Evaluating Transformer Models for Punctuation Restoration in Italian

Alessio Miaschi[1,2], Andrea Amelio Ravelli[2], and Felice Dell'Orletta[2]

[1] Department of Computer Science, Università di Pisa
[2] Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC–CNR), ItaliaNLP Lab, www.italianlp.it
alessio.miaschi@phd.unipi.it,
{andreaamelio.ravelli, felice.dellorletta}@ilc.cnr.it

Abstract. In this paper, we propose an evaluation of a Transformer-based punctuation restoration model for the Italian language. Experimenting with a BERT-base model, we perform several fine-tuning with different training data and sizes and tested them in an in- and cross-domain scenario. Moreover, we offer a comparison in a multilingual setting with the same model fine-tuned on English transcriptions. Finally, we conclude with an error analysis of the main weaknesses of the model related to specific punctuation marks.

Keywords: punctuation restoration · transformers · speech transcription

1 Introduction

Nowadays, Automatic Speech Recognition (ASR) and Speech-to-Text technologies and services have reached an incredible level of accuracy in transcribing recorded (or live) speech audio streams. A simple but effective test can be run, with any modern smartphone, by using the dictation feature to write a text message.[1] However, we can immediately notice that the audio stream is transcribed as a word stream, lacking any punctuation or sentence segmentation, and sometimes pieces of text are difficult to understand without some attempts to mentally insert punctuation marks in the flow of words.[2]

Lack of punctuation may be a minor problem in everyday short-text messaging, but correctly inserted punctuation is crucial in long speech transcription, live subtitling or any NLP processing of speech data, especially for downstream processes such as parsing, information extraction, dialog modeling. Many major

[1] Obviously, the speech must have a close-to-standard accent without using dialectal or slang words.

[2] Other than difficult, unpunctuated text can be also ambiguous. Here is an amusing example of two completely different letters, with the same words but different punctuation: https://www.nationalpunctuationday.com/dearjohn.html.

commercial services such as Google Cloud[3] or Microsoft Azure[4] offer the option of including automatically generated punctuation. As well, it is possible to train a public ASR model, such as wav2vec [25] or Vosk,[5] and then apply a Punctuation Restoration technique on the output of the first. Both alternatives come at a cost: on one side, commercial services requires a payment fee; on the other, training requires computational power, time and, above all, good and enough training data in the form of aligned audio sources, transcriptions and phonetic annotations. By assuming of working on already transcribed data, recent Transformers models could be a convenient way of tackling punctuation restoration in standard language transcriptions, as they can be easily fine-tuned on many tasks, including the insertion of commas, periods and question marks. The objective of this paper is to verify if it could be possible to obtain good results in transcription by post-processing raw text from everyday Speech-to-Text technologies (e.g. dictation on a smartphone) with a Transformers model fine-tuned for Punctuation Restoration. More specifically, we set our experiments on Italian language and we verify the impact of different domains and sizes of fine-tuning data on the performance of a Transformer-based punctuation restoration model. Then, we tested its performances on an in- and cross-domain scenario and we also offer a comparison with the same model trained on the English language.

The rest of the paper is organized as follows: in Sec. 2 we present related works, in Sec. 3 we introduce the setting, models and data used for the experiments, in Sec. 4 discuss the obtained results and in Sec. 5 we conclude the paper.

Contributions In this paper we: i) investigate the impact of different training sizes on the performance of a punctuation restoration model based on the Transformer architecture; ii) we test the performance of the model in different scenarios (in- and cross-domain); iii) we compare the results obtained in Italian with those obtained with an English model; iv) we inspect the most common errors emerged during the experiments.

2 Related Work

Punctuation restoration is a well known task, especially in Speech Processing and Machine Translation, where many approaches have been tested to tackle the problem in the past decades. In early attempts, acoustic features has been exploited to train finite-states or Hidden Markov Models [10, 5, 12]: the basic idea was to model prosody from speech data and use pauses as cues for sentence boundary, thus as signal of full stop punctuation marks. While prosody is useful in some cases, most of the time cannot be used to place punctuation in an ASR output because speakers use pauses in speech not only to shape the rhythm of their communication, but also for physical needs (e.g. breathing) or hesitations.

[3] https://cloud.google.com/speech-to-text/
[4] https://azure.microsoft.com/en-us/services/cognitive-services/speech-to-text/
[5] https://alphacephei.com/vosk/

To solve this problem, multimodal models have been proposed, making use of parallel audio and transcripts as training [26, 11]. Most of these approaches take Language Models scores, tokens or POS tags of a huge amount of continuous words as the textual features, and exploit pause, pitch contour, energy and prosody as principal acoustic features [16]. With the rise of Deep Learning techniques, many works reported good performances by training Deep Neural Networks with parallel acoustic and textual features [4, 29, 14, 15]

Obviously, multimodal models need a discrete amount of parallel audios and texts, and outside the English World it is not trivial at all to find such data. Given that, many works have focused on textual-only approaches [30, 34, 13].

More recently, the potential of Transformer-based Neural Language Models (NLMs) have been exploited in several studies [32, 33, 19]. For instance, [18] used a pre-trained BERT model [7] with bidirectional LSTM and a CRF layer to achieve state-of-the-art results on the reference transcriptions of the IWSLT2012 dataset[6]. [33], instead, proposed an adversarial multitask learning approach with auxiliary part-of-speech tagging using a pre-trained BERT model.

While the vast majority of this research is focused on the English language, relatively little work has been done to inspect the potential of these models on other languages. [9] proposed a method based on Chinese punctuation prediction by combining the BERT model with a BiLSTM that outperformed the baseline by up to 31% absolute in overall micro-F1 on a Chinese news dataset. The study by [1], from which we built our experiments for the Italian language, explored different Transformer-based models and propose an augmentation strategy for the punctuation restoration task both on high- (English) and low-resource (Bangla) languages.

3 Experimental Setting

We explored the potential of transformer based language models for the punctuation restoration task on the Italian language. Specifically, we defined two sets of experiments. The first consists in evaluating the impact of the fine-tuning set size on the task performances. For that purpose, we tested the performance of a state-of-the-art transformer based architecture for punctuation restoration [1] with incremental fine-tuning sizes.

In the second set of experiments, we compared the performances of two differently fine-tuned models on 4 test datasets, as explained in 3.2. Moreover, we proposed an error analysis in order to investigate strength and weakness of the proposed methodology.

Model and datasets used for the experiments are described below.

3.1 Model

We relied on the architecture previously defined in [1]. The architecture is based on a Transformer model from which the internal representations are then used

[6] http://hltc.cs.ust.hk/iwslt/index.php/evaluation-campaign/ted-task.html

as input for a BiLSTM layer, consisting of 768 hidden units. The outputs of the BiLSTM layer are then concatenated at each time step to a fully connected layer with four output neurons: one for the O (Other) class and three for the punctuation marks of Comma (C), Period (P) and Question (Q). Thus, this model casts the punctuation restoration problem as a classification problem: the output is basically a class assigned to each token.

The pre-trained Transformer used in our experiments is the XXL uncased version of the BERT model for the Italian language developed by the MDZ Digital Library Team and available trough the Huggingface's *Transformers* library [31][7]. The model was trained on Italian Wikipedia and texts from the OPUS [28] and OSCAR [27] corpora. We will refer to the model as BERT-BiLSTM.

3.2 Data

The model has been fine-tuned on two corpora, in order to evaluate divergences in the results with respect to the domain variation deriving from different data. The first corpus is a large collection of authentic contemporary texts in Italian derived from the web, and it is the *de-facto* reference corpus for Italian in many NLP applications: the Italian Web as Corpus (ItWaC) [3]. It counts 2 billion words and it has been built from the Web by limiting the crawl to the *.it* domain, and using as seeds medium-frequency words from La Repubblica journalistic corpus [2] and *Il Nuovo Vocabolario di Base* (NVdB - list of basic words of Italian) [6]. Given the extension and the origin, the ItWaC corpus spans across many domains. It contains texts with registries that vary from colloquial (i.e. texts derived from forums and social media) to highly formal (i.e. official documents, newspapers, technical descriptions),and the use of punctuation varies accordingly.

The second corpus used for the fine-tuning is the Italian sub-corpus of the Opensubtitles Multilingual Corpus [17].[8] This huge corpus has been compiled from a large database of movie and TV subtitles collected from the Opensubtitles website,[9] and includes a total of 1,689 bi-texts spanning 2.6 billion sentences across 60 languages. The Italian-only subcorpus consists of a total of 769.5 millions of words. Language of movies and television has been often defined as *broadcast-spoken* [20, 22], that is a variety of language that sits in the middle between written and spoken. More specifically, broadcast-spoken is characterised by the fact that it is a well programmed language, based on pre-written texts, and performed to mimic spoken variety. Obviously, it lacks features specific of the spontaneous speech, such as hesitations, retracting and fillers, and it shows high regularity, especially in the use punctuation as marks of pauses in the transcription.

By creating two fine-tuned models, we want to investigate if the language diversity observable in the two corpora (i.e. average written language and multi-registry from ItWaC, close-to-spoken but highly regular from Opensubtitles) is reflected in the way the models handle punctuation.

[7] https://huggingface.co/dbmdz/bert-base-italian-xxl-cased
[8] https://opus.nlpl.eu/OpenSubtitles-v2018.php
[9] http://www.opensubtitles.org

Dataset	Sentences	Tokens	Commas	Periods	Questions
ItWaC	765,491	19,226,715 (25.12)	1,403,527 (1.83)	729,806 (0.95)	35,685 (0.05)
Opensubtitles_it	1505279	14,468,346 (9.61)	754,951 (0.5)	1,265,306 (0.84)	239,973 (0.16)
ParlaMint_it	134,887	3,203,374 (23.75)	238,960 (1.77)	130,386 (0.97)	4,501 (0.03)
TEDx_it	1,139	21,667 (19.02)	1,823 (1.6)	1,070 (0.94)	69 (0.06)
TEDx_en	1,210	21,383 (17.67)	1,636 (1.35)	1,142 (0.94)	68 (0.06)

Table 1. Statistics on the datasets used for fine-tuning and test. In parenthesis, the average distribution per sentence.

Class	Punctuation Marks
COMMA	, ; — - ( )
PERIOD	. : ! ...
QUESTION	?

Table 2. Mapping of punctuation marks to reduced classes for model fine-tuning.

Moreover, we considered other two resources for the purpose of evaluating the two models performances in a cross-domain scenario. The first resource is the Italian part of the ParlaMint Comparable Corpora [8], which contains transcriptions of parliamentary debates from 2015 to mid-2020, counting about 20 millions of words.[10] Given the context of the texts, language is highly formal, and thus also the use of punctuation in the transcripts is precise and regular.

With the second *test-only* resource we also introduce a multilingual setting, useful to compare models performances with reference systems available for English language. We used the Italian-English alignment of the Multilingual TEDx Dataset [24, 23],[11], which is a collection of audio recordings from TEDx talks in 8 source languages. The Italian-English alignment derives from transcriptions of Italian TEDx speeches with aligned English translations, and it counts about 18 thousands words in both languages.

Table 1 reports some numbers about the datasets herein described. These statistics refer to the whole set of texts processed, and for all the experiments conducted with different size of fine-tuning a random selection of sentences has been collected.

Data pre-processing The model implemented in our experiments is trained on a classification of tokens on the basis of the presence or absence of a punctuation mark immediately after the target token. It is important to remember that punctuation is a feature of the written language modality, and it is used to mimic oral pauses in the transcription of speech: commas are used for short pauses, periods for long pauses at the end of an utterance and question marks for questions. For this reason, we collapsed all the possible punctuation marks to these 3 classes, reducing the complexity of the fine-tuning data.

[10] The complete collection of comparable corpora in 17 languages is available at: https://www.clarin.si/repository/xmlui/handle/11356/1432

[11] The full dataset is available at: http://www.openslr.org/100/

Table 2 shows how each punctuation mark has been mapped to the corresponding class. The majority of the symbols have been mapped to COMMA because normally they are used to signal parenthetical clauses and do not interrupt the sentence, while exclamation mark and suspension points, which signals sentence boundaries, are assimilated to the PERIOD class, but question marks have been considered as a separate class (QUESTION), in order to keep the distinction between questions and assertions. Along with these 3 classes, the class OTHER have been used to annotate tokens not followed by a punctuation mark. We are aware that this mapping and reduction could be simplistic, but, again, we are targeting our experiments towards speech transcription, where no punctuation at all exists, and we need to account for all possible punctuation found in the training data.

To feed the model with data in the correct format, we previously pos-tagged all the corpora with Stanza [21] in order to easily recognise punctuation marks and thus label correctly tokens followed by them. To better explain, consider the following sentence taken from the TEDx English test dataset data: *During my career, I had responsibilities and many satisfactions.*. This sentence has been converted as shown in Table 3. During the process, we also lower-cased all tokens to avoid the possibility of predicting full stops (such as periods, exclamation marks and suspension points) on the basis of the casing of the following token.

During my career, I had responsibilities and many satisfactions.									
during	my	career	i	had	responsibilities	and	many	satisfactions	
O	O	COMMA	O	O	O	O	O	PERIOD	

Table 3. Example of a pre-processed sentence for the fine-tuning.

4 Results

We first investigate the impact of different training sizes on the performance of BERT-BiLSTM. In order to do so, we fine-tuned our punctuation restoration models in parallel, with increasingly large portions of the two corpora, from 100k to 10 million tokens, and then tested them on a previously unseen portion of the two datasets consisting of 200k words. Results (in terms of micro F-score) are reported in Figure 1. As a general remark, we found that, for both models, the curve tends to flatten out when the fine-tuning process is performed with portions larger than 2 million tokens. As regards the differences between the two datasets, we can notice that the model fine-tuned on Opensubtitles performed slightly better than the one trained on ItWac. For instance, focusing on the results obtained in the last run (10 million tokens) we can observe that the difference between the two models in terms of F-score is about 0.05 points (0.75 vs. 0.70). Moreover, it is interesting to note that while the Opensubtitles model obtained

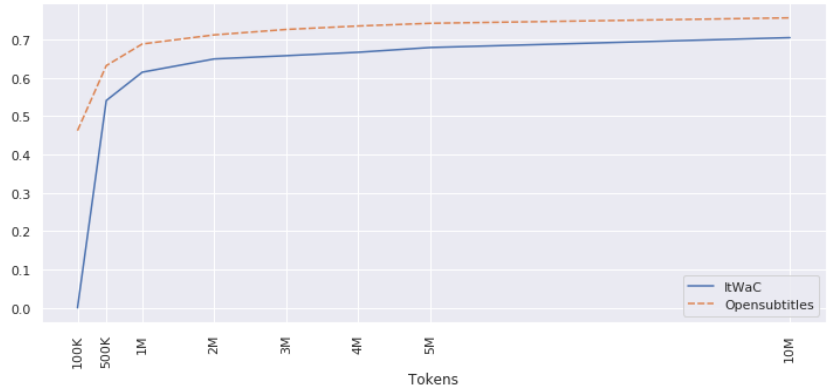

Fig. 1. Average micro F-scores obtained with increasing ItWaC and Opensubtitles dataset sizes.

quite good results even with very small portions of the dataset (e.g. 100K), the itWac model requires at least one million words to achieve comparable results. This behaviour is quite predictable due to the fact that Opensubtitles texts are extremely regular and minimum variation is appreciable through the whole set of data. On the contrary, using ItWaC that is more heterogeneous, the system need more data to start correctly modelling the distribution of punctuation marks.

To better investigate their performances, we report in Table 4 the results obtained by the two models fine-tuned with 10 millions words from Opensubtitles/ItWac and tested in two different scenarios: i) in-domain, i.e. testing on the same dataset; ii) cross domain, i.e. testing on the other domain. Moreover, in order to provide a direct comparison between the two models, we tested both their performance on the Parlamint datasets.

As it can bee seen by looking at the average scores (column *Avg*), the in-domain configuration always achieves the best results (ItWaC: 0.65; Opensubtitles: 0.73). By focusing on the cross-domain configurations, it is interesting to notice that the high variability of ItWaC texts strengthens the model and enables it to handle punctuation with better performances with respect to the model fine-tuned on Opensubtitles. Specifically, observing the performances of ItWaC model on Opensubtitles testset and viceversa, we notice a difference of 0.05 points. While, looking at both models (ItWaC and Opensubtitles) tested on ParlaMint, the gap increases to 0.07 points in favour of the ItWaC model. We can explain this behaviour on the basis of the nature of the ParlaMint dataset, where regularity and formality leads to longer sentences with punctuation usage closer to average written texts. Thus, ItWaC model, which is based on an heterogeneous collection of texts larger than Opensubtitles, is capable of predicting punctuation in a more robust way.

Looking at per-class scores, it is possible to notice that all systems perform better in predicting the PERIOD class: with exclusion of the Opensubtitles

Test Set	Other			Comma			Period			Question			Avg (CPQ)		
ItWaC Fine-tuning															
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
ItWaC	.96	.97	.97	.73	.67	.70	.70	.72	.71	.52	.59	.55	.65	.66	.65
Opensubtitles	.96	.98	.97	.62	.36	.46	.68	.74	.71	.54	.61	.58	.61	.57	.58
ParlaMint	.97	.98	**.98**	.80	.70	**.74**	.77	.83	**.80**	.54	.63	.58	.70	.72	.70
Opensubtitles Fine-tuning															
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
Opensubtitles	.97	.98	**.98**	.74	.64	.69	.80	.80	**.80**	.75	.69	**.72**	.76	.71	**.73**
ItWaC	.95	.96	.96	.62	.45	.52	.50	.67	.57	.56	.45	.50	.56	.52	.53
ParlaMint	.97	.98	**.98**	.75	.59	.66	.65	.80	.72	.48	.55	.51	.63	.64	.63

Table 4. Results (Precision, Recall and F-score) on Opensubtitles/ItWaC and Parlamint datasets when the fine-tuning is performed on 10 million words of the ItWac and Opensubtitles datasets. Average scores (*Avg* column) are computed by averaging *C*, *P* and *Q* scores. Higher F_1 scores per class, across all the models and runs, are in bold.

model tested on ItWaC, all scores are above 0.70. This result is encouraging because periods, exclamation marks and other full stops are used to signal the end of a sentence, thus a similar model can be effectively exploited to tackle the task of segmenting the continuous flow of speech transcription, enabling better subsequent sentence-based methods of analysis (e.g. part-of-speech tagging, dependency parsing and so on).

We register lower figures on the QUESTION class. It is probably due to the unpredictability of these in Italian only on the basis of transcribed text, without considering intonation. We further investigate this problem in 4.2.

4.1 Model comparison in multilingual setting

As already mentioned in Sec. 3.2, we also decided to compare the performance of our fine-tuned models with a reference system available for the English language. Specifically, we compared the results obtained with the ItWaC/Opensubtitles models when tested on the Italian transcriptions of TEDx speeches with the ones obtained by the system devised in [1] and tested on the TEDx aligned English translations. Results are reported in Table 5.

Test Set	Other	Comma	Period	Question	Avg (CPQ)
Alam et al.[1] (EN)	0.97	0.65	0.78	0.80	0.74
ItWaC (IT)	0.97	0.60	0.70	0.64	0.64
Opensubtitles (IT)	0.97	0.56	0.69	0.58	0.61

Table 5. F-scores obtained respectively by the [1] model tested on English TEDx translations and by the ItWaC/Opensubtitles models tested on Italian TEDx transcriptions.

As noticed in the previous experiments, the model fine-tuned on ItWaC data achieve better results when tested in a cross-domain scenario. In fact, we can observe a difference of about 0.03 points in terms of average F-scores. In this respect, it is interesting to note that the main difference between the performance of the two models is due to the classification of question marks. Focusing instead on the comparison between Italian and English models, we can clearly observe that the latter outperforms the Italian ones. Also in this case, the classification of question marks is the one that contributed most to the score difference between the models. This result is quite expected, since a question in English, beside the presence of a question mark at the end of the sentence, is usually characterised by an inversion of the subject and the verb in the principal clause. On the contrary, in Italian the punctuation mark is the only discriminating feature of questions in the written modality, while intonation plays the main role in spoken Italian. Therefore, the identification of question marks tends to be much easier for the English language.

4.2 Error Analysis

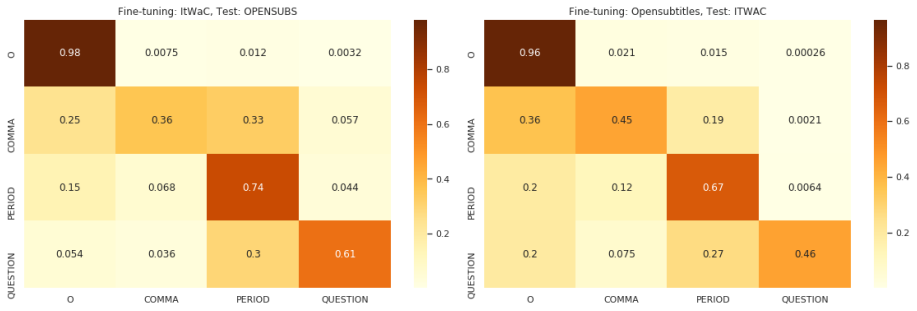

Fig. 2. Confusion matrices of the results obtained by the model fine-tuned on ItWaC and tested on Opensubtitles (*Fine-tuning: ItWaC, Test: OPENSUBS*) and vice-versa (*Fine-tuning: Opensubtitles, Test: ITWAC*).

In order to further deepen our analysis, in this section we investigate in more detail the main errors made by the two models when predicting the different punctuation marks. In Figure 2 we report the confusion matrices (in terms of accuracy) of the results obtained by the model fine-tuned on ItWaC and tested on Opensubtitles and vice-versa. As a general remark, we can highlight that the COMMA class is the most confused in both models. Due to the unbalanced distribution of the O (Other) class with respect to the punctuation classes, the high confusion of every class with this one is easily predictable. Thus, if we exclude the O class from the figure, we can notice that the class with which the COMMA is often confused is PERIOD, for both models tested on the opposite

dataset (e.g. model fine-tuned on ItWaC and tested on Opensubtitles and vice versa). We can ascribe this problem to the average length of sentences, that diverges between the two: in ItWaC, the average sentence counts 25.12 tokens with about 1.83 commas per sentence; in Opensubtitles, the average sentence is 9.61 tokens long, with a distribution of commas of 0.5. For this reason, we can assume that the Opensubtitles model tends to create shorter sentences, thus using the full stop mark more frequently than the ItWaC one.

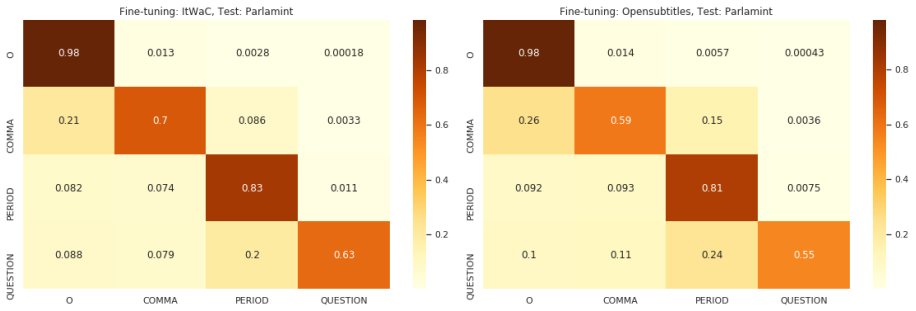

Fig. 3. Confusion matrices of the results obtained by the models fine-tuned on ItWac/Opensubtitles and tested on ParlaMint.

Figure 3 reports instead the confusion matrices of the results obtained by the two fine-tuned models (ItWaC and Opensubtitles) on ParlaMint test data. As we have seen previously, the model fine-tuned on ItWaC is the one that achieved better results regardless of the class taken into account. In fact, with the exception of *Other* (*O*), in all the other classes we observe a performance gap that goes from 0.2 (*PERIOD*) to 0.11 (*COMMA*) accuracy points. Focusing on the mismatched classes, we can see once again that commas are often mistaken as other tokens (*O*), while question marks are classified as periods, as in the following example:

> **Original**: Vorrei ricordarvi i fallimenti ai quali siete andati incontro e state continuamente andando incontro con i bonus. Devo ricordarvi, forse il bonus vacanze? [en. *I would like to remind you of the failures you have experienced and you are continually experiencing with the bonuses. Should I remind you the holiday bonus?*]
> **ItWaC/Opensubtitles**: Vorrei ricordarvi i fallimenti ai quali siete andati incontro e state continuamente andando incontro con i bonus. devo ricordarvi, forse il bonus vacanze.

If we look at the differences between the two models, we can clearly notice that the one fine-tuned on Opensubtitles tends to wrongly classify in 0.15 of the cases a comma also as a full stop, as in the following example:

Original: Scusate la digressione: pure io sono un mancato operaio, due braccia rubate all'agricoltura - allora lo si diceva in senso denigratorio, mentre oggi tale definizione si è qualificata un po' di più - e ho potuto permettermi di studiare e di laurearmi. [en. *Sorry for the digression: I am too a non-working class person, two arms stolen from agriculture - at the time this was said in a derogatory sense, whereas today this definition has been requalified - and I was able to afford my studies and my degree.*]
ItWaC: Scusate la digressione, pure io sono un mancato operaio, due braccia rubate all'agricoltura, allora lo si diceva in senso denigratorio, mentre oggi tale definizione si è qualificata un po' di più e ho potuto permettermi di studiare e di laurearmi
Opensubtitles: Scusate la digressione. pure io sono un mancato operaio. due braccia rubate all'agricoltura. allora lo si diceva in senso denigratorio, mentre oggi tale definizione si è qualificata un po' di più e ho potuto permettermi di studiare e di laurearmi

From the previous example, we can also highlight that the hyphens were correctly classified as COMMA by the ItWaC model (punctuation marks to class mapping in Table 2), while they were identified as full stops by the Opensubtitles one. This could be due to the fact that since the Opensubtites dataset is composed of shorter sentences (derived from transcribed dialogic turns), the model tends to extend this behaviour on its inferences. Conversely, the colon were correctly identified as PERIOD by the Opensubtitles models.

5 Conclusions

In this paper we verified if it could be possible to obtain good results in restoring punctuation in raw transcription texts by means of a fine-tuned Transformers model. We chose to exploit 2 corpora as fine-tuning, namely ItWaC and Opensubtitles, in order to observe the differences emerging from domain variety and their projection on performances.

First, we evaluated the impact of different sizes of fine-tuning datasets, and we observed that the model fine-tuned on highly regular data (i.e. Opensubtitles) need less information to start modelling punctuation with regards to the model fine-tuned on more heterogeneous data (i.e. ItWaC); for both models, the curve tends to flatten out with fine-tuning portions larger than 2 million tokens. Moreover, the model fine-tuned on ItWaC obtains the best results when tested cross-domain on ParlaMint dataset, which is used as neutral testing field for both models.

Lately, we offered a comparison between the Italian models herein fine-tuned and the English model originally presented in [1], tested on the parallel it-en part of the TEDx dataset. With this comparison it has been possible to easily interpret the errors deriving from the confusion between question marks and periods, that is problematic in Italian due to the lack of strong syntactic cues, such as in English, and for this reason it is extremely difficult to distinguish between question marks and periods considering textual information uniquely. In conclusion,

a precise punctuation restoration with Transformers based models is a difficult task, but considering the good results in predicting periods positioning, we can confirm that it is possible to mark sentence boundaries and thus segmenting in sentences the continuous flow of speech transcription.

References

1. Alam, T., Khan, A., Alam, F.: 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). pp. 132–142. Association for Computational Linguistics, Online (2020). https://doi.org/10.18653/v1/2020.wnut-1.18, https://aclanthology.org/2020.wnut-1.18
2. Baroni, M., Bernardini, S., Comastri, F., Piccioni, L., Volpi, A., Aston, G., Mazzoleni, M.: Introducing the La Repubblica Corpus: A Large, Annotated, TEI (XML)-compliant Corpus of Newspaper Italian. In: LREC (2004)
3. Baroni, M., Bernardini, S., Ferraresi, A., Zanchetta, E.: The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. Language Resources and Evaluation **43**(3), 209–226 (2009)
4. Che, X., Luo, S., Yang, H., Meinel, C.: Sentence boundary detection based on parallel lexical and acoustic models. In: Interspeech. pp. 2528–2532 (2016)
5. Christensen, H., Gotoh, Y., Renals, S.: Punctuation annotation using statistical prosody models. (2001)
6. De Mauro, T.: Il nuovo vocabolario di base della lingua italiana. In: Guida all'uso delle parole. Editori Riuniti (1980)
7. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: 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). pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019). https://doi.org/10.18653/v1/N19-1423, https://aclanthology.org/N19-1423
8. Erjavec, T., Ogrodniczuk, M., Osenova, P., Ljubešić, N., Simov, K., Grigorova, V., Rudolf, M., Pančur, A., Kopp, M., Barkarson, S., Steingrímsson, S., van der Pol, H., Depoorter, G., de Does, J., Jongejan, B., Haltrup Hansen, D., Navarretta, C., Calzada Pérez, M., de Macedo, L.D., van Heusden, R., Marx, M., Çöltekin, Ç., Coole, M., Agnoloni, T., Frontini, F., Montemagni, S., Quochi, V., Venturi, G., Ruisi, M., Marchetti, C., Battistoni, R., Sebők, M., Ring, O., Darġis, R., Utka, A., Petkevičius, M., Briedienė, M., Krilavičius, T., Morkevičius, V., Diwersy, S., Luxardo, G., Rayson, P.: Multilingual comparable corpora of parliamentary debates ParlaMint 2.1 (2021), http://hdl.handle.net/11356/1432, slovenian language resource repository CLARIN.SI
9. Fang, M., Zhao, H., Song, X., Wang, X., Huang, S.: Using bidirectional lstm with bert for chinese punctuation prediction. In: 2019 IEEE International Conference on Signal, Information and Data Processing (ICSIDP). pp. 1–5 (2019). https://doi.org/10.1109/ICSIDP47821.2019.9172986
10. Gotoh, Y., Renals, S.: Sentence boundary detection in broadcast speech transcripts. In: ASR2000-Automatic Speech Recognition: Challenges for the new Millenium ISCA Tutorial and Research Workshop (ITRW) (2000)

11. Gravano, A., Jansche, M., Bacchiani, M.: Restoring punctuation and capitalization in transcribed speech. In: 2009 IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 4741–4744. IEEE (2009)
12. Kim, J.H., Woodland, P.C.: A combined punctuation generation and speech recognition system and its performance enhancement using prosody. Speech Communication **41**(4), 563–577 (2003)
13. Kim, S.: Deep recurrent neural networks with layer-wise multi-head attentions for punctuation restoration. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 7280–7284. IEEE (2019)
14. Klejch, O., Bell, P., Renals, S.: Punctuated transcription of multi-genre broadcasts using acoustic and lexical approaches. In: 2016 IEEE Spoken Language Technology Workshop (SLT). pp. 433–440. IEEE (2016)
15. Klejch, O., Bell, P., Renals, S.: Sequence-to-sequence models for punctuated transcription combining lexical and acoustic features. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 5700–5704. IEEE (2017)
16. Levy, T., Silber-Varod, V., Moyal, A.: The effect of pitch, intensity and pause duration in punctuation detection. In: 2012 IEEE 27th Convention of Electrical and Electronics Engineers in Israel. pp. 1–4. IEEE (2012)
17. Lison, P., Tiedemann, J.: Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles (2016)
18. Makhija, K., Ho, T.N., Chng, E.S.: Transfer learning for punctuation prediction. In: 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). pp. 268–273. IEEE (2019)
19. Nagy, A., Bial, B., Ács, J.: Automatic punctuation restoration with bert models. arXiv preprint arXiv:2101.07343 (2021)
20. Nencioni, G.: Parlato-parlato, parlato-scritto, parlato-recitato. Strumenti critici **29** (1976)
21. Qi, P., Zhang, Y., Zhang, Y., Bolton, J., Manning, C.D.: Stanza: A python natural language processing toolkit for many human languages. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. pp. 101–108 (2020)
22. Sabatini, F.: La comunicazione orale, scritta e trasmessa. In: Boccafurni, A.M., Serromani, S. (eds.) Educazione linguistica nella scuola superiore: sei argomenti per un curricolo, pp. 105–27 (1982)
23. Salesky, E., Wiesner, M., Bremerman, J., Cattoni, R., Negri, M., Turchi, M., Oard, D.W., Post, M.: Multilingual tedx corpus for speech recognition and translation (2021), http://www.openslr.org/100/
24. Salesky, E., Wiesner, M., Bremerman, J., Cattoni, R., Negri, M., Turchi, M., Oard, D.W., Post, M.: The multilingual TEDx corpus for speech recognition and translation. arXiv:2102.01757 (2021)
25. Schneider, S., Baevski, A., Collobert, R., Auli, M.: wav2vec: Unsupervised pre-training for speech recognition. arXiv preprint arXiv:1904.05862 (2019)
26. Stolcke, A., Shriberg, E., Bates, R.A., Ostendorf, M., Hakkani, D.Z., Plauche, M., Tür, G., Lu, Y.: Automatic detection of sentence boundaries and disfluencies based on recognized words. In: ICSLP. vol. 2, pp. 2247–2250. Citeseer (1998)
27. Suárez, P.J.O., Sagot, B., Romary, L.: Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures. Challenges in the Management of Large Corpora (CMLC-7) 2019 p. 9 (2019)
28. Tiedemann, J., Nygaard, L.: The opus corpus-parallel and free: http://logos. uio. no/opus. Citeseer (2004)

29. Tilk, O., Alumäe, T.: Lstm for punctuation restoration in speech transcripts. In: Sixteenth annual conference of the international speech communication association (2015)
30. Tilk, O., Alumäe, T.: Bidirectional recurrent neural network with attention mechanism for punctuation restoration. In: Interspeech. pp. 3047–3051 (2016)
31. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., Rush, A.: Transformers: State-of-the-art natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 38–45. Association for Computational Linguistics, Online (Oct 2020). https://doi.org/10.18653/v1/2020.emnlp-demos.6, https://www.aclweb.org/anthology/2020.emnlp-demos.6
32. Yi, J., Tao, J.: 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). pp. 7270–7274. IEEE (2019)
33. Yi, J., Tao, J., Bai, Y., Tian, Z., Fan, C.: Adversarial transfer learning for punctuation restoration. arXiv preprint arXiv:2004.00248 (2020)
34. Yi, J., Tao, J., Wen, Z., Li, Y., et al.: Distilling knowledge from an ensemble of models for punctuation prediction. In: Interspeech. pp. 2779–2783 (2017)