Punctuation Restoration in Spanish Customer Support Transcripts using Transfer Learning

Xiliang Zhu, Shayna Gardiner, David Rossouw
Tere Roldán, Simon Corston-Oliver
Dialpad Canada Inc.
{xzhu, sgardiner, davidr}@dialpad.com
{tere.roldan, scorston-oliver}@dialpad.com

Abstract

Automatic Speech Recognition (ASR) systems typically produce unpunctuated transcripts that have poor readability. In addition, building a punctuation restoration system is challenging for low-resource languages, especially for domain-specific applications. In this paper, we propose a Spanish punctuation restoration system designed for a real-time customer support transcription service. To address the data sparsity of Spanish transcripts in the customer support domain, we introduce two transfer-learning-based strategies: 1) domain adaptation using out-of-domain Spanish text data; 2) cross-lingual transfer learning leveraging in-domain English transcript data. Our experiment results show that these strategies improve the accuracy of the Spanish punctuation restoration system.

1 Introduction

Automatic Speech Recognition (ASR) systems play an increasingly important role in our daily lives, with a wide range of applications in different domains such as voice assistant, customer support and healthcare. However, ASR systems usually generate an unpunctuated word stream as the output. Unpunctuated speech transcripts are difficult to read and reduce overall comprehension (Jones et al., 2003) . Punctuation restoration is thus an important post-processing task on the output of ASR systems to improve general transcript readability and facilitate human comprehension.

Punctuation restoration for transcripts of Spanish-speaking customer support telephone dialogue is a non-trivial task. First, real-world human conversation transcripts have unique characteristics compared to common written text, e.g., filler words and false starts are common in spoken dialogue. Moreover, further challenges arise when addressing noisy ASR transcripts in a specific domain, as the lexical data distribution can be quite different compared to public Spanish datasets. Examples of Spanish sentences from different sources are shown below:

- **Written text in Wikipedia**: *El español o castellano es una lengua romance procedente del latín hablado, perteneciente a la familia de lenguas indoeuropeas.* (Spanish or Castilian is a Romance language derived from spoken Latin, belonging to the Indo-European language family.)
- **Written text in customer support**: *Mire, quería ver si me podían ayudar.* (Look, I wanted to see if you guys could help me)
- **Noisy ASR transcript in customer support**: *Mire, este, es que, que- quería ver si me podían ayudar.* (Look, well, so I, I wanted to see if you could help me)

Recent advances in transformer-based pre-trained models have been proven successful in many NLP tasks across different languages. For Spanish, available pre-trained resources include multilingual models such as multilingual BERT (mBERT) (Devlin et al., 2019) and XLM-RoBERTa (XLM-R) (Conneau et al., 2020), as well as monolingual models such as BETO (Cañete et al., 2020). However, large pre-trained models are trained on various written text sources such as Wikipedia and CommonCrawl (Wenzek et al., 2019), which are very distant from what we are trying to address in noisy ASR transcripts in the customer support domain. While Spanish is not usually considered a low-resource language in many NLP tasks, it is much more challenging to acquire sufficient training data in Spanish for our domain-specific task, since most of the publicly-available Spanish datasets do not come from natural human conversations, and have little coverage in the customer support domain.

In addressing the challenge of in-domain data sparsity we make the following contributions:

Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing, pages 80 - 89
July 14, 2022 ©2022 Association for Computational Linguistics

1. We propose a punctuation restoration system dedicated for Spanish based on pre-trained models, and examine the feasibility of various pre-trained models for this task.

2. We adopt a domain adaptation approach utilizing out-of-domain Spanish text data.

3. We implement a data modification strategy and match in-domain English transcripts with Spanish punctuation usage, and propose a cross-lingual transfer approach using English transcripts.

4. We demonstrate that our proposed transfer learning approaches (domain adaptation and cross-lingual transfer) can sufficiently improve the overall performance of Spanish punctuation restoration in our customer support domain, without any model-level modifications.

2 Background

Punctuation restoration is the task of inserting appropriate punctuation marks in the appropriate position on the unpunctuated text input. A variety of approaches have been used for punctuation restoration, most of which are built and evaluated on one language: English. The use of classic machine learning models such as n-gram language model (Gravano et al., 2009) and conditional random fields (Lu and Ng, 2010) are common in early studies. More recently, deep neural networks such as Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) and transformers (Vaswani et al., 2017) have been adopted in (Tilk and Alumäe, 2015) and (Courtland et al., 2020).

Punctuation conventions differ between Spanish and English. Namely, in addition to the equivalents of English and Spanish periods, commas, terminating question marks and terminating exclamation marks, we must also account for the inverted question marks (¿) and inverted exclamation marks (¡) used to introduce these respective clauses in Spanish. There has been limited work done in Spanish punctuation restoration and in most cases Spanish is covered as part of the multilingual training. (Li and Lin, 2020) proposed a multilingual LSTM including the support for Spanish. (González-Docasal et al., 2021) uses a transformer-based model with both lexical and acoustic inputs for Spanish and Basque.

Transfer learning has been widely studied and applied in NLP applications for low-resource languages (Alyafeai et al., 2020). Domain adaptation and cross-lingual learning both fall under the category of transductive transfer learning, where source and target share the same task but labeled data is only available in source (Ruder et al., 2019). Data selection is among the data-centric methods used in domain adaptation, which aims to select the best matching data for a new domain (Ramponi and Plank, 2020). (Fu et al., 2021) uses data selection to improve English punctuation restoration with out-of-domain datasets. Recent advances in multilingual language models such as mBERT and XLM-R have shown great potential in cross-lingual zero-shot learning, wherein a multilingual model can be trained on the target task in a high-resource language, and afterwards applied to the unseen target languages by zero-shot learning (Hedderich et al., 2021). (Wu and Dredze, 2019) and (Pires et al., 2019) demonstrate the effectiveness of mBERT as a zero-shot cross-lingual transfer model in various NLP tasks, such as classification and natural language inference.

3 Methods

3.1 System Description

Pre-trained transformer-based models have been widely adopted for various NLP tasks since the introduction of BERT (Devlin et al., 2019). Publicly available pre-trained models for Spanish include the multilingual models mBERT and XLM-R and the BERT-like monolingual model BETO. In this work, we evaluate all three pre-trained models in our experiments and compare their performance in both proposed domain adaptation and cross-lingual transfer approaches.

Using pre-trained models as a starting point, we formulate the Spanish punctuation restoration problem as a sequence labeling task, where the model predicts one punctuation class for each input word token. Instead of covering all possible Spanish punctuation marks, we only include nine target punctuation classes that are commonly used and important in terms of improving transcript readability:

- OPEN_QUESTION: **¿** should be added at the start of this word token.

- CLOSE_QUESTION: **?** should be added at the end of this word token.

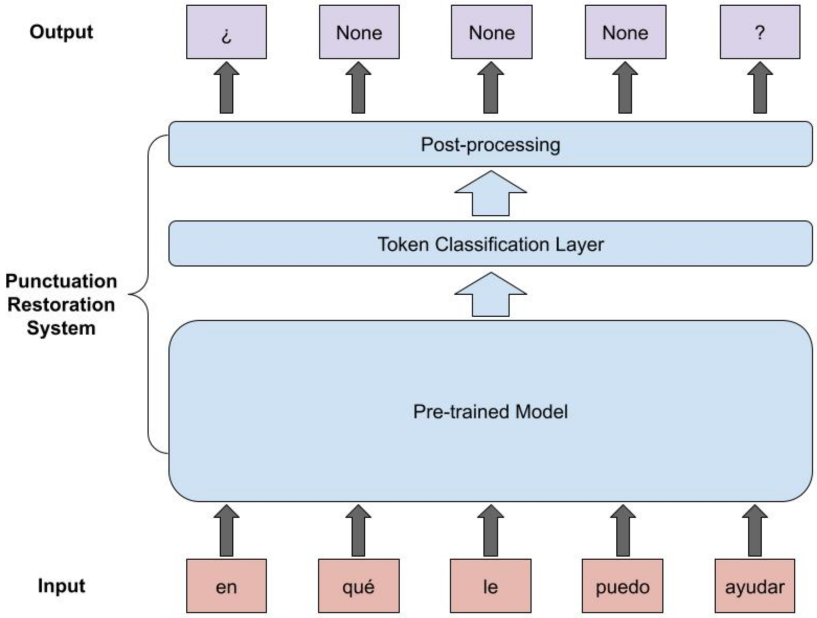


Figure 1: Our punctuation restoration system, showing the process of predicting "*en qué le puedo ayudar*" as "*¿En qué le puedo ayudar?*" (How can I help you?).

- FULL_QUESTION: **¿** and **?** should be added at the start and end of this word token respectively.
- OPEN_EXCLAMATION: **¡** should be added at the start of this word token.
- CLOSE_EXCLAMATION: **!** should be added at the end of this word token.
- FULL_EXCLAMATION: **¡** and **!** should be added at the start and end of this word token respectively.
- COMMA: **,** should be added at the end of this word token.[1]
- PERIOD: **.** should be added at the end of this word token.
- NONE: no punctuation should be associated with this word token.

The input to the Spanish punctuation restoration system is a transcribed utterance emitted by the ASR system. The ASR system outputs an utterance if an endpoint (long pause or speaker change) is detected in the audio. The length of a given utterance can vary, each utterance can contain multiple sentences, meaning that there can be multiple terminating punctuation marks – period, question mark and exclamation mark – in a single utterance.

The punctuation restoration model structure is illustrated in Figure 1. We add a token classification layer on top of the pre-trained models. Raw model prediction results are also post-processed by a set of simple heuristics to mitigate the error caused by unmatched predictions for paired punctuation marks. For instance, a predicted OPEN_QUESTION class will be changed to NONE if there is no matched CLOSE_QUESTION prediction in the same utterance. [2]

3.2 Datasets

It is essential to acquire in-domain manual transcripts that come from real customer support scenarios to build a punctuation restoration model that fits the customer support domain. However, only around 5,000 in-domain transcribed Spanish utterances from call recordings could be obtained at this early product development stage. Addition-

[1] The insertion of commas as decimal separators is not included here.

[2] This post-processing step may not always produce the correct result, but the overall prediction accuracy was improved by adding this post-processing in our experiments.

Spanish out-of-domain (LDC) examples
Ah, qué bueno, yo conozco mucho cubano pero más que todo en Filadelfia. (Ah, how good, I know many Cubans but especially in Philadelphia.)
Bueno, mira, eh, ¿sus papás, cuántos años llevan casados? (Well, look, uhm, your parents, how long have they been married?)
Spanish out-of-domain (OpenSubtitle) examples
Sé que lo que estoy pidiéndote es difícil. (I know that what I'm asking you is hard.)
Sí, da un poco de tristeza. (Yes, it makes you a little bit sad.)
Spanish in-domain examples
Buenas tardes, ¿cómo le puedo ayudar? (Good afternoon, how can I help you?)
Pues no me funciona y lo he intentado varias veces. (So, it doesn't work and I've tried several times)
English in-domain examples
I don't find this app very helpful, I'm calling to cancel my subscription.
Hi, this is Tom, how can I help you today?

Table 1: Examples of Spanish and English utterances.

ally, there are around 200,000 in-domain manually transcribed English utterances from our call center product.

We supplemented this in-domain Spanish data with the Linguistic Data Consortium (LDC) Fisher Spanish Speech and Fisher Spanish Transcripts corpora (Graff et al., 2010). These corpora consist of audio files and transcripts for approximately 163 hours of telephone conversations from native Spanish speakers. These recordings are a good match to the acoustic properties of our telephone conversations, but the transcripts, which are mostly social calls with predefined topics, do not match the domain of customer support conversations.

The Spanish portion of the OpenSubtitle corpus (Lison and Tiedemann, 2016) also contains a variety of human-to-human conversation, albeit from movies rather than from spontaneous conversational speech. Spanish OpenSubtitle offers 179 million sentences from 192,000 subtitle files, and can provide our models with good exposure to exclamation marks, which are not included in the LDC dataset. However, the movie topics are generally distant from our business-specific, customer support domain.

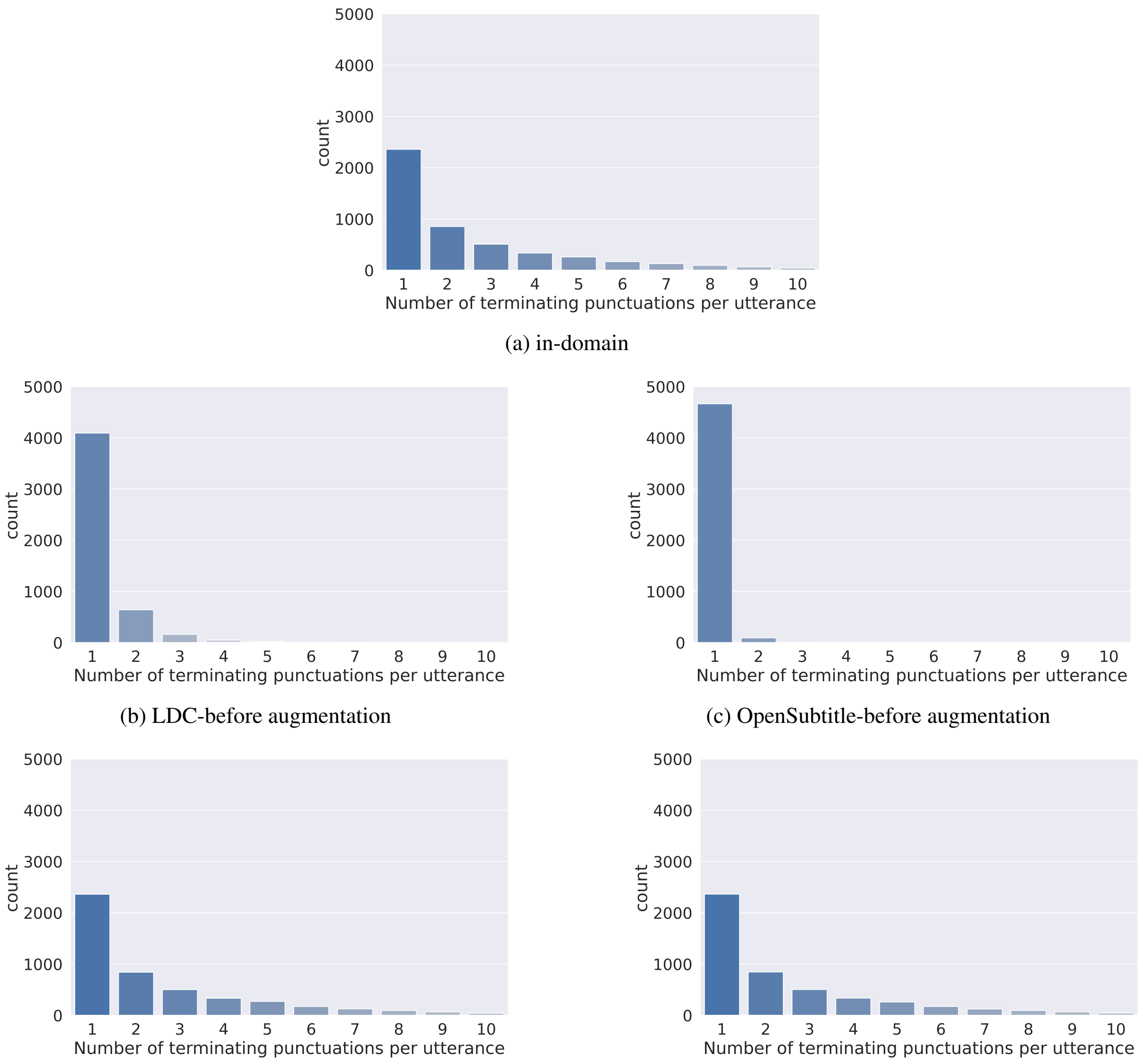
Some examples from both in-domain and out-of-domain data sources are illustrated in Table 1. External out-of-domain datasets usually have various Spanish punctuation marks outside our supported range as described in 3.1. After reviewing the datasets from a linguistic perspective, we first apply a set of conversion rules to those unsupported punctuation marks without affecting the readability and semantic meanings: we delete quotation marks, replace colons and semicolons with commas, and replace ellipses with periods.

3.3 Domain Adaptation

Many machine learning applications have the assumption that training and testing datasets follow the same underlying distribution. But for our target task in the customer support domain, we mostly have to rely on external data such as LDC and Spanish OpenSubtitle during the training process, due to the lack of in-domain Spanish data. This will therefore cause a mismatch between our training and testing data in terms of its distribution, and consequently, performance will drop in our target task. Therefore, to mitigate this distribution mismatch, we apply domain adaptation on external Spanish datasets from two directions: data selection and data augmentation.

3.3.1 Data Selection

As described in 3.2, Spanish OpenSubtitle has a total of over 179 million sentences, which is much larger than our other data sources. However, the vast majority of the data in the Spanish OpenSubtile corpus are fundamentally distinct from our target customer support domain, and randomly sampling from out-of-domain datasets could hurt the model performance. Thus, following the procedure in (Fu et al., 2021), we first train a 4-gram language model using our Spanish in-domain data, and then sample the 100,000 utterances from the OpenSubtitle corpus with lowest perplexity (i.e. the highest language model similarity to the in-domain data).



(a) in-domain

(b) LDC-before augmentation

(c) OpenSubtitle-before augmentation

(d) LDC-after augmentation

(e) OpenSubtitle-after augmentation

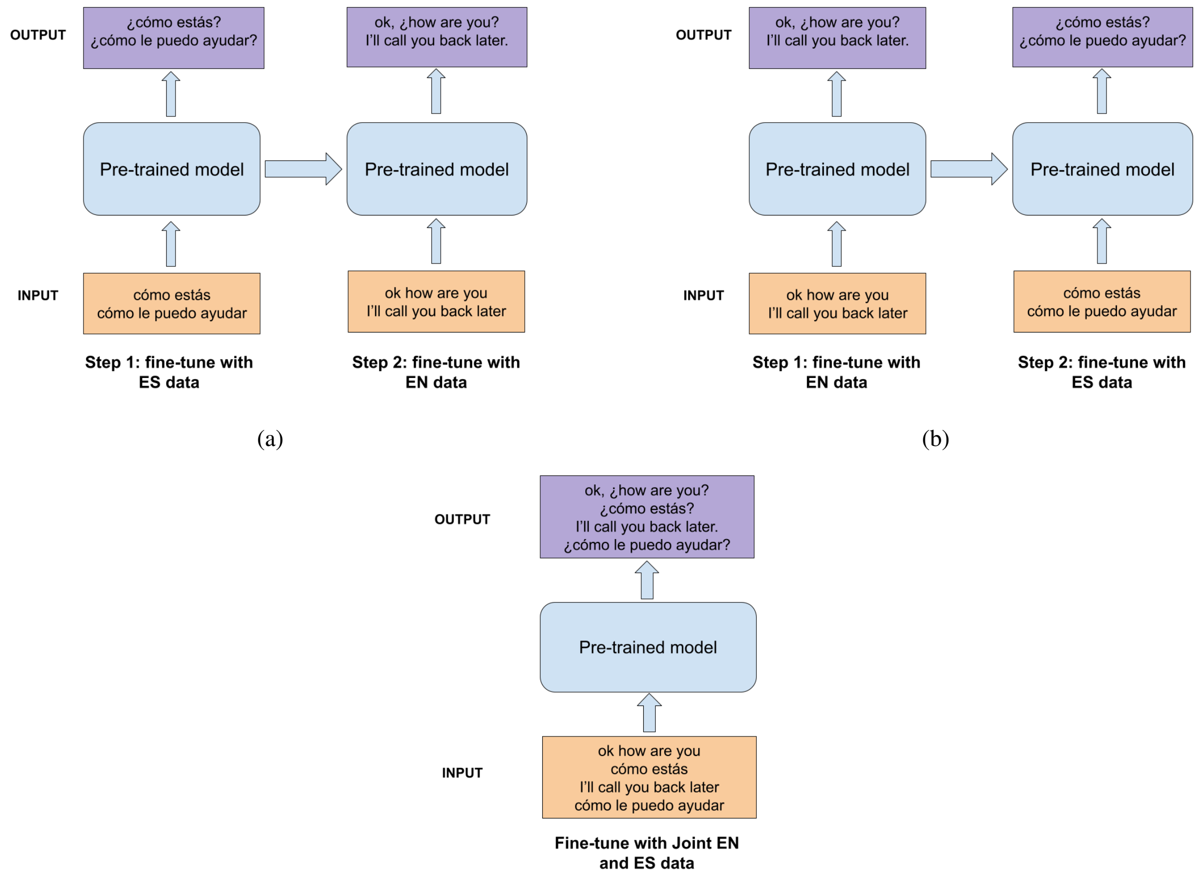
Figure 2: Comparison of number of terminating punctuations per utterance distribution in in-domain, LDC and OpenSubtitle datasets, before and after data augmentation.

Since the telephone conversation transcripts in the LDC corpora are closer to our target domain and there are only 130,000 utterances in this dataset, we do not perform further data selection on the LDC data for training purposes.

3.3.2 Data Augmentation

Most of the data in LDC and OpenSubtitle datasets is segmented into single sentences. However, as described in 3.1, the input to our punctuation restoration system will be composed of larger blocks of utterances rather than single sentences. To illustrate this difference, we investigate how many terminating punctuation marks occur in each input from external datasets and in-domain data, respectively.

As shown in Figure 2(a)(b)(c), our in-domain data has a much wider distribution in terms of the number of terminating punctuation marks in a single utterance. However, the majority of samples in both LDC and OpenSubtitle consist of only one sentence each. It is necessary to augment the out-of-domain datasets to cover the wider spread of distribution exhibited in our in-domain data, based on the fact that this will affect how many terminating punctuation marks the model tends to predict per input utterance. We therefore apply data augmentation by concatenating sentences in these corpora, in proportion to the spread seen in our in-domain dataset, so that the overall terminating punctuation distribution in out-of-domain datasets matches our in-domain data. As Figure 2(d)(e) shows, the



(a)

(b)

(c)

Figure 3: Diagram of three proposed fine-tuning strategies. (a) **ES->EN**, (b) **EN->ES**, (c) **Joint EN, ES**

augmented results for the LDC and OpenSubtitle corpora more closely match the distribution of our in-domain Spanish data.

3.4 Cross-lingual Transfer

Multilingual language models such as mBERT and XLM-R advanced zero-shot cross-lingual transfer learning for low-resource languages (Hedderich et al., 2021). Instead of using cross-lingual transfer as zero-shot, we utilize our English in-domain data (described in 3.2) to fine-tune multilingual pre-trained models in addition to our available Spanish datasets to improve our Spanish punctuation restoration system. However, punctuation conventions differ between languages; to better leverage cross-lingual transfer learning, we first convert the punctuation usage in the source language to appropriately match the punctuation conventions in the target language.

Since this study involves matching English punctuation to Spanish, the task is not insurmountable: most of the punctuation marks and their usages are the same across these two languages. Periods are used to terminate a declarative sentence in both languages, and the usage of commas to separate words or phrases is very similar. Therefore, no modifications are required for these two punctuation marks.

One more significant challenge for this task is the fact that question marks and exclamation marks do work somewhat differently in Spanish writing than in English. Namely, in addition to the terminating role played in both languages by standard question marks (to denote the end of an interrogative sentence) and standard exclamation marks (to denote the end of an exclamatory sentence), Spanish writing conventions also require the addition of an inverted question mark or an inverted exclamation mark, which occur at the beginning of the clause that contains the question or exclamation. For example:

- **English**: *Hi, how are you today?*
- **Spanish**: *Hola, ¿cómo estás hoy?*

For each question mark and exclamation mark in our English training data, we add an open question

Training Data	BETO	mBERT	XLM-R
LDC	51.3%	50.2%	51.8%
LDC + Selected OpenSubtitle	52.1%	51.5%	53.2%
Augmented (LDC + Selected OpenSubtitle)	**53.7%**	**52.1%**	**54.7%**

Table 2: F1 score performance comparison using the LDC and OpenSubtitle datasets, before and after our domain adaptation approaches.

mark or exclamation mark, respectively, at the start of the word chunk that the terminating question or exclamation mark is in.

For example, consider the following English utterance:

“OK, how can I help you?”

For cross-lingual transfer training, it will be modified to:

“OK, ¿how can I help you?”

By doing this conversion, the model will learn to predict punctuation as it should occur in Spanish contexts during the fine-tuning phase, even though what it actually sees are English utterances with Spanish punctuation.

To determine the best way to transfer the in-domain distribution from English (**EN**) to Spanish (**ES**) in the punctuation restoration task, we investigate three fine-tuning strategies for cross-lingual transfer learning:

1. Fine-tune the pre-trained models in two steps, Spanish first then English. Noted as “**ES->EN**”.

2. Fine-tune the pre-trained models in two steps, English first then Spanish. Noted as “**EN->ES**”.

3. Fine-tune the pre-trained models in one step, with joint English and Spanish data. Noted as “**Joint EN, ES**”

Diagrams of three fine-tuning strategies are illustrated in Figure 3. Note that our objective is to build a model for Spanish, but it is still worth experimenting with “**ES->EN**” setting to establish the impact of more in-domain data albeit in a different language.

4 Evaluation

4.1 Evaluation Setup

We evaluate our proposed transfer learning approaches using the datasets described in 3.2. Using the model architecture shown in Figure 1, we fine-tune pre-trained models using various data combinations and fine-tuning strategies to demonstrate the effectiveness of our proposed approaches. Pre-trained models including both monolingual (BETO) and multilingual (MBERT and XLM-R) are explored and evaluated.

The Spanish punctuation restoration system is intended to operate in real-time so that customer-support agents can review prior information communicated by a customer and to provide the input to product features such as automatically retrieving information to assist the agent. As shown in (Fu et al., 2021), reducing the number of layers from deep pre-trained models does not significantly impact accuracy for the punctuation restoration task. To reduce the computation time during inference, we take only the first six layers from the pre-trained models as our starting point.

To evaluate the model accuracy in our target customer support domain, we split our in-domain Spanish manual transcripts into three parts: the training set (60%), the validation set (10%) and the test set (30%). The Spanish in-domain training set is over-sampled to make the size comparable to the other datasets. The performance of every model is evaluated on the in-domain test set after being fine-tuned on various combinations of training sources and processes.

4.2 Performance with Domain Adaptation

We evaluate the F1 score performance before and after the domain adaptation approaches proposed in 3.3. Pre-trained models are fine-tuned using the combinations of LDC and selected OpenSubtitle datasets only, and then evaluated on our in-domain test set. The results are shown in Table 2. Both data selection and data augmentation improve the overall F1 score performance for all three pre-trained models, which demonstrates the effectiveness of our domain adaptation approaches for the Spanish punctuation restoration task. Among three different models, XLM-R shows the best performance under this setup, and outperforms the monolingual BETO model after domain adaptation.

Training data and strategy	BETO	mBERT	XLM-R
ES only (no cross-lingual transfer)	62.8%	61.5%	62.9%
ES->EN	N/A	59.1%	60.7%
EN->ES	N/A	62.0%	63.5%
Joint EN,ES	N/A	62.4%	**64.4%**

Table 3: F1 score performance comparison with and without cross-lingual transfer. ***ES***: the combination of Spanish datasets including (1) Augmented (LDC + Selected OpenSubtitle) as described in Table 2; (2) Spanish in-domain transcripts. ***EN***: English in-domain transcripts.

Prediction Gold	CLOSE_QUESTION	PERIOD
CLOSE_QUESTION	223	106
PERIOD	37	2177

Table 4: Confusion matrix of CLOSE_QUESTION and PERIOD on test set, using best performing XLM-R in 4.3

4.3 Performance with Cross-lingual transfer

To understand the effect of cross-lingual transfer, we use all the available data sources described in 3.2. We separate the Spanish datasets (LDC, selected OpenSubtitle and Spanish in-domain transcripts) from the English one (English in-domain transcripts), and fine-tune the pre-trained models using three different strategies described in 3.4 ("**ES->EN**", "**EN->ES**" and "**Joint EN, ES**") as shown in Figure 3.

Table 3 shows our results on cross-lingual transfer learning: multilingual models (mBERT and XLM-R) both show performance gain with "**Joint EN, ES**" and "**EN->ES**" training. However, "**ES->EN**" training actually results in lower accuracy than models trained without cross-lingual transfer. As for the comparison with the monolingual model (BETO) which is not feasible for the direct cross-lingual transfer, XLM-R produces similar results as BETO without cross-lingual transfer, but XLM-R outperforms BETO by 1.5% F1 score after joint training with both Spanish and English datasets. mBERT becomes comparable to BETO after cross-lingual transfer as well.

5 Future Work

When analysing the prediction errors, we found that many CLOSE_QUESTION classes are predicted as PERIOD by the model, as shown in Table 4. This is a common behavior across all three pre-trained models, and is possibly due to the linguistic properties of Spanish. Because Spanish clauses do not require an overt subject noun phrase, and because Spanish has considerable variability in constituent order, it is often the case that there is no structural indication of whether an utterance should be interpreted as a declarative or as a question. Instead, intonation is used to make this distinction. For example, "*hablan español*" ("they speak Spanish" or "do they speak Spanish") becomes a question with rising intonation. Future work in this area might focus on incorporating such acoustic information into punctuation restoration tasks.

6 Conclusion

For this study, we trained and tested a Spanish punctuation restoration system for the customer support domain based on pre-trained transformer models. To address in-domain data sparsity in Spanish, transfer learning approaches were applied in two directions: domain adaptation and cross-lingual transfer. We explored and fine-tuned three different pre-trained models with our transfer learning approaches for this task; our results demonstrate that the domain adaptation method improves the accuracy of all three pre-trained models. Cross-lingual transfer with joint training of English and Spanish datasets improves the performance of both multilingual pre-trained models. XLM-R substantially outperforms the monolingual BETO after cross-lingual transfer and achieves the best F1 score in our Spanish punctuation restoration task.

References

Zaid Alyafeai, Maged Saeed AlShaibani, and Irfan Ahmad. 2020. A survey on transfer learning in natural language processing. *CoRR*, abs/2007.04239.

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*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Maury Courtland, Adam Faulkner, and Gayle 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. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina 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. Association for Computational Linguistics.

Xue-Yong Fu, Cheng Chen, Md Tahmid Rahman Laskar, Shashi Bhushan, and Simon Corston-Oliver. 2021. Improving punctuation restoration for speech transcripts via external data. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 168–174, Online. Association for Computational Linguistics.

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(0):59–68.

David Graff, Shudong Huang, Ingrid Cartagena, Kevin Walker, and Christopher Cieri. 2010. Fisher spanish - transcripts ldc2010t04.

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.

Michael A. Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2021. A survey on recent approaches for natural language processing in low-resource scenarios. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2545–2568, Online. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Douglas Jones, Florian Wolf, Edward Gibson, Elliott Williams, Evelina Fedorenko, Douglas Reynolds, and Marc Zissman. 2003. Measuring the readability of automatic speech-to-text transcripts.

Xinxing Li and Edward Lin. 2020. A 43 Language Multilingual Punctuation Prediction Neural Network Model. In *Proc. Interspeech 2020*, pages 1067–1071.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Wei Lu and Hwee Tou Ng. 2010. Better punctuation prediction with dynamic conditional random fields. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 177–186, Cambridge, MA. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Alan Ramponi and Barbara Plank. 2020. Neural unsupervised domain adaptation in NLP—A survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6838–6855, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Sebastian Ruder, Anders Søgaard, and Ivan Vulić. 2019. Unsupervised cross-lingual representation learning. In *Proceedings of ACL 2019, Tutorial Abstracts*, pages 31–38.

Ottokar Tilk and Tanel Alumäe. 2015. LSTM for punctuation restoration in speech transcripts. In *Proc. Interspeech 2015*, pages 683–687.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2019. CCNet: Extracting high quality monolingual datasets from web crawl data. *CoRR*.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical*

Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 833–844, Hong Kong, China. Association for Computational Linguistics.