3R: Word and Phoneme Edition based Data Augmentation for Lexical Punctuation Prediction

Aihua Zheng[12], Naipeng Ye[13], Xiao Wang[1], Xiao Song[3*]
[1]*Anhui Provincial Key Laboratory of Multimodal Cognitive Computation,*
School of Computer Science and Technology, Anhui University, Hefei, China
[2]*Key Lab of Intelligent Computing and Signal Processing of Ministry of Education, Hefei, China*
[3]*Peking University Shenzhen Institute, Shenzhen, China*
ahzheng214@foxmail.com, naipengye@gmail.com, wangxiaocvpr@foxmail.com, xiao.song@imsl.org.cn

***Abstract*—Existing Lexical Punctuation Prediction methods are mainly trained on the standard clean data while losing the generalization in practical automatic speech recognition (ASR) system with ubiquitous transcription errors. To bridge the gap between clean training data and noisy testing data, we propose three random (3R) data augmentation strategies: random word deletion (RWD), random word substitution (RWS), and random phoneme edition (RPE) in both word and phoneme levels on the training dataset. Specifically, we contribute an acoustically similar vocabulary with phoneme level editions for acoustically similar word substitution. In addition, we first introduce the RoBERTa-large model into a punctuation prediction task to capture the semantics and the long-distance dependencies in language. Extensive experiments on the English dataset IWSLT2011 yield to a new state-of-the-art comparing to the prevalent punctuation prediction methods.**

***Keywords*-Automatic Speech Recognition; Punctuation Prediction; Data Augmentation;**

I. Introduction

Punctuation prediction, as a post-processing technique in automatic speech recognition (ASR), plays an important role in nature language processing related communities. Existing punctuation prediction methods fall into three categories: acoustic or prosodic features based methods [1], [2], lexical feature based methods [3], [4], and both [5], [6]. We focus on lexical feature based methods, which one can easily obtain the massive training data through the Internet such as Wikipedia, and thus avoid the intonation and pause diversities among the speakers.

Early works on lexical feature based punctuation prediction [7], [8] integrate punctuation as hidden events in N-gram language model or hidden markov model (HMM) [9] during ASR. More efforts utilize maximum entropy (EM) model [10] and conditional random fields (CRFs) [11] to model the punctuation prediction problem as a post-processing in ASR. With the blossom of deep learning, Tilk et al. [12] propose LSTM based punctuation prediction method, which can predict the punctuation of long sentences better. Recently, Kim et al. [4] propose a multi-directional deep recurrent neural network structure to better model the context from different perspectives. However, existing lexical punctuation prediction mainly train their models on the standard text data, which significantly limits their capability in real yet more challenging scenario with word errors. In fact, the transcription errors are ubiquitous in a practical ASR system. Therefore, due to the assumption of Independent Identical Distribution (i.e. IID) in deep learning based methods, the punctuation prediction models which learn their parameters on the standard clean training data, are unfortunately not competent on such noisy testing data with massive transcription errors.

On the one hand, data augmentation is an effective way to mitigate the distribution diversity between the training and testing data, avoid overfitting issue and improve the generalization of deep models, which has been successfully employed in all kinds of computer vision and nature language processing tasks [13], [14]. In this paper, we creatively propose three data augmentation strategies to bridge the gap between clean training data and noisy testing data. First, substitution errors (SE) and deletion errors (DE) are the two common types of word errors in ASR. To generate the hard samples with these two errors, we propose two word level data augmentation strategies, named Random Word Deletion (RWD) and Random Word Substitution (RWS) by randomly deleting and substituting some words respectively in the sentence with a certain proportion. Meanwhile, based on the observation that the most substitution errors in ASR derive from the wrong phonemes with acoustically similar pronunciation, we further propose a phoneme level data augmentation strategy named Random Phoneme Edition (RPE) by randomly substituting the certain words with their acoustically similar words constructed with similar phonemes.

On the other hand, existing methods mainly employ LSTM or RNN to model punctuation prediction as a sequence labeling task. Recently, BERT [15] has shown its superiority in various NLP tasks. More recently, Liu et al. [16] propose a robustly optimized BERT pre-training approach, RoBERTa, which can better capture the bidirectional context due to the dynamic masking to learn different language representations, and the multi-layer bidirectional

978-0-7381-4311-8/20/$31.00 ©2020 IEEE
DOI 10.1109/CIS52066.2020.00009

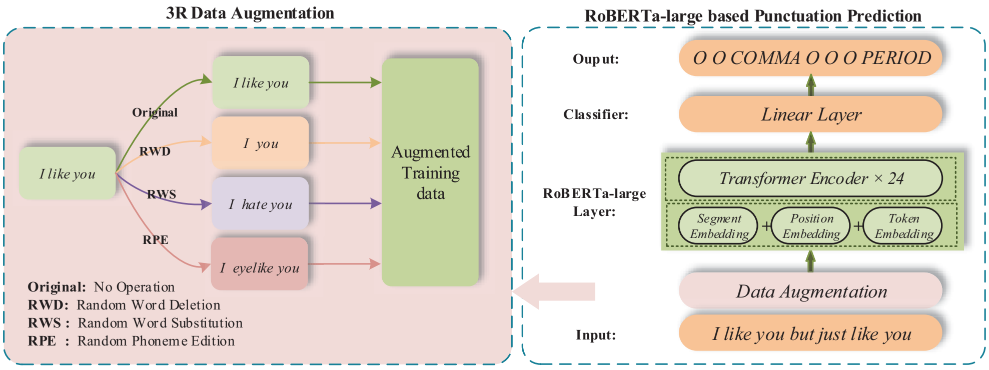

Figure 1. We first perform 3R operations, including two word edition RWD, RWS and one phoneme edition RPE for data augmentation. Then we feed the augmented data into the RoBERTa-large module for feature learning, followed by a linear layer for the final punctuation prediction.

transformer [17] to model longer-distance dependencies in a sequence. RoBERTa-large is a RoBERTa model with 24 layers of bidirectional transformer and it has been shown to be quite effective in [16]. Herein, we propose to employ the RoBERTa-large to model our punctuation prediction task.

In summary, we make following contributions in this paper:

- We propose three random (3R) data augmentation strategies in both word and phoneme levels for the punctuation prediction task. These three simple but rather effective strategies can significantly bridge the gap between training data and testing data in ASR.
- We contribute an acoustically similar vocabulary based on the phoneme level edition. This vocabulary can support the acoustically similar words substitution evaluation and other phoneme level operation in NLP related communities.
- We first propose to introduce RoBERTa-large into the punctuation prediction task to better capture the longer distance dependencies for sentences.
- Extensive experiments on benchmark punctuation prediction dataset validate the superiority of our method in both written and spoken languages.

II. The Proposed Methodology

In this paper, we propose three random (3R) data augmentation strategies for punctuation prediction based on RoBERTa-large. The pipeline of the proposed method is shown in Figure 1.

A. Punctuation Prediction based Sequence labeling

The most common way to solve the problem of punctuation prediction is to define it as a sequence labeling task, and to predict punctuation label sequence $Y = \{y_1 \cdots y_t \cdots y_n\}$ in a given word sequence $X = \{x_1 \cdots x_t \cdots x_n\}$, with the formula as follows:

$$y_t = \begin{cases} s \in S & \text{If } x_t \text{ is followed by one punctuation mark,} \\ O & \text{otherwise,} \end{cases} \quad (1)$$

where S is a closed set of punctuation marks including “PERIOD”, “COMMA” and “QUESTION”. The label “O” indicates the corresponding word followed by another word, while “PERIOD”, “COMMA” and “QUESTION” indicate the corresponding words followed by the specific punctuation marks.

B. 3R Data Augmentation Strategies

To bridge the gap between training data and testing data, we propose three kinds of random editions as the data augmentation strategies.

1) Random Word Deletion (RWD): According to our observation, the deleting error (DE) is one of most common errors in the transcribed text from ASR system. Based on this observation, we first propose to randomly delete some words to bridge the gap between *clean* training data and *noisy* testing data.

Table I
Two possible scenarios of Random Word Deletion (RWD).

Operation		Sentence					
Original (sequence)		*I*	*like*	*cooking*	*Mom*	*and*	*Dad*
Original (label)		O	O	COMMA	O	O	PERIOD
Scenario 1	sequence	*I*	*like*	*cooking*	~~*Mom*~~	*and*	*Dad*
	label	O	O	COMMA	~~O~~	O	PERIOD
Scenario 2	sequence	*I*	*like*	*cooking*	*Mom*	*and*	~~*Dad*~~
	label	O	O	COMMA	O	~~O~~	PERIOD

Specifically, we randomly delete some labels for the corresponding words in the sequence with a certain probability P_{RD} to generate the hard samples similar to the transcribed text sequence from ASR system. Table I shows two possible scenarios of random word deleting, where “Original” means no operation implemented on the sequence. There are two main types of word deleting: 1) Deleting the word followed by another word, as shown in Table I Scenario 1, where we directly delete the corresponding label. 2) Deleting the word followed by a specific punctuation, as shown in Table I Scenario 2. Instead of directly deleting the corresponding label which is obviously a punctuation label, we delete the forward label which refers to a certain word, therefore to remain the punctuation for the sentence.

2) Random Word Substitution (RWS): Another issue in ASR system is the substitution errors between the transcribed sequence and the real text sequence. To handle this issue, we actively substitute the randomly selected word in the sentence with another random word in the dictionary.

Table II
THE EXAMPLE OF RANDOM WORD SUBSTITUTION (RWS).

Operation	Sentence
Original (sequence)	*I like cooking mom and Dad*
Original (label)	O O COMMA O O PERIOD
RWS (sequence)	*I like ~~cooking~~ → play mom and Dad*
RWS (label)	O O COMMA O O PERIOD

Specifically, we randomly substitute certain probability P_{RS} word as shown in Table II.

3) Random Phoneme Edition (RPE): Despite of data augmentation via RWS at the word level, there exists another important form of substitution errors deriving from the phoneme misinterpretation. To handle this issue, we propose a phoneme level edition named random phoneme edition (RPE) in this paper for data augmentation.

Specifically, we first employ the grapheme to phoneme (G2P) tools[1] to convert the word to corresponding phoneme(s) lexicon, as illustrated in Table III, for 235886 words from a dictionary. Then we construct acoustically similar vocabularies [2] for each word via Edit Distance [18] between their phoneme strings. The top 3 corresponding words with closer Edit Distance to the phoneme string of each word are selected to construct the acoustically similar vocabularies for each word. As shown in Table IV, the candidate words generally sound acoustically similar to the word "*like*" especially reading in the context "*I like you*", which may indeed confuse both human and machine algorithms. Note that the acoustically similar vocabularies may be the none existing words.

To achieve the phoneme level substitution, we randomly choose a certain probability P_{RPE} words and substitute each word with one of its top 3 acoustically similar candidates.

C. RoBERTa-large based Punctuation Prediction

After obtaining the augmented training data via the proposed 3R edition, we propose to employ RoBERTa-large [16] to fulfill the punctuation prediction due to its powerful ability to capture semantics and long-distance dependencies in NLP. Given the input sentence in the text, we first obtain its feature embedding, i.e. token embedding, segment embedding and position embedding via the input representation layer, as shown in Fig. 1. Specifically, the token embedding is a general wordpiece used for representing a word. The segment embedding is to judge whether a sentence is continuous in BERT [15]. The position embedding [17] is to preserve the location information of each word while capturing the sequential information of the sentence. Then we add these three embeddings as the feature representation and feed into the subsequent RoBERTa-large module. RoBERTa-large model consists of 24 bidirectional transformer encoder modules. The transformer encoder module [17] adopts fully-connected self-attention and multi-head attention to model long-distance dependencies in a sequence. As shown in Fig. 1, after learning the contextualized word representations from RoBERTa-large model, we employ a linear layer for the final punctuation prediction and train by Cross Entropy (CE) loss.

Table III
THE EXAMPLES OF WORD-TO-PHONEMES CONVERTING.

Word	Phoneme String
I	"AY"
like	"L" "AY" "K"
you	"Y" "UW"

Table IV
THE TOP 3 CANDIDATE WORDS IN THE ACOUSTICALLY SIMILAR VOCABULARIES FOR THE WORD "LIKE".

Word	Phoneme String
alike	"AH" "L" "AY" "K"
liker	"L" "AY" "K" "ER"
eyelike	"AY" "L" "AY" "K"

We choose the pre-trained RoBERTA-large model from open source projects (e.g., HuggingFace's Transformers[3]). In the training stage, We set the length of the text sequence to 256. Adam is selected as the model optimizer, and the initial learning rate is 0.00001. The batch size and training epoch is set to 8 and 10 respectively.

III. EXPERIMENTAL EVALUATION

We evaluated our methods in both written and spoken text in English. Note that one can easily apply our method to other languages. The implementation platform is Python on Ubuntu 16.04 with the 10GB GTX 1080Ti GPU.

A. Datasets

We evaluate our method on the public challenging punctuation prediction dataset IWSLT [3], collected from TED talks. These datasets consist of three parts: training set, development set and testing set. The training set and the development sets contain about 144K sentences with 2.1M words and 21K sentences with 296K words respectively.

[1] https://github.com/cmusphinx/g2p-seq2seq
[2] https://github.com/learnxy/3R
[3] https://github.com/huggingface/transformers

Table V
EXPERIMENTAL RESULTS ON ASR TESTING SET (IN %). THE TOP THREE RESULTS ARE HIGHLIGHTED IN RED, BLUE AND GREEN RESPECTIVELY.

Algorithm	COMMA			PERIOD			QUESTION			OVERALL		
	P	R	$F1$	P	R	$F1$	P	R	$F1$	P	R	$F1$
T-LSTM [12]	41.8	37.8	39.7	56.4	49.3	52.6	55.6	42.9	48.4	49.1	43.6	46.2
T-BRNN-pre [19]	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 [20]	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 [20]	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 [4]	—	—	—	—	—	—	—	—	—	—	—	—
SAPP [6]	64.0	59.6	61.7	75.5	75.8	75.6	72.6	65.9	69.1	70.7	67.1	68.8
Baseline	56.10	72.06	63.08	79.51	83.44	81.42	47.17	71.43	56.82	60.92	75.64	67.11
Ours	59.91	70.05	64.59	80.17	83.44	81.77	61.36	77.14	68.35	67.15	76.88	71.57

Table VI
EXPERIMENTAL RESULTS ON $Ref.$ TEST SET (IN %). THE TOP THREE RESULTS ARE HIGHLIGHTED IN RED, BLUE AND GREEN RESPECTIVELY.

Algorithm	COMMA			PERIOD			QUESTION			OVERALL		
	P	R	$F1$	P	R	$F1$	P	R	$F1$	P	R	$F1$
T-LSTM [12]	49.6	41.4	45.1	60.2	53.4	56.6	57.1	43.5	49.4	55.0	47.2	50.8
T-BRNN-pre [19]	65.5	47.4	54.8	73.3	72.5	72.9	70.7	63.0	66.7	70.0	59.7	64.4
BLSTM-CRF [20]	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 [20]	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 [4]	62.9	60.8	61.9	77.3	73.7	75.5	69.6	69.6	69.6	69.9	67.2	68.6
SAPP [6]	67.4	61.1	64.1	82.5	77.4	79.9	80.1	70.2	74.8	76.7	69.6	72.9
Baseline	76.09	77.83	76.95	87.27	89.22	88.24	82.00	89.13	85.42	81.77	85.39	83.54
Ours	72.75	77.83	75.20	86.74	88.35	87.54	78.85	89.13	83.67	79.44	85.10	82.14

The testing set consists of two settings, the ASR testing set with spoken transcript set with 18% word error rate (WER), and the reference ($Ref.$) testing set with the ground truth is written transcripts of the ASR testing set. Either ASR or $Ref.$ testing set contains about 860 sentences with 13K words.

B. Evaluation Metrics

Following the protocols in [6], we evaluate the prediction results by the following three metrics, precision (P), recall (R) and F-measure ($F1$),

$$P = \frac{TP}{TP+FP}, R = \frac{TP}{TP+FN}, F1 = \frac{2PR}{P+R} \qquad (2)$$

where FP, TP, FN, and TN denote the number of false positives, true positives, false negatives, and true negatives, respectively.

C. Comparison Results

We compare our method to prevalent lexical punctuation prediction methods on both ASR and $Ref.$ testing sets. The edition probabilities P_{RWD}, P_{RWS} and P_{RPE} are empirically fixed to 5% in all the experiments.

1) Results on ASR testing set: Table V reports the comparison results on ASR testing set, where, "Baseline" indicates the method of RoBERTa-large on the original training set without any data augmentation. From Table V we can see, 1) our baseline outperforms most of the prevalent methods on overall $F1$ score, which verifies the effectiveness of RoBERTa for punctuation prediction. 2) SAPP [6] achieves superior performance on the "QUESTION" mark, which leads to a slightly higher overall $F1$ score than our baseline. The main reason is "QUESTION" marks are generally with distinguishing acoustic characteristics while SAPP [6] introduces the acoustic information into lexical features. Note that we only use the lexical features and still achieve comparable performance to SAPP [6] on overall $F1$ score and even much higher $F1$ scores on "COMMA" and "PERIOD". 3) By introducing the proposed 3R data augmentation, our method significantly beats all the existing methods on either lexical features or combination of the acoustic/prosodic and lexical features, which promises the contribution of 3R strategies.

2) Results on $Ref.$ testing set: To further demonstrate the robustness of our method on the standard transcription scenario, we further compare the results on $Ref.$ testing set, as shown in Table VI. First of all, our method (either with or without data augmentation) surpasses the state-of-the-art methods by a large margin, which verifies the effectiveness of our method. Although introducing data augment slightly declines the baseline by overall 1.4% score in $F1$, our method still significantly surpasses the state-of-the-art methods (by about 10% overall $F1$ score) and significantly boosts the performance in the more challenging scenario with transcription errors as shown in Table V.

D. Ablation Study

We implement the ablation study on our method with three variants on ASR testing set as reported in Table VII. Specifically, Ours-I, Ours-II and Ours-III indicate the variants by progressively introducing random word delete (RWD), random word substitute (RWS) and random phoneme edition (RPD) to the baseline. Clearly, all three augment strategies contribute to the task of punctuation prediction. The contribution can be descendingly ordered as RPE, RWD and RWS, while the combination of the three strategies achieves the best performance.

IV. CONCLUSION

In this paper, we have creatively designed three random data augmentation strategies (RWD, RWS, RPE) to bridge

Table VII
THE TOP THREE RESULTS ARE HIGHLIGHTED IN RED, BLUE AND GREEN RESPECTIVELY.

Component		Ours	Ours-I	Ours-II	Ours-III	Baseline
(a) RWD		✓	✓	×	×	×
(b) RWS		✓	×	✓	×	×
(c) RPE		✓	×	×	✓	×
Metrics						
OVERALL	P (%)	67.15	63.75	62.65	62.67	60.92
	R (%)	76.88	74.38	74.43	76.21	75.64
	$F1$ (%)	71.57	68.41	67.80	68.60	67.11

the gap between clean training data and noisy testing data for lexical punctuation prediction, and first introduce the RoBERTa-large into the prediction task. After data augmentation via the proposed three strategies, our method significantly improves the performance especially on the more challenging scenarios with massive substitution and deletion errors in ASR. While the powerful RoBERTa-large promises the performance of punctuation prediction on standard written language. Our future work will focus on real-time punctuation prediction with more efficient models.

ACKNOWLEDGMENT

This research is supported in part by the National Natural Science Foundation of China (61976002) and the Shenzhen Basic Research Program (JCYJ20170817155939233).

REFERENCES

[1] F. Batista, H. Moniz, I. Trancoso, and N. Mamede, "Bilingual experiments on automatic recovery of capitalization and punctuation of automatic speech transcripts," the IEEE Transactions on Audio Speech and Language Processing, vol. 20, no. 2, pp. 474–485, 2012.

[2] A. Moró and G. Szaszák, "A phonological phrase sequence modelling approach for resource efficient and robust real-time punctuation recovery," in Annual Conference of the International Speech Communication Association (INTERSPEECH), 2017.

[3] X. Che, C. Wang, H. Yang, and C. Meinel, "Punctuation prediction for unsegmented transcript based on word vector," in the Tenth International Conference on Language Resources and Evaluation (LREC), 2016, pp. 654–658.

[4] S. Kim, "Deep recurrent neural networks with layer-wise multi-head attentions for punctuation restoration," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019.

[5] G. Szaszák and M. kos Tündik, "Leveraging a character, word and prosody triplet for an asr error robust and agglutination friendly punctuation approach," in Annual Conference of the International Speech Communication Association (INTERSPEECH), 2019.

[6] J. Yi and J. Tao, "Self-attention based model for punctuation prediction using word and speech embeddings," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019.

[7] A. Gravano, M. Jansche, and M. Bacchiani, "Restoring punctuation and capitalization in transcribed speech," in IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2009.

[8] D. Beeferman, A. Berger, and J. Lafferty, "Cyberpunc: A lightweight punctuation annotation system for speech," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), vol. 2. IEEE, 1998, pp. 689–692.

[9] H. Christensen, Y. Gotoh, and S. Renals, "Punctuation annotation using statistical prosody models," in ISCA tutorial and research workshop on prosody in speech recognition and understanding (ITRW), 2001.

[10] J. Huang and G. Zweig, "Maximum entropy model for punctuation annotation from speech," in International Conference on Spoken Language Processing (ICSLP), 2002.

[11] M. Hasan, "Noise-matched training of crf based sentence end detection models," in Annual Conference of the International Speech Communication Association (INTERSPEECH), 2015.

[12] O. Tilk and T. Alume, "Lstm for punctuation restoration in speech transcripts," in Annual Conference of the International Speech Communication Association (INTERSPEECH), 2015.

[13] X. Wang, C. Li, B. Luo, and J. Tang, "Sint++: Robust visual tracking via adversarial positive instance generation," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4864–4873.

[14] J. Wei and K. Zou, "EDA: Easy data augmentation techniques for boosting performance on text classification tasks," in the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019, pp. 6383–6389.

[15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018.

[16] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," arXiv preprint arXiv:1907.11692, 2019.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in Advances in neural information processing systems (NIPS), 2017, pp. 5998–6008.

[18] G. Navarro, "A guided tour to approximate string matching," ACM computing surveys (CSUR), vol. 33, no. 1, pp. 31–88, 2001.

[19] O. Tilk and T. Alume, "Bidirectional recurrent neural network with attention mechanism for punctuation restoration," in Annual Conference of the International Speech Communication Association (INTERSPEECH), 2016.

[20] J. Yi, J. Tao, Z. Wen, and Y. Li, "Distilling knowledge from an ensemble of models for punctuation prediction," in Annual Conference of the International Speech Communication Association (INTERSPEECH), 2017.