Incorporating External POS Tagger for Punctuation Restoration

Ning Shi, Wei Wang, Boxin Wang, Jinfeng Li, Xiangyu Liu, and **Zhouhan Lin**

{**shining.shi**, luyang.ww, jinfengli.ljf, eason.lxy}@alibaba-inc.com, boxinw2@illinois.edu, **lin.zhouhan@gmail.com**

Punctuation Restoration

Punctuation restoration is one of the many post-processing steps in automatic speech recognition (ASR) that are non-trivial to be dealt with.

Since we can restore the target sequence based on the predicted punctuation tags, the task can be treated as a sequence labeling problem.

Huge efforts have been devoted to investigating better model structures to recover punctuation from raw lexical ASR output, including MLP, CRF, RNNs, CNNs, Transformers, and top layers with pre-train LMs.

An example of pre-processed data to align with BERT (bert-base-uncased).

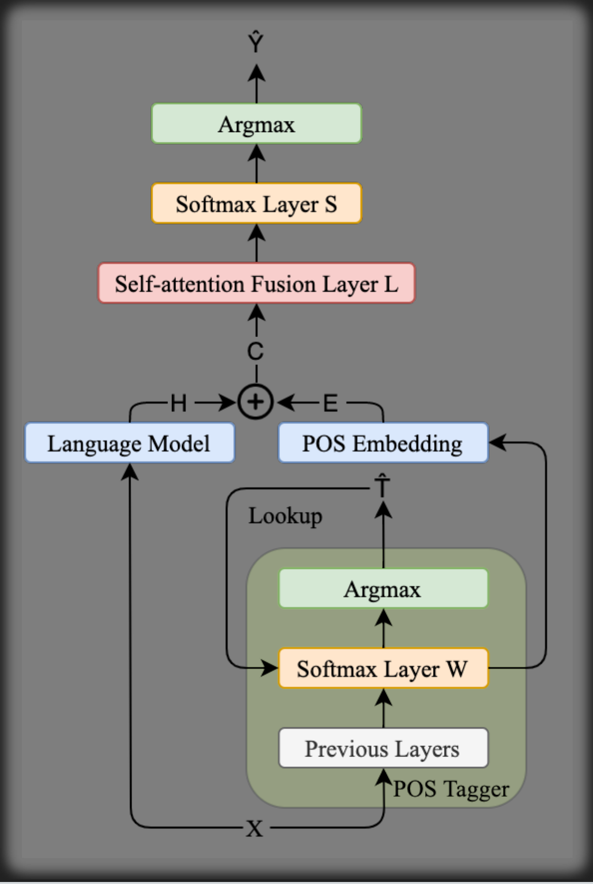
Raw Word Sequence	adrian kohler well we ’re here today to talk about the puppet horse
Raw Label Sequence	O COMMA COMMA O O O O O O O O O PERIOD
Token Sequence (X)	(BOS) [CLS] adrian ko ##hler well we ’ re here today to talk about the puppet horse [SEP] (EOS)
Label Sequence (Y)	O O O COMMA COMMA O O O O O O O O O O PERIOD
Position Mask	0 1 0 1 1 1 0 1 1 1 1 1 1 1 1 1 0

POS Tagger

Whether a word needs to be followed by punctuation is closely related to its grammatical role. For instance, a comma is often placed before the coordinating conjunction to join two independent clauses.

We propose a novel framework that brings POS knowledge via a self-attention based fusion layer for punctuation restoration.

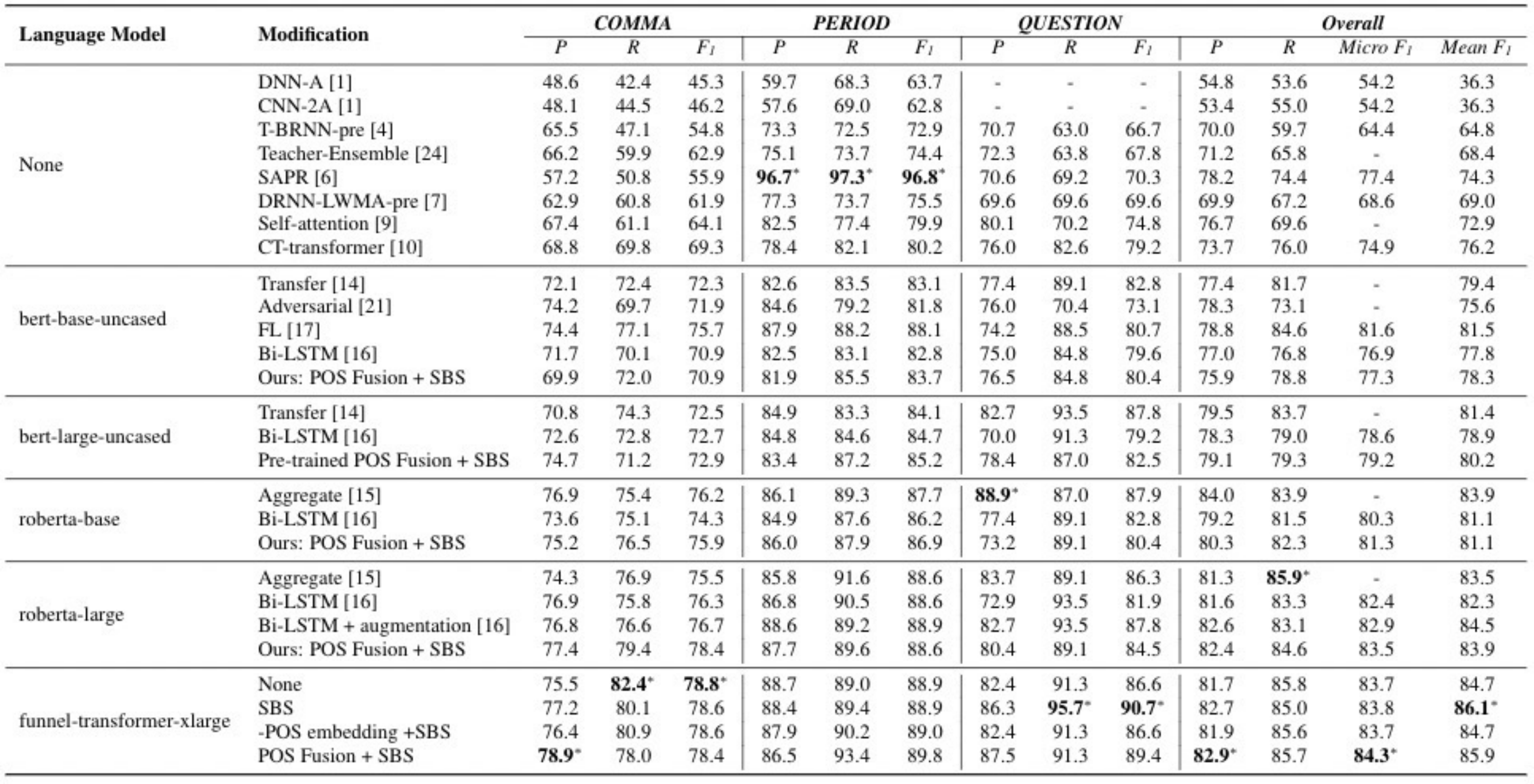
To incorporate T_hat into the LM, we utilize the softmax layer weights W from the POS tagger, and elements in T_hat serve as indexes to lookup for the corresponding columns in W to form a pre-trained POS tag Embedding E.



An example of POS tag sequence (T_hat)

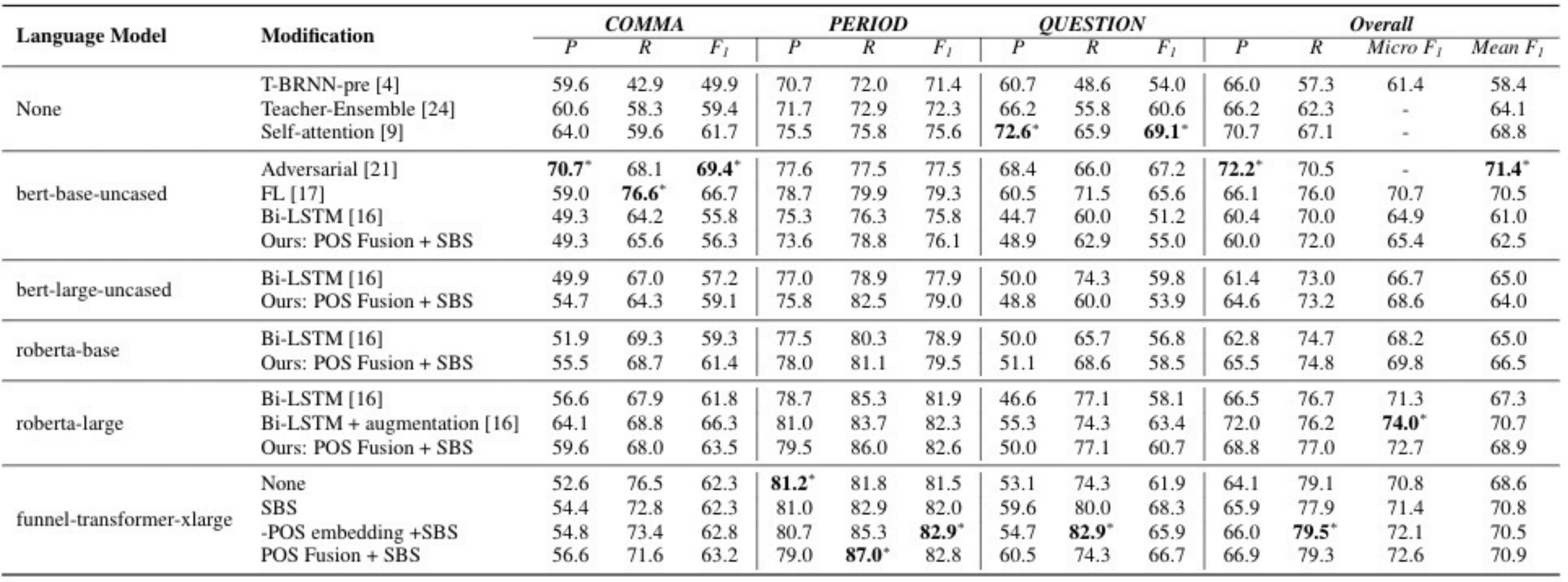
Raw Word Sequence	adrian kohler well we ’re here today to talk about the puppet horse
Raw Label Sequence	O COMMA COMMA O O O O O O O O O PERIOD
POS Tag Sequence (T_hat)	X PROPN X PROPN INTJ PRON X VERB ADV NOUN PART VERB ADP DET NOUN NOUN X

Experiments Results (REF)



Language Model	Modification	*COMMA*			*PERIOD*			*QUESTION*			*Overall*			
		P	*R*	*F_1*	*P*	*R*	*F_1*	*P*	*R*	*F_1*	*P*	*R*	*Micro F_1*	*Mean F_1*
None	DNN-A [1]	48.6	42.4	45.3	59.7	68.3	63.7	-	-	-	54.8	53.6	54.2	36.3
	CNN-2A [1]	48.1	44.5	46.2	57.6	69.0	62.8	-	-	-	53.4	55.0	54.2	36.3
	T-BRNN-pre [4]	65.5	47.1	54.8	73.3	72.5	72.9	70.7	63.0	66.7	70.0	59.7	64.4	64.8
	Teacher-Ensemble [24]	66.2	59.9	62.9	75.1	73.7	74.4	72.3	63.8	67.8	71.2	65.8	-	68.4
	SAPR [6]	57.2	50.8	55.9	**96.7***	**97.3***	**96.8***	70.6	69.2	70.3	78.2	74.4	77.4	74.3
	DRNN-LWMA-pre [7]	62.9	60.8	61.9	77.3	73.7	75.5	69.6	69.6	69.6	69.9	67.2	68.6	69.0
	Self-attention [9]	67.4	61.1	64.1	82.5	77.4	79.9	80.1	70.2	74.8	76.7	69.6	-	72.9
	CT-transformer [10]	68.8	69.8	69.3	78.4	82.1	80.2	76.0	82.6	79.2	73.7	76.0	74.9	76.2
bert-base-uncased	Transfer [14]	72.1	72.4	72.3	82.6	83.5	83.1	77.4	89.1	82.8	77.4	81.7	-	79.4
	Adversarial [21]	74.2	69.7	71.9	84.6	79.2	81.8	76.0	70.4	73.1	78.3	73.1	-	75.6
	FL [17]	74.4	77.1	75.7	87.9	88.2	88.1	74.2	88.5	80.7	78.8	84.6	81.6	81.5
	Bi-LSTM [16]	71.7	70.1	70.9	82.5	83.1	82.8	75.0	84.8	79.6	77.0	76.8	76.9	77.8
	Ours: POS Fusion + SBS	69.9	72.0	70.9	81.9	85.5	83.7	76.5	84.8	80.4	75.9	78.8	77.3	78.3
bert-large-uncased	Transfer [14]	70.8	74.3	72.5	84.9	83.3	84.1	82.7	93.5	87.8	79.5	83.7	-	81.4
	Bi-LSTM [16]	72.6	72.8	72.7	84.8	84.6	84.7	70.0	91.3	79.2	78.3	79.0	78.6	78.9
	Pre-trained POS Fusion + SBS	74.7	71.2	72.9	83.4	87.2	85.2	78.4	87.0	82.5	79.1	79.3	79.2	80.2
roberta-base	Aggregate [15]	76.9	75.4	76.2	86.1	89.3	87.7	**88.9***	87.0	87.9	84.0	83.9	-	83.9
	Bi-LSTM [16]	73.6	75.1	74.3	84.9	87.6	86.2	77.4	89.1	82.8	79.2	81.5	80.3	81.1
	Ours: POS Fusion + SBS	75.2	76.5	75.9	86.0	87.9	86.9	73.2	89.1	80.4	80.3	82.3	81.3	81.1
roberta-large	Aggregate [15]	74.3	76.9	75.5	85.8	91.6	88.6	83.7	89.1	86.3	81.3	**85.9***	-	83.5
	Bi-LSTM [16]	76.9	75.8	76.3	86.8	90.5	88.6	72.9	93.5	81.9	81.6	83.3	82.4	82.3
	Bi-LSTM + augmentation [16]	76.8	76.6	76.7	88.6	89.2	88.9	82.7	93.5	87.8	82.6	83.1	82.9	84.5
	Ours: POS Fusion + SBS	77.4	79.4	78.4	87.7	89.6	88.6	80.4	89.1	84.5	82.4	84.6	83.5	83.9
funnel-transformer-xlarge	None	75.5	**82.4***	**78.8***	88.7	89.0	88.9	82.4	91.3	86.6	81.7	85.8	83.7	84.7
	SBS	77.2	80.1	78.6	88.4	89.4	88.9	86.3	**95.7***	**90.7***	82.7	85.0	83.8	**86.1***
	-POS embedding +SBS	76.4	80.9	78.6	87.9	90.2	89.0	82.4	91.3	86.6	81.9	85.6	83.7	84.7
	POS Fusion + SBS	**78.9***	78.0	78.4	86.5	93.4	89.8	87.5	91.3	89.4	**82.9***	85.7	**84.3***	85.9

Experiments Results (ASR)



Language Model	Modification	*COMMA*			*PERIOD*			*QUESTION*			*Overall*			
		P	*R*	*F_1*	*P*	*R*	*F_1*	*P*	*R*	*F_1*	*P*	*R*	*Micro F_1*	*Mean F_1*
None	T-BRNN-pre [4]	59.6	42.9	49.9	70.7	72.0	71.4	60.7	48.6	54.0	66.0	57.3	61.4	58.4
	Teacher-Ensemble [24]	60.6	58.3	59.4	71.7	72.9	72.3	66.2	55.8	60.6	66.2	62.3	-	64.1
	Self-attention [9]	64.0	59.6	61.7	75.5	75.8	75.6	**72.6***	65.9	**69.1***	70.7	67.1	-	68.8
bert-base-uncased	Adversarial [21]	**70.7***	68.1	**69.4***	77.6	77.5	77.5	68.4	66.0	67.2	**72.2***	70.5	-	**71.4***
	FL [17]	59.0	**76.6***	66.7	78.7	79.9	79.3	60.5	71.5	65.6	66.1	76.0	70.7	70.5
	Bi-LSTM [16]	49.3	64.2	55.8	75.3	76.3	75.8	44.7	60.0	51.2	60.4	70.0	64.9	61.0
	Ours: POS Fusion + SBS	49.3	65.6	56.3	73.6	78.8	76.1	48.9	62.9	55.0	60.0	72.0	65.4	62.5
bert-large-uncased	Bi-LSTM [16]	49.9	67.0	57.2	77.0	78.9	77.9	50.0	74.3	59.8	61.4	73.0	66.7	65.0
	Ours: POS Fusion + SBS	54.7	64.3	59.1	75.8	82.5	79.0	48.8	60.0	53.9	64.6	73.2	68.6	64.0
roberta-base	Bi-LSTM [16]	51.9	69.3	59.3	77.5	80.3	78.9	50.0	65.7	56.8	62.8	74.7	68.2	65.0
	Ours: POS Fusion + SBS	55.5	68.7	61.4	78.0	81.1	79.5	51.1	68.6	58.5	65.5	74.8	69.8	66.5
roberta-large	Bi-LSTM [16]	56.6	67.9	61.8	78.7	85.3	81.9	46.6	77.1	58.1	66.5	76.7	71.3	67.3
	Bi-LSTM + augmentation [16]	64.1	68.8	66.3	81.0	83.7	82.3	55.3	74.3	63.4	72.0	76.2	**74.0***	70.7
	Ours: POS Fusion + SBS	59.6	68.0	63.5	79.5	86.0	82.6	50.0	77.1	60.7	68.8	77.0	72.7	68.9
funnel-transformer-xlarge	None	52.6	76.5	62.3	**81.2***	81.8	81.5	53.1	74.3	61.9	64.1	79.1	70.8	68.6
	SBS	54.4	72.8	62.3	81.0	82.9	82.0	59.6	80.0	68.3	65.9	77.9	71.4	70.8
	-POS embedding +SBS	54.8	73.4	62.8	80.7	85.3	**82.9***	54.7	**82.9***	65.9	66.0	**79.5***	72.1	70.5
	POS Fusion + SBS	56.6	71.6	63.2	79.0	**87.0***	82.8	60.5	74.3	66.7	66.9	79.3	72.6	70.9

Other Contributions

- Sequence Boundary Sampling (SBS) to better adapt to pre-trained LMs.

Since sentence boundaries are not explicit in raw ASR output, the raw ASR can be viewed as a continuous word stream. Thus, we propose SBS, where we uniformly select a range in the corpus to form a token sequence of length L instead of truncation or sliding window.

SBS provides a computationally more efficient process than earlier ways by both weakening the connection between positions and tokens and allowing mini-batches of samples to represent the entire corpus.

- With RoBERTa, our method sets a new state-of-the-art on IWSLT datasets in terms of Micro F1.
- We introduce Funnel Transformer to further push the gap between our method and previous studies.
- As ablation study, we examine a wide range of pre-trained LMs in a fair and comparable setting, which provides a wide set of benchmarks on this task.

Thanks for your attention.

Q&A

Code: https://github.com/ShiningLab/POS-Tagger-for-Punctuation-Restoration