Deep Contextual Punctuator for NLG Text

Vandan Mujadia Pruthwik Mishra Dipti Misra Sharma
Language Technologies Research Center
Kohli Center On Intelligent Systems, IIIT Hyderabad
{vandan.mu, pruthwik.mishra}@research.iiit.ac.in, dipti@iiit.ac.in

Abstract

This paper describes our team oneNLP's (LTRC, IIIT-Hyderabad) participation for the SEPP-NLG 2021 tasks[1], Sentence End and Punctuation Prediction in NLG Text-2021. We applied sequence to tag prediction over contextual embedding as fine-tuning for both of these tasks. We also explored the use of multilingual Bert and multitask learning for these tasks on English, German, French and Italian.

1 Introduction

Generally, the output of automatic speech recognition (ASR) systems ignore the prediction of punctuation marks. Similarly, output of OCR systems (Nguyen et al., 2019) need automatic validation for punctuation. Apart from the omission of punctuation markers, some automatic tools generated texts e.g. PDF to text extraction may erroneously displace sentences for several reasons. Here, detecting the end of a sentence and placing an appropriate punctuation mark significantly improves the quality of such outputs by preserving the original meaning. Thus, missing punctuation or inappropriate punctuation degrade the readability of the presented text and leads to poor user experiences in real-world scenarios (Che et al., 2016; Ueffing et al., 2013) as well as erroneous input to the subsequent automatic systems such as Machine Translation, Summarization, Question Answering, NLU etc. Therefore it is necessary to restore or correct punctuation marks for these automatic outputs.

Traditionally, automatic punctuation marking approaches (Lu and Ng, 2010) can be divided into three broad categories (Vandeghinste et al., 2018) based on the used features. They can be prosody based features (Kim and Woodland, 2001; Christensen et al., 2001), lexical features (Augustyniak et al., 2020; Peitz et al., 2014) or combined or hybrid features of the previous two features based methods. Recent lexical based punctuation prediction methods build upon deep neural networks where it is modeled as a sequence to tag prediction task (Li and Lin, 2020) or a sequence to sequence prediction task (Vandeghinste et al., 2018).

The simplest and basic form of punctuation prediction is the discovery of sentence boundaries, here the problem is the binary classification (where classes are period or empty as label). An incremental and a bit harder problem is the prediction of each individual punctuation, here the class labels for subtask2 are ": - , ? . 0" (0 indicating no punctuation). SEPP-NLG 2021 presents both these tasks as a challenge for the English, German, French and Italian languages.

In a recent advance of deep learning, pretrained language models such as ELMo (Peters et al., 2018), ULMFiT (Howard and Ruder, 2018), OpenAI Transformer (Lee and Hsiang, 2020) and BERT (Devlin et al., 2018) have resulted in a massive jump in the state-of-the-art performance for many NLP tasks, i.e text classification (Büyüköz et al., 2020), natural language inference and question-answering, dialogue system (Budzianowski and Vulić, 2019) etc. All these approaches pre-train an unsupervised language model on a large corpus of data such as all wikipedia articles, news articles and then fine-tune these pretrained models on different downstream tasks.

Here, for our experiments on the two punctuation prediction tasks, we try to use multi-lingual Bert and ALBert (for English) as a fine-tuning task along with the baseline experiments with CRF.

2 Dataset

As a part of SEPP-NLG 2021, the organizers released an Europarl corpus of spoken texts by lower-

[1] https://sites.google.com/view/sentence-segmentation/

Lang	#Train_Sents	#Train_Toks	Avg_Train_Sent_Len	#Dev_Sents	#Dev_Toks
English	1406577	33779095	24.015	321333	7743489
German	1308508	28645112	21.891	291443	6358683
French	1236504	32690367	26.438	332330	8781593
Italian	1132554	28167993	24.871	290089	7194189

Table 1: Training and Development Data Detail

Class	#Count	#Percentage
:	43133	0.128
-	80916	0.240
,	1759686	5.209
?	44290	0.131
.	1396166	4.133
0	30454904	90.159
Total	**33779095**	

Table 2: English : Class Details for Training Data

Class	#Count	#Percentage
:	51192	0.179
-	81710	0.285
,	2208970	7.712
?	40511	0.141
.	1290282	4.504
0	24972447	87.179
Total	**28645112**	

Table 3: German : Class Details for Training Data

Class	#Count	#Percentage
:	46128	0.141
-	68523	0.210
,	1657880	5.071
?	41005	0.125
.	1223802	3.744
0	29653029	90.709
Total	**32690367**	

Table 4: French : Class Details for Training Data

Class	#Count	#Percentage
:	55080	0.196
-	52983	0.188
,	1503502	5.338
?	38807	0.138
.	1138669	4.042
0	25378952	90.099
Total	**28167993**	

Table 5: Italian : Class Details for Training Data

casing and removing all punctuations in the transcripts available in multiple languages. Table 1 describes the corpora details for Training and Development corpus for all languages in terms of sentences and Tokens. Table 2, Table 3, Table 4 and Table 5 describe the training corpora details in terms of punctuation classes and their respective distribution. Here data numbers are given after ‘’! ;‘’ are mapped to ‘’.‘’.

3 Approach

We primarily used two broad categories of approaches. We model the problem as a sequence labeling task. In Machine Learning approaches, we trained a CRF model to identify the different kinds of labels correctly. Transformer based BERT fine tuning is also used as the other technique.

3.1 CRF

We split the training data in English into sequences of 25 tokens each. This decision of setting the maximum sequence length to 25 was based on the average sentence length of the training data in English. We only used words as features and utilized a continuous window of 5 words over the full corpus as the required features for the CRF.

3.2 Fine-tuning Contextual Embedding

Multi-task learning (MTL) is a technique which aims to improve generalization, strengthen representations and enable adaptation in machine learning (Worsham and Kalita, 2020) for related tasks. For our case, we enabled multi-task learning for our AlBert and mBert based contextual experiments as presented in Figure 1 where contextual embeddings are shared across the sub-tasks. We applied a sequence to tag classifier on the output contextualized token embeddings of Albert/mBert for the tag prediction. Here, we have used Albert [2] for English

[2]https://tfhub.dev/google/albert_base/3

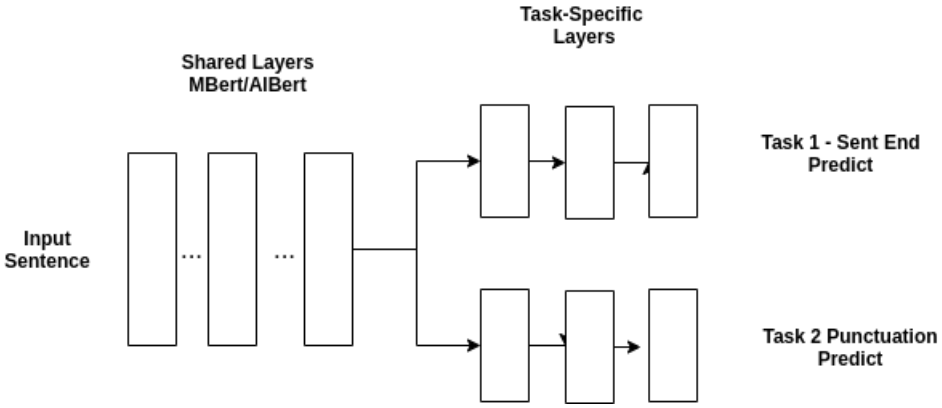

Figure 1: Multi Task Learning

and mBert[3] for other languages.

The following configuration is used for Bert fine-tuning.

- Input : Subword tokens (same as Bert/Albert vocab)
- Embedding size : 512
- Transformer Config : layers (6), hidden_size (2048) attention heads (8)
- Dropout : 0.30, Optimizer : Adam
- max_word_sequence_length : 300
- Fine Tuning Steps : 90K with 40 Batch size

Due to the algorithmic limitations, we were not able to apply MTL for CRF as it does not facilitate the joint learning of multiple classification tasks at one go.

4 Results and Discussion

As the results of CRF with word level features for English were poor (shown in table 7), we did not conduct CRF experiments on other languages.

We could observe that the results of Bert with Multi task learning is superior to the results of CRF. This is due to the better sentence or sequence representations learnt from the transformers. Simple surface level word features fail to capture the end sentence or punctuation markers in CRF.

5 Conclusion and Future work

We have successfully applied contextual embedding for the task of punctuation prediction and achieved comparable results on both of the subtasks. We believe that fine-tuning Bert on more data would benefit the overall punctuation task. Also, the language specific contextual embedding would improve performance in other languages. We will be incorporating both of these points in our future work.

[3]https://tfhub.dev/tensorflow/bert_multi_cased_L-12_H-768_A-12/4

Dataset	Lang	Pr	Re	F1
Test	EN	0.92	0.92	0.92
	DE	0.93	0.95	0.94
	FR	0.9	0.89	0.9
	IT	0.88	0.89	0.89
	AVG	0.91	0.91	0.91
Surprise Test	EN	0.81	0.67	0.73
	DE	0.85	0.72	0.78
	FR	0.77	0.62	0.69
	IT	0.78	0.58	0.67
	AVG	0.8	0.65	0.72
Dev	EN	0.92	0.92	0.92
	DE	0.94	0.95	0.94
	FR	0.9	0.89	0.9
	IT	0.88	0.89	0.89
	AVG	0.91	0.91	0.91

Table 6: Subtask1 Results using BERT MTL

Subtask#	Pr	Re	F1-score
1	0.73	0.52	0.61
2	0.71	0.32	0.35

Table 7: CRF Results of Subtask 1 and 2 for English Dev data

References

Łukasz Augustyniak, Piotr Szymanski, Mikołaj Morzy, Piotr Zelasko, Adrian Szymczak, Jan Mizgajski, Yishay Carmiel, and Najim Dehak. 2020. Punctuation prediction in spontaneous conversations: Can

Dataset	Lang	Pr	Re	F1
Test	EN	0.79	0.69	0.72
	DE	0.8	0.74	0.77
	FR	0.79	0.65	0.68
	IT	0.78	0.62	0.66
	AVG	0.79	0.68	0.71
Surprise Test	EN	0.62	0.52	0.56
	DE	0.61	0.57	0.58
	FR	0.61	0.48	0.51
	IT	0.54	0.43	0.46
	AVG	0.8	0.65	0.72
Dev	EN	0.8	0.69	0.72
	DE	0.81	0.74	0.77
	FR	0.79	0.65	0.69
	IT	0.78	0.62	0.66
	AVG	0.8	0.68	0.71

Table 8: Subtask2 Results using BERT MTL

we mitigate asr errors with retrofitted word embeddings? *arXiv preprint arXiv:2004.05985*.

Paweł Budzianowski and Ivan Vulić. 2019. Hello, it's gpt-2–how can i help you? towards the use of pretrained language models for task-oriented dialogue systems. *arXiv preprint arXiv:1907.05774*.

Berfu Büyüköz, Ali Hürriyetoğlu, and Arzucan Özgür. 2020. Analyzing elmo and distilbert on sociopolitical news classification. In *Proceedings of the Workshop on Automated Extraction of Socio-political Events from News 2020*, pages 9–18.

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

Heidi Christensen, Yoshihiko Gotoh, and Steve Renals. 2001. Punctuation annotation using statistical prosody models.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.

Ji-Hwan Kim and Philip C Woodland. 2001. The use of prosody in a combined system for punctuation generation and speech recognition. In *Seventh European conference on speech communication and technology*.

Jieh-Sheng Lee and Jieh Hsiang. 2020. Patent claim generation by fine-tuning openai gpt-2. *World Patent Information*, 62:101983.

Xinxing Li and Edward Lin. 2020. A 43 language multilingual punctuation prediction neural network model. *Proc. Interspeech 2020*, pages 1067–1071.

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.

Thi-Tuyet-Hai Nguyen, Adam Jatowt, Mickael Coustaty, Nhu-Van Nguyen, and Antoine Doucet. 2019. Deep statistical analysis of ocr errors for effective post-ocr processing. In *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 29–38.

Stephan Peitz, Markus Freitag, and Hermann Ney. 2014. Better punctuation prediction with hierarchical phrase-based translation. In *Proc. of the Int. Workshop on Spoken Language Translation (IWSLT), South Lake Tahoe, CA, USA*.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proc. of NAACL*.

Nicola Ueffing, Maximilian Bisani, and Paul Vozila. 2013. Improved models for automatic punctuation prediction for spoken and written text. In *Interspeech*, pages 3097–3101.

Vincent Vandeghinste, Lyan Verwimp, Joris Pelemans, and Patrick Wambacq. 2018. A comparison of different punctuation prediction approaches in a translation context. In *Proceedings of the 21st Annual Conference of the European Association for Machine Translation: 28-30 May 2018, Universitat d'Alacant, Alacant, Spain*, pages 269–278. European Association for Machine Translation.

Joseph Worsham and Jugal Kalita. 2020. Multi-task learning for natural language processing in the 2020s: Where are we going? *Pattern Recognition Letters*, 136:120–126.