INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Focal Loss for Punctuation Prediction

Jiangyan Yi[1], *Jianhua Tao*[1,2,3], *Zhengkun Tian*[1,3], *Ye Bai*[1,3], *Cunhang Fan*[1,3]

[1]National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences
[2]Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences
[3]University of Chinese Academy of Sciences, Beijing

{jiangyan.yi, jhtao, zhengkun.tian, ye.bai, cunhang.fan}@nlpr.ia.ac.cn

Abstract

Many approaches have been proposed to predict punctuation marks. Previous results demonstrate that these methods are effective. However, there still exists class imbalance problem during training. Most of the classes in the training set for punctuation prediction are non-punctuation marks. This will affect the performance of punctuation prediction tasks. Therefore, this paper uses a focal loss to alleviate this issue. The focal loss can down-weight easy examples and focus training on a sparse set of hard examples. Experiments are conducted on IWSLT2011 datasets. The results show that the punctuation predicting models trained with a focal loss obtain performance improvement over that trained with a cross entropy loss by up to 2.7% absolute overall F_1-score on test set. The proposed model also outperforms previous state-of-the-art models.

Index Terms: focal loss, class imbalance, punctuation prediction, speech recognition

1. Introduction

Automatic speech recognition (ASR) systems mostly don't generate punctuated sequences. This will degrade the readability of the outputs and result in poor user experiences [1]. Thus it is important to predict punctuation marks for speech transcripts. Many efforts have been made to restore punctuation marks automatically. There are about three classes of these approaches: prosody features [2, 3], lexical features [1, 4, 5, 6, 7, 8, 9, 10] and the combination of the prior two [11, 12, 13, 14, 15, 16] based methods. Since it is not difficult to get large scale text data, this paper only focuses on lexical features based approaches.

Many methods [1, 4, 5, 6, 7, 8, 9, 10] are proposed to predict punctuation marks only using text data. One kind of methods is that punctuation marks are viewed as hidden inter-word events [17, 18, 19]. An n-gram language model (LM) is used to predict punctuation marks. The other kind of methods is that punctuation restoration is treated as a sequence labeling task [1, 7], in which a punctuation mark is assigned to each word. Previous studies [20, 1, 21] show that conditional random fields (CRF) are better-suited to predict punctuation marks than the n-gram LM based methods. Recently, some studies [22] demonstrate that neural networks based models outperform the CRF based models over purely lexical features. Unlike the previous methods, the lexical features of the neural networks are word embeddings. Bidirectional recurrent neural network with attention mechanism (T-BRNN) is introduced by Tilk et al. [5] to improve the performance of the punctuation prediction task. Yi et al. [6] propose to distill knowledge from an ensemble of models for predicting punctuation marks. Most recently, deep recurrent neural networks with layer-wise multi-head attention-

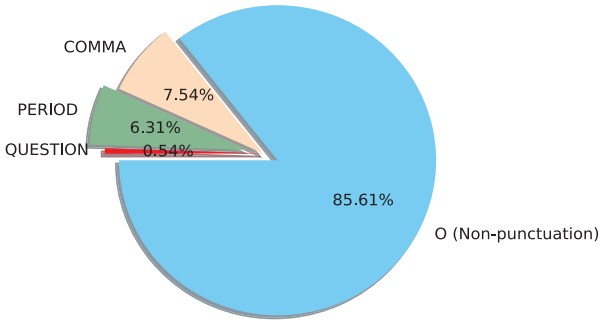

Figure 1: *Class distribution of the training set in IWSLT2011 datasets.*

s are used for punctuation restoration by Kim [8]. Some researchers [9, 10] use bidirectional encoder representations from transformers (BERT) models to initialize models for predicting punctuation marks.

The results demonstrate that the above-mentioned methods are effective and promising. However, datasets for predicting punctuation marks exist class imbalance problem, such as IWSLT2011 datasets. Figure 1 shows the class distribution of the training set in IWSLT2011 datasets. There are four classes in this datasets: *O*, *COMMA*, *PERIOD* and *QUESTION*. The proportion of non-punctuation mark *O* is about 85.61%, while the proportion of all punctuation marks (*COMMA*, *PERIOD* and *QUESTION*) is only about 14.39%. Furthermore, the examples of punctuation mark *QUESTION* are much less than that of other punctuation marks. This class imbalance problem will affect the performance of punctuation prediction tasks.

Motivated by the success of focal loss for dense object detection [23], this paper introduces a focal loss to alleviate the class imbalance issue for punctuation prediction. The focal loss focuses training on a sparse set of hard examples and down-weights the loss assigned to well-classified examples. In addition, inspired by the state-of-the-art performance of the pre-trained BERT model on many tasks [24, 10], this paper uses a pre-trained BERT model to initialize a punctuation prediction model as shown in Figure 2. The BERT model is trained by fusing context from both left and right directions.

The main contributions of this paper is that the focal loss is used to address class imbalance problem for punctuation prediction tasks. Experiments are conducted on IWSLT2011 datasets. The results show that the models trained with a focal loss acquire performance gains against the models trained with a cross entropy loss by up to 2.7% absolute overall F_1-score on test

http://dx.doi.org/10.21437/Interspeech.2020-1638

set. The proposed method also obtains better performance gains compared to previous state-of-the-art methods.

The rest of this paper is organized as follows. Section 2 reviews focal loss briefly. A pre-trained model trained with a focal loss is presented in Section 3. Experiments and results are described in Section 4. This paper is concluded in Section 5.

2. Review of focal loss

This section briefly introduces the definition of cross entropy loss and focal loss.

2.1. Cross entropy loss

The cross entropy (CE) loss is introduced for multi-class classification. The CE loss is defined as:

$$\mathcal{L}_{\mathrm{CE}} = -\sum_{k=1}^{K} t_k \log p_k \tag{1}$$

where k denotes the index of a class label, K is the total number of categories, p_k is the predicted probability of label k, t_k denotes the target probability of label k.

As shown in equation 2, where $t_k = 1$ if k belongs to the corresponding ground-truth class y_k, else it is 0.

$$t_k = \begin{cases} 1 & \text{if} \quad k = y_k \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

2.2. Focal loss

In order to address the class imbalance problem encountered during training, Lin et al. [23] propose to add a modulating factor $\beta = (1 - p_k)^{\gamma}$ to the standard CE loss $\mathcal{L}_{\mathrm{CE}}$, with a tunable focusing parameter $\gamma \geqslant 0$. So the focal loss (FL) is defined as:

$$\mathcal{L}_{\mathrm{FL}} = -\sum_{k=1}^{K} (1 - p_k)^{\gamma} t_k \log p_k \tag{3}$$

Thus the reshaped loss function $\mathcal{L}_{\mathrm{FL}}$ can down-weight easy examples and focus training on hard examples. When an example is misclassified and p_k is small, the modulating factor β is near 1 and the loss is unaffected. As p_k is near 1, β goes to 0 and the loss for well-classified examples is down-weighted.

The focusing parameter γ smoothly adjusts the rate at which easy examples are down-weighted. When $\gamma = 0$, FL is equivalent to CE loss. The effect of the modulating factor β is increased when γ is increased. Intuitively, β reduces the loss contribution from easy examples. Meanwhile, it extends the range in which an example receives low loss.

3. Focal loss for pre-trained model

Generally, datasets for predicting punctuation marks exist class imbalance problem. Most of the classes in the training set are non-punctuation marks. This will affect the performance of punctuation prediction tasks. Inspired by the success of focal loss for dense object detection [23], the focal loss is used to alleviate the class imbalance problem for the punctuation prediction task.

Additionally, motivated by the state-of-the-art performance of the pre-trained BERT model on many tasks [24, 9, 10], this paper tries to transfer parameters from a pre-trained BERT to a punctuation prediction model. Thus the model can learn bidirectional knowledge.

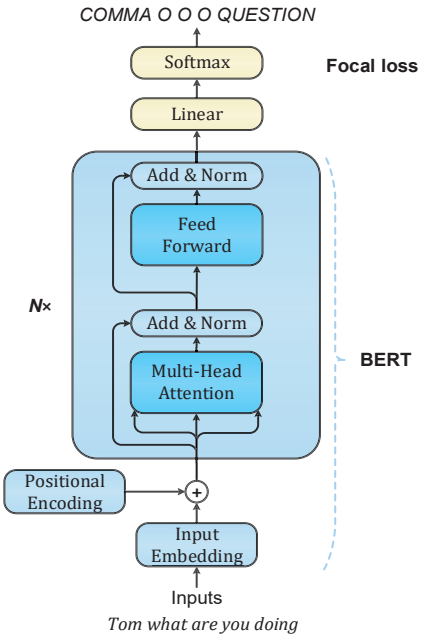

Figure 2: *Focal loss for punctuation prediction model with transferred BERT parameters.*

The architecture of the proposed model is shown as Figure 2. The inputs of the model are words, e. g. "*Tom what are you doing*", while the outputs of the model are punctuation marks, such as "*COMMA O O O QUESTION*". More details of punctuation marks are presented in Section 4.1. The model consists of task shared layers and task specific layers.

The task shared layers are transferred from the pre-trained BERT model [24], which has a stack of N identical layers as shown at the bottom of Fig. 2. Each layer has two sub-layers. The first is a multi-head self-attention mechanism. The second is a fully connected feed-forward network. A residual connection is employed around each of the two sub-layers, followed by layer normalization. More details are introduced in [25, 24].

The task specific layers are used for a punctuation predicting task. It is a simple classifier, which has a linear layer and a softmax layer. The output labels of the classification layer are punctuation marks and one non-punctuation mark.

The proposed model is trained with a focal loss. Intuitively, the modulating factor $\beta = (1 - p_k)^{\gamma}$ in the focal loss $\mathcal{L}_{\mathrm{FL}}$ reduces the loss contribution from easy examples and increases the importance of correcting misclassified examples.

4. Experiments

A series of experiments are conducted on English IWSLT datasets [22] to evaluate our proposed method.

4.1. Datasets

IWSLT datasets are from TED Talks, which are reorganized for predicting punctuation marks by Che et al. [22]. There are three kinds of datasets: training set, validation set and test set.

The training and validation sets are provided by the training

Table 1: *Overall data distributions of IWSLT datasets.*

Dataset	#Talks	#Sentences	#Tokens	*COMMA*	*PERIOD*	*QUESTION*	*O*
Training	1,690	143,991	2,102,417	158,499 (7.54%)	132,680 (6.31%)	11,311 (0.54%)	1,799,927 (85.61%)
Validation	20	20,635	295,800	22,475 (7.60%)	1,8940 (6.40%)	1,695 (0.57%)	252,690 (85.43%)
Test (*Ref.*)	8	861	12,626	830 (6.57%)	808 (6.40%)	53 (0.42%)	10,935 (86.61%)
Test (*ASR*)	8	852	12,822	798 (6.22%)	810 (6.32%)	42 (0.33%)	11,172 (87.13%)

Table 2: *The results of models with focal loss (FL) in terms of $P(\%)$,$R(\%)$, $F_1(\%)$ on test sets of IWSLT2011 datasets.*

Test set	Loss	γ	*COMMA*			*PERIOD*			*QUESTION*			*Total*		
			P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
Ref.	CE	0.0	71.9	74.2	73.0	85.6	85.8	85.7	71.2	85.3	77.6	76.2	81.8	78.9
	FL	0.1	72.1	74.5	73.3	85.7	86.0	85.9	71.6	85.6	78.0	76.5	82.0	79.1
	FL	0.2	72.2	74.7	73.4	85.9	86.2	86.1	71.9	86.1	78.4	76.7	82.3	79.4
	FL	0.5	72.7	74.9	73.8	86.3	86.7	86.5	72.4	86.6	78.9	77.1	82.7	79.8
	FL	1.0	73.1	75.6	74.3	86.6	87.0	86.8	72.7	87.0	79.2	77.5	83.2	80.2
	FL	2.0	**74.4**	**77.1**	**75.7**	**87.9**	**88.2**	**88.1**	**74.2**	**88.5**	**80.7**	**78.8**	**84.6**	**81.6**
	FL	5.0	73.1	75.9	74.5	86.8	87.5	87.2	73.1	87.1	79.5	77.7	83.5	80.5
ASR	CE	0.0	56.9	74.3	64.5	76.8	77.6	77.2	57.9	68.7	62.9	63.9	73.5	68.4
	FL	0.1	57.0	74.4	64.6	76.9	77.7	77.3	58.1	68.8	63.0	64.0	73.6	68.5
	FL	0.2	57.1	74.6	64.7	77.1	78.0	77.6	58.6	69.0	63.4	64.3	73.9	68.8
	FL	0.5	57.3	74.8	64.9	77.5	78.3	77.9	59.0	69.9	64.0	64.6	74.3	69.1
	FL	1.0	57.6	75.0	65.2	77.8	78.9	78.4	59.5	70.4	64.5	65.0	74.8	69.5
	FL	2.0	**59.0**	**76.6**	**66.7**	**78.7**	**79.9**	**79.3**	**60.5**	**71.5**	**65.6**	**66.1**	**76.0**	**70.7**
	FL	5.0	58.0	76.0	65.8	78.1	78.7	78.4	59.6	70.6	64.7	65.3	75.1	69.8

data of IWSLT2012 machine translation track, which consists of 1,710 TED Talks. Che et al. [22] further split it into training and validation sets according to the ID of TED talks. The two test sets are *Ref.* and *ASR*, which provided by the test data of IWSLT2011 ASR track. *Ref.* is from manual transcripts of audio files. *ASR* is from transcripts of the ASR system. More details of the datasets can be found in [22].

The datasets have four kinds of labels: *O*, *COMMA*, *PERIOD* and *QUESTION*. *O* denotes a non-punctuation mark. *COMMA* denotes the kind of colons or dashes. Exclamation marks or semicolons are denoted by *PERIOD*. *QUESTION* is the kind of question marks. Table 1 describes data statistics of IWSLT datasets. Table 1 shows that this datasets exist class imbalance problem. The proportion of non-punctuation mark *O* is much more than the proportion of all punctuation marks (*COMMA*, *PERIOD* and *QUESTION*) in all datasets. Moreover, the examples of punctuation mark *QUESTION* are much less than that of other punctuation marks.

4.2. Metrics

All models are evaluated in terms of precision (P), recall (R), F_1-score (F_1) in our experiments. We focus on the performance of the punctuation marks. So the correctly predicted non-punctuation marks *O* are ignored. We only evaluate the performance of *COMMA*, *PERIOD* and *QUESTION* on two test sets: *Ref.* and *ASR*, respectively. More details of metrics can be found in [22].

4.3. Experimental setup

The pre-trained BERT models are released by Google [1], implemented with the TensorFlow toolkit [26]. The pre-trained models include two kinds of models [2]: BERT-Large and BERT-Base. The size of our experimental dataset is small. Therefore, we use the Uncased BERT-Base model to initialize the models for predicting punctuation marks. Uncased means that any case and accent markers are stripped out.

The basic architecture of the BERT-Base model is shown at the bottom of Figure 2. The encoder has a stack of $N = 12$ identical layers. The heads of the parallel self-attention are 12. The total parameters of the BERT-Base is 110M. Please see [24] for pre-training details of the BERT-Base model.

The validation sets are utilized for selecting models and hyper parameters. The training terminates when a little improvement between two epochs on the validation set has been observed.

The results are reported on the two test sets of IWSLT datasets: *Ref.* and *ASR*. "*Total*" denotes the performance of all the three punctuation marks.

4.4. The baseline model trained with a cross entropy loss

The architecture of the baseline model for punctuation prediction is identical to that in Figure 2. The number of the output labels is four. The output labels of the classification layer are three punctuation marks and one non-punctuation mark *O*. The baseline model is trained with a cross entropy (CE) loss.

The model is first initialized with the parameters of the pre-trained BERT model. Then all of the parameters are jointly fine-tuned using the training data. The model is fine-tuned for 3 epochs over the training data. The $batch\_size$ is set to 32. The rate of dropout is set to 0.1. We use a linear learning rate decay schedule with warmup over 0.2% of training. we select the best fine-tuning learning rate of 5e-5 on the development set. The

[1] https://github.com/google-research/bert

[2] https://github.com/google-research/bert#pre-trained-models

Table 3: *Compared with other models on IWSLT2011 datasets. The results of punctuation predicting models in terms of $P(\%)$,$R(\%)$, $F_1(\%)$ on test sets .*

Test set	Model	*COMMA*			*PERIOD*			*QUESTION*			*Total*		
		P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
Ref.	T-BRNN-pre [5]	65.5	47.1	54.8	73.3	72.5	72.9	70.7	63.0	66.7	70.0	59.7	64.4
	BLSTM-CRF [6]	58.9	59.1	59.0	68.9	72.1	70.5	71.8	60.6	65.7	66.5	63.9	65.1
	Teacher-Ensemble [6]	66.2	59.9	62.9	75.1	73.7	74.4	72.3	63.8	67.8	71.2	65.8	68.4
	DRNN-LWMA-pre [8]	62.9	60.8	61.9	77.3	73.7	75.5	69.6	69.6	69.6	69.9	67.2	68.6
	Self-attention [16]	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-Punct_BASE [10]	72.1	72.4	72.3	82.6	83.5	83.1	77.4	89.1	82.8	77.4	81.7	79.4
	FL (Ours)	**74.4**	**77.1**	**75.7**	**87.9**	**88.2**	**88.1**	**74.2**	**88.5**	**80.7**	**78.8**	**84.6**	**81.6**
ASR	T-BRNN-pre [5]	59.6	42.9	49.9	70.7	72.0	71.4	60.7	48.6	54.0	66.0	57.3	61.4
	BLSTM-CRF [6]	55.7	56.8	56.2	68.7	71.5	70.1	63.8	53.4	58.1	62.7	60.6	61.5
	Teacher-Ensemble [6]	60.6	58.3	59.4	71.7	72.9	72.3	66.2	55.8	60.6	66.2	62.3	64.1
	DRNN-LWMA-pre [8]	-	-	-	-	-	-	-	-	-	-	-	-
	Self-attention [16]	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-Punct_BASE [10]	-	-	-	-	-	-	-	-	-	-	-	-
	FL (Ours)	**59.0**	**76.6**	**66.7**	**78.7**	**79.9**	**79.3**	**60.5**	**71.5**	**65.6**	**66.1**	**76.0**	**70.7**

results of the model trained with the CE loss on two test sets (*Ref.* and *ASR*) are reported in Table 2. When $\gamma = 0$, the FL is equivalent to the CE loss.

4.5. Our proposed models trained with a focal loss

The architecture of the proposed model for punctuation prediction is shown as Figure 2. The number of output labels of the classification layer is also four. The proposed models are trained with a focal loss (FL).

The models are also first initialized with the parameters of the pre-trained BERT model. Then all of the parameters are jointly fine-tuned using the training data. The models are fine-tuned for 3 epochs over the training data. We use a linear learning rate decay schedule with warmup over 0.2% of training. The $batch\_size$ is set to 32. The rate of dropout is set to 0.1. we select the best fine-tuning learning rate of 5e-5 on the development set.

The results of our models trained with the FL on two test sets are reported in Table 2. The FL introduces the focusing parameter γ that controls the strength of the modulating term $\beta = (1 - p_k)^\gamma$. The focusing parameters γ are set to be within [0, 5] in our experiments. When $\gamma = 0$, the FL is equivalent to the CE loss. The results in Table 2 show that the FL obtains performance improvements over the CE as γ is increased. With $\gamma = 2$, the FL yields the best gains on both two test sets. The models trained with the FL obtain performance improvement over that trained with the CE loss by 2.7% and 2.3% absolute overall F_1-score on test sets *Ref.* and *ASR*, respectively. The main reason of the performance improvement is that the focal loss focuses training on a sparse set of hard examples and down-weights the loss assigned to well-classified examples.

4.6. Comparison to other methods

We also compare our best model with $\gamma = 2$ to previous models on IWSLT2011 datasets. The previous results are listed in Table 3.

T-BRNN-pre is the best attention model proposed by Tilk et al. [5]. *BLSTM-CRF* denotes the best single model introduced in [6]. *Teacher-Ensemble* is the best ensemble model proposed by Yi et al. [6]. *DRNN-LWMA-pre* represents the best multi-head attention based model from [8]. *Self-attention* [27] is the transform based model trained with word and speech embeddings proposed by Yi et al. [6]. *Bert-Punct_BASE* denotes the model initialized by the BERT-Base model in [10].

T-BRNN-pre, *BLSTM-CRF*, *Teacher-Ensemble*, *DRNN-LWMA-pre* and *Bert-Punct_BASE* models in Table 3 are trained only with text data. Whereas *Self-attention* model is trained using both lexical and prosody features. Our models are trained only using text data.

The results in Table 3 show that our best model with $\gamma = 2$ using purely lexical features outperform all the previous state-of-the-art models. Our best model obtains better performance improvement over the model *T-BRNN-pre* [5] by 17.2% and 9.3% absolute overall F_1-score on Ref. and ASR test set, respectively. When compared with the model initialized by BERT-Base in [10], the overall F_1-score of our best model improves absolutely by 2.2% on *Ref.* test set. Our best model also outperforms the lexical and prosody model*Self-attention* [27] by 8.7% and 1.9% absolute overall F_1-score on Ref. and ASR test set, respectively.

5. Conclusions

This paper uses the focal loss to alleviate class imbalance problem encountered during training for the punctuation prediction task. The focal loss can focus training on a sparse set of hard examples and down-weight easy examples. Experiments are conducted on IWSLT2011 datasets. The results show that the models trained with the focal loss yields performance improvement over the models trained with the cross entropy loss by 2.7% and 2.3% absolute overall F_1-score on two test sets, respectively. Our best model also outperforms the previous state-of-the-art models. Future work includes applying the focal loss to other speech signal processing tasks.

6. Acknowledgements

This work is supported by the National Key Research & Development Plan of China (No.2018YFB1005003), the National Natural Science Foundation of China (NSFC) (No.61831022, No.61901473, No.61771472, No.61773379) and Inria-CAS Joint Research Project (No.173211KYSB20170061 and No.173211KYSB20190049).

7. References

[1] N. Ueffing, M. Bisani, and P. Vozila, "Improved models for automatic punctuation prediction for spoken and written text," in *INTERSPEECH*, 2013, pp. 3097–3101.

[2] H. Christensen, Y. Gotoh, and S. Renals, "Punctuation annotation using statistical prosody models," *Proc Isca Workshop on Prosody in Speech Recognition and Understanding*, pp. 35–40, 2001.

[3] J. Kim and P. C. Woodland, "A combined punctuation generation and speech recognition system and its performance enhancement using prosody," *Speech Communication*, vol. 41, no. 4, pp. 563–577, 2003.

[4] E. Cho, K. Kilgour, N. J., and W. A., "Combination of nn and crf models for joint detection of punctuation and disfluencies," in *INTERSPEECH*, 2015, pp. 3650–3654.

[5] O. Tilk and T. Alumae, "Bidirectional recurrent neural network with attention mechanism for punctuation restoration," in *INTERSPEECH*, 2016, pp. 3047–3051.

[6] J. Yi, J. Tao, Z. Wen, and Y. Li, "Distilling knowledge from an ensemble of models for punctuation prediction," in *INTERSPEECH*, 2017, pp. 2779–2783.

[7] P. elasko, P. Szymaski, J. Mizgajski, A. Szymczak, Y. Carmiel, and N. Dehak, "Punctuation prediction model for conversational speech," in *INTERSPEECH*, 2018, pp. 2633–2637.

[8] S. Kim, "Deep recurrent neural networks with layer-wise multi-head attentions for punctuation restoration," in *ICASSP*, 2019, pp. 7280–7284.

[9] A. Vravs and A. S., "Restoring punctuation and capitalization using transformer models," in *International Conference on Statistical Language and Speech Processing*, 2018.

[10] M. Karan, H. Thi-Nga, and C. Eng-Siong, "Transfer learning for punctuation prediction," in *APSIPA*, 2019, pp. 268–273.

[11] O. Tilk and T. Alumae, "Lstm for punctuation restoration in speech transcripts," in *INTERSPEECH*, 2015, pp. 683–687.

[12] X. Che, S. Luo, H. Yang, and C. Meinel, "Sentence boundary detection based on parallel lexical and acoustic models," in *INTERSPEECH*, 2016, pp. 2528–2532.

[13] O. Klejch, P. Bell, and S. Renals, "Sequence-to-sequence models for punctuated transcription combining lexical and acoustic features," in *ICASSP*, 2017, pp. 5700–5704.

[14] G. Szaszak and M. Tundik, "Leveraging a character, word and prosody triplet for an asr error robust and agglutination friendly punctuation approach," in *INTERSPEECH*, 2019, pp. 2988–2992.

[15] A. Nanchen and P. N. Garner, "Empirical evaluation and combination of punctuation prediction models applied to broadcast news," in *ICASSP*, 2019, pp. 7275–7279.

[16] J. Yi and J. Tao, "Self-attention based model for punctuation prediction using word and speech embeddings," in *ICASSP*, 2019, pp. 7270–7274.

[17] D. Beeferman, A. Berger, and J. Lafferty, "Cyberpunc: a lightweight punctuation annotation system for speech," in *ICASSP*, 1998, pp. 689–692 vol.2.

[18] E. Liu, Y. nd Shriberg, A. Stolcke, D. Hillard, M. Ostendorf, and M. Harper, "Enriching speech recognition with automatic detection of sentence boundaries and disfluencies," *IEEE Trans Audio Speech Language Process*, vol. 14, no. 5, pp. 1526–1540, 2006.

[19] A. Gravano, M. Jansche, and M. Bacchiani, "Restoring punctuation and capitalization in transcribed speech," in *ICASSP*, 2009, pp. 4741–4744.

[20] W. Lu and H. Ng, "Better punctuation prediction with dynamic conditional random fields." in *EMNLP*, 2010, pp. 177–186.

[21] M. Hasan, "Noise-matched training of crf based sentence end detection models," in *INTERSPEECH*, 2015, pp. 349–353.

[22] X. Che, C. Wang, H. Yang, and C. Meinel, "Punctuation prediction for unsegmented transcript based on word vector," in *LREC*, 2016, pp. 654–658.

[23] T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in *ICCV*, 2017, pp. 2980–2988.

[24] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *arXiv preprint arXiv:1603.00786*.

[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *NIPS*, 2017.

[26] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, and M. Isard, "Tensorflow: a system for large-scale machine learning," in *12th USENIX Symposium on Operating Systems Design and Implementation*, 2016, pp. 265–283.

[27] J. Yi, J. Tao, and Y. Bai, "Language-invariant bottleneck features from adversarial end-to-end acoustic models for low resource speech recognition," in *ICASSP*, 2019, pp. 6071–6075.