa with data augmentation for Galician and using BERTa with data augmentation for Catalan. Both have achieved excellent performance with a macro-average *F1* score of 70.94% and overall performance of 90.2% for the Galician, and a macro-average *F1* score of 69.34% and 90.86% of overall performance for the Catalan.

As future work, we would like to use the same approach to create a capitalization and punctuation restoration system for Spanish and compare the performance with other models, such as Zhu et al. (2022), and González-Docasal et al. (2021). And the last proposal is to test a new pre-training model called *Whisper* and see if it works in Catalan and Galician and compare the results with other models, and develop a model that takes into account the relationships of the previous sentence with the following sentence to increase the accuracy of the models and resolve the errors discussed in Section 5.

The models are available on the Huggingface platform[1,2]. In addition, a demo application[3] of these models has also been created for the user to test them in real time. Additional resources concerning to this paper can be accessed.[4]

Acknowledgements

This work is part of the research project (2021/C005/00150076) funded by Spanish Government - Ministerio de Asuntos

[1] https://huggingface.co/UMUTeam/catalan_capitalization_punctuation_restoration

[2] https://huggingface.co/UMUTeam/galician_capitalization_punctuation_restoration

[3] https://huggingface.co/spaces/UMUTeam/punctuation_and_capitalization_restoration

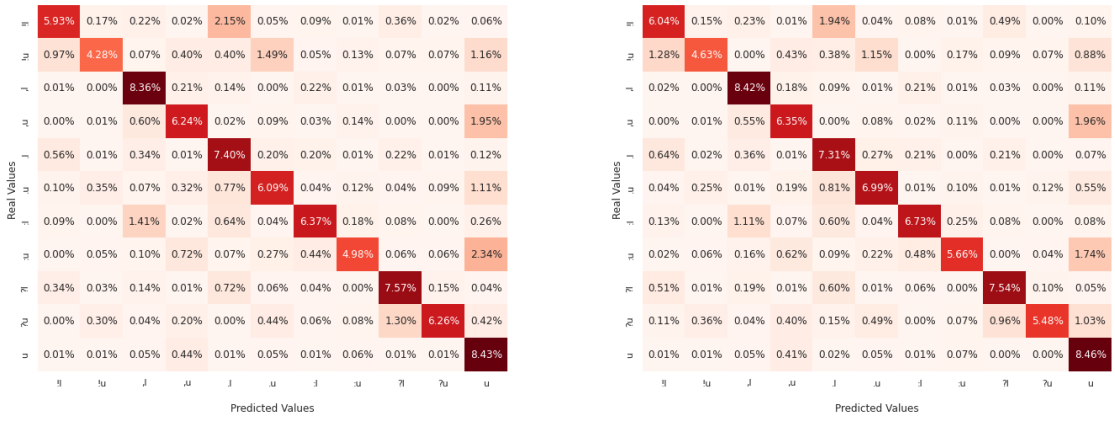
[4] https://github.com/NLP-UMUTeam/capitalization-and-punctuation-restoration

Predicted	Real
Què?	Què!
Com?	Com!
On?	On!
Qui?	Qui!
Quin?	Quin!
Estic bé!	Estic bé.
Et vaig fer el sopar: sopa i truita!	Et vaig fer el sopar: sopa i truita.
Fresca, neta i pura, així és l'aigua de font.	Fresca, neta i pura: així és l'aigua de font.
Aquesta feina no és el meu somni, és una feina amb prou feines.	Aquesta feina no és el meu somni: és una feina amb prou feines.

Table 7: A set of examples misclassified by the BERTa model trained with the augmented dataset for Catalan.

Predicted	Real
que?	Que!
Quen?	Quen!
Cal?	Cal!
Canto?	Canto!
onde?	Onde!
Estou ben!	Estou ben.
Hoxe chegou tarde!	Hoxe chegou tarde.
Querido amigo, hai moito que non sei nada de ti!	Querido amigo: Hai moito que non sei nada de ti.
Naquela libraría había de todo, libros, xornais, cómics.	Naquela libraría había de todo: libros, xornais, cómics.

Table 8: A set of examples misclassified by the XLM-RoBERTa model trained with the augmented dataset for Galician.



(a) Catalan (b) Galician

Figure 5: Confusion matrix of Catalan and Galician capitalization and punctuation restoration system.

Económicos y Transformación and by the European Union NextGenerationEU/PRTR. This work is also part of the research project LaTe4PSP (PID2019-107652RB-I00/AEI/ 10.13039/501100011033) funded by MCIN/AEI/10.13039/501100011033.

References

Alam, T., A. Khan, and F. 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, Online, November. Association for Computational Linguistics.

Armengol-Estapé, J., C. P. Carrino, C. Rodriguez-Penagos, O. de Gibert Bonet, C. Armentano-Oller, A. Gonzalez-Agirre, M. Melero, and M. Villegas. 2021. Are multilingual

models the best choice for moderately under-resourced languages? A comprehensive assessment for Catalan. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4933–4946, Online, August. Association for Computational Linguistics.

Bannard, C. and C. Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 597–604, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Bañón, M., P. Chen, B. Haddow, K. Heafield, H. Hoang, M. Esplà-Gomis, M. L. Forcada, A. Kamran, F. Kirefu, P. Koehn, S. Ortiz Rojas, L. Pla Sempere, G. Ramírez-Sánchez, E. Sarrías, M. Strelec, B. Thompson, W. Waites, D. Wiggins, and J. Zaragoza. 2020. ParaCrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567, Online, July. Association for Computational Linguistics.

Basili, R., C. Bosco, R. Delmonte, A. Moschitti, and M. Simi, editors. 2015. *Harmonization and Development of Resources and Tools for Italian Natural Language Processing within the PARLI Project*, volume 589 of *Studies in Computational Intelligence*. Springer.

Bostrom, K. and G. Durrett. 2020. Byte pair encoding is suboptimal for language model pretraining. *CoRR*, abs/2004.03720.

Cañete, J., G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, and J. Pérez. 2020. Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020*.

Che, X., C. Wang, H. Yang, and C. 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, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Conneau, A., K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Courtland, M., A. Faulkner, and G. McElvain. 2020. Efficient automatic punctuation restoration using bidirectional transformers with robust inference. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 272–279, Online, July. Association for Computational Linguistics.

David Vilares, Marcos Garcia, C. G.-R. 2021. Bertinho: Galician bert representations. *Procesamiento del Lenguaje Natural*, 66(0):13–26.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Federico, M., M. Cettolo, L. Bentivogli, M. Paul, and S. Stüker. 2012. Overview of the IWSLT 2012 evaluation campaign. In *Proceedings of the 9th International Workshop on Spoken Language Translation: Evaluation Campaign*, pages 12–33, Hong Kong, Table of contents, December 6-7.

González-Docasal, A., A. García-Pablos, H. Arzelus, and A. Álvarez. 2021. Autopunct: A bert-based automatic punctuation and capitalisation system for spanish and basque. *Procesamiento del Lenguaje Natural*, 67(0):59–68.

Jones, D., F. Wolf, E. Gibson, E. Williams, E. Fedorenko, D. Reynolds, and M. Zissman. 2003. Measuring the readability of automatic speech-to-text transcripts. In *8th European Conference on Speech Communication and Technology, EUROSPEECH 2003 - INTERSPEECH 2003, Geneva, Switzerland, September 1-4, 2003*. ISCA, 09.

Ljubešić, N. and A. Toral. 2014. cawac - a web corpus of catalan and its ap-

plication to language modeling and machine translation. In N. C. C. Chair), K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).

Ortiz Suárez, P. J., L. Romary, and B. Sagot. 2020. A monolingual approach to contextualized word embeddings for mid-resource languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1703–1714, Online, July. Association for Computational Linguistics.

Peitz, S., M. Freitag, A. Mauser, and H. Ney. 2011. Modeling punctuation prediction as machine translation. In *Proceedings of the 8th International Workshop on Spoken Language Translation: Papers*, pages 238–245, 12.

Sanh, V., L. Debut, J. Chaumond, and T. Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Tiedemann, J. 2012. Parallel data, tools and interfaces in opus. In N. C. C. Chair), K. Choukri, T. Declerck, M. U. Dogan, B. Maegaard, J. Mariani, J. Odijk, and S. Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Tilk, O. and T. Alumäe. 2016. Bidirectional recurrent neural network with attention mechanism for punctuation restoration. In *INTERSPEECH*.

Yi, J. and J. Tao. 2019. Self-attention based model for punctuation prediction using word and speech embeddings. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7270–7274.

Yi, J., J. Tao, Y. Bai, Z. Tian, and C. Fan. 2020. Adversarial transfer learning for punctuation restoration.

Zhu, X., S. Gardiner, D. Rossouw, T. Roldán, and S. Corston-Oliver. 2022. Punctuation restoration in Spanish customer support transcripts using transfer learning. In *Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing*, pages 80–89, Hybrid, July. Association for Computational Linguistics.