Punctuation Restoration using Transformer Models for High-and Low-Resource languages

Tanvirul Alam, Akib Khan, Firoj Alam

Abstract

Punctuation restoration is a common post-processing problem for Automatic Speech Recognition (ASR) systems. It is important to improve the readability of the transcribed text for the human reader and facilitate NLP tasks. Current state-of-art address this problem using different deep learning models. Recently, transformer models have proven their success in downstream NLP tasks, and these models have been explored very little for the punctuation restoration problem. In this work, we explore different transformer based models and propose an augmentation strategy for this task, focusing on high-resource (English) and low-resource (Bangla) languages. For English, we obtain comparable state-of-the-art results, while for Bangla, it is the first reported work, which can serve as a strong baseline for future work. We have made our developed Bangla dataset publicly available for the research community.

Dataset

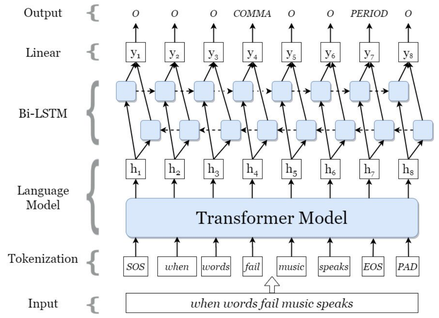
We used IWSLT dataset for English punctuation restoration, which consists of transcriptions from TED Talks. For Bangla, we collected training and development data from news articles. We prepared three test datasets for Bangla: one from news articles, one from manual transcriptions of 65 minutes of speech excerpts, and another from ASR transcriptions of the same speech. There are four classes of interest: (i) Comma: includes commas, colons and dashes, (ii) Period: includes full stops, exclamation marks and semicolons, (iii) Question: only question mark, and (iv) O: for any other token.

Dataset	Total	Period	Comma	Question
Train	1379986	78791(7.16%)	65235(4.73%)	4555(0.33%)
Dev	179371	13161(7.34%)	7544(4.21%)	534(0.3%)
Test(news)	87721	6263(7.14%)	4102(4.68%)	305(0.35%)
Test(Ref.)	6821	996(14.6%)	279(4.09%)	170(2.49%)
Test(ASR)	6417	887(13.82%)	253(3.94%)	125(1.95%)

Distributions of English and Bangla datasets. The number in parenthesis represents Percentage.

Architecture

We fine-tune Transformer language models for the task. Embedding obtained from the models are used as input to a bidirectional LSTM layer. This allows the network to make effective use of both past and future contexts for prediction. The outputs from the forward and backward LSTM layers are concatenated at each time step and fed to a fully connected layer with four output neurons, which correspond to 3 punctuation marks and one O token.



A general model architecture for our experiments.

Augmentation

Due to the lack of large-scale manual transcriptions, punctuation restoration models are typically trained using written text, which is well-formatted and correctly punctuated. Hence, the trained model lacks the knowledge of the typical errors that ASR makes. To train the model with such characteristics, we use an augmentation technique that simulates such errors.

We use three different kinds of augmentation corresponding to three possible errors are as follows:

1. First (i.e., substitution), we replace random tokens with the unknown token.
2. Second (i.e., deletion), we delete some tokens randomly from the processed input sequence.
3. Finally, we add (i.e., insertion) the unknown token at some random position of the input.

We used three tunable parameters: (i) a parameter to determine token change probability, α , (ii) a parameter, α_{sub}, to control the probability of substitution, (iii) a parameter, α_{del} , to control the probability of deletion. Probability of insertion is given by $\mathbf{1 - (\alpha_{sub} + \alpha_{del})}$.

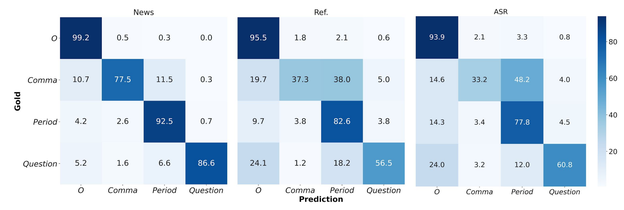
Results

Test	Model	Comma			Period			Question			Overall		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1
Prev	Self-attention	67.4	61.1	64.1	82.5	77.4	79.9	80.1	70.2	74.8	76.7	69.6	72.9
	BERT-Adversarial	76.2	71.2	73.6	87.3	81.1	84.1	79.1	72.7	75.8	80.9	75	77.8
Ref. (Ours)	RoBERTa	76.9	75.8	76.3	86.8	90.5	88.6	72.9	93.5	81.9	81.6	83.3	82.4
	RoBERTa+Aug	76.8	76.6	76.7	88.6	89.2	88.9	82.7	93.5	87.8	82.6	83.1	82.9
Prev	Self-attention	64	59.6	61.7	75.5	75.8	75.6	72.6	65.9	69.1	70.7	67.1	68.8
	BERT-Adversarial	72.4	69.3	70.8	80	79.1	79.5	71.2	68	69.6	74.5	72.1	73.3
ASR (Ours)	RoBERTa	56.6	67.9	61.8	78.7	85.3	81.9	46.6	77.1	58.1	66.5	76.7	71.3
	RoBERTa+Aug	64.1	68.8	66.3	81	83.7	82.3	55.3	74.3	63.4	72	76.2	74

Result on English test datasets.

Test	Model	Comma			Period			Question			Overall		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1
News	XLM-RoBERTa	86	77	81.2	89.4	92.3	90.8	77.4	85.6	81.3	87.8	86.2	87
	XLM-RoBERTa+Aug	85.8	77.5	81.4	88.8	92.5	90.6	77.9	86.6	82	87.4	86.6	87
Ref.	XLM-RoBERTa	39.3	36.9	38.1	76.9	81.4	79.1	54.3	58.8	56.5	67.6	70.2	68.8
	XLM-RoBERTa+Aug	43.3	37.3	40.1	76.5	82.6	79.4	53	56.5	54.7	68.3	70.8	69.5
ASR	XLM-RoBERTa	38.3	35.6	36.9	69.2	77.2	73	38.5	52	44.2	60.3	66.4	63.2
	XLM-RoBERTa+Aug	37.2	33.2	35.1	69.1	77.8	73.2	45.5	60.8	52.1	61.1	67.2	64

Result on Bangla test datasets.



Confusion Matrix (in percentage) for Bangla test datasets.

Conclusion

In this study, we explore different transformer models for high-and low-Resource languages (i.e., English and Bangla). In addition, we propose an augmentation technique, which improves performance on noisy ASR texts. There has not been any reported result and resources for punctuation restoration on Bangla. Our study, findings, and developed resources will enrich and push the current state-of-art for this low-resource language. We have released the created Bangla dataset and code for the research community.