A Small and Fast BERT for Chinese Medical Punctuation Restoration

Tongtao Ling[1,2], Yutao Lai[1], Lei Chen[1], Shilei Huang[3], Yi Liu[2,]*

[1]Guangdong University of Technology, Guangzhou, China
[2]IMSL Shenzhen Key Lab, PKU-HKUST Shenzhen Hong Kong Institute, Shenzhen, China
[3]Shenzhen Raisound Technologies, Co., Ltd, Shenzhen, China

`ltt_rick@163.com, 2112114007@mail2.gdut.edu.cn, chenlei3@gudt.edu.cn,`
`shilei.huang@raisound.com, yi.liu@imsl.org.cn`

Abstract

In clinical dictation, utterances after automatic speech recognition (ASR) without explicit punctuation marks may lead to the misunderstanding of dictated reports. To provide a precise and understandable clinical report with ASR, automatic punctuation restoration (APR) is required. Considering a practical scenario, we propose a fast and lightweight pre-trained model for Chinese medical punctuation restoration based on the 'pre-training and fine-tuning' paradigm. In this work, we distill pre-trained models by incorporating supervised contrastive learning and a novel auxiliary pre-training task (Punctuation Mark Prediction) to make it well-suited for punctuation restoration. We then reformulate APR as a slot tagging problem in the fine-tuning stage to bridge the gap between pre-training and fine-tuning. Our experiments on various distilled models reveal that our model can achieve 95% performance with a 10% model size relative to the state-of-the-art Chinese RoBERTa.

Index Terms: punctuation restoration, knowledge distillation, contrastive learning, slot tagging

1. Introduction

Automatic punctuation restoration is a crucial post-processing step of ASR systems. In the medical field, Electronic Medical Record systems (EMRs) leverage ASR technology to assist doctors in inputting medical reports rather than typing them manually, thereby enhancing the efficiency of medical professionals [1, 2]. However, the absence of punctuation marks in ASR-generated transcripts significantly impacts the comprehensiveness of medical reports. Consequently, it is non-trivial to predict and insert punctuation marks for ASR-generated transcripts or numerous text processing applications [3].

Over the past few years, the mainstream of punctuation restoration can be divided into three categories: 1) The first line [4, 5, 6] views punctuation restoration as a sequence labeling or token classification problem, which predicts whether each token is followed by a punctuation mark or not. 2) Second, some researchers [7, 8] treated it as a Seq2Seq task, which inputs a non-punctuation sequence and outputs a punctuated one. 3) The others [9, 10] propose a multi-modal framework to learn hybrid representations from audio and text samples for downstream classifiers to perform punctuation restoration.

Considering that text data is easier to access than audio data, BERT [11] and its variants [12, 13] have been widely applied to enhance the effectiveness of token classification, we focus on the first method. However, BERT was trained on large-scale punctuated dataset with punctuations but fine-tuned on non-punctuated dataset, which may critically damage the se-

mantics in downstream task. Besides, the data imbalance problem of token classification should be addressed. For instance, in the benchmark dataset IWSLT2011 [14], over 85% of tokens are no-punctuation tokens. Concretely, we aim to train a small and fast model that can be applied to the Chinese medical field, demonstrating high performance.

Inspired by these observations, in this paper, we propose a novel auxiliary pre-training task to alleviate the damage of semantics and incorporate supervised contrastive learning (SCL) [15] to solve the problem of data imbalance. To train a light model that has smaller hidden layer dimensions and transformer layers, we employ knowledge distillation (KD) [16] during the pre-training stage. In the fine-tuning stage, we re-formalize punctuation restoration as a Slot Tagging [17] task to narrow the gap between pre-training and fine-tuning. The main contributions of this paper:

- We train a fast and lightweight Chinese BERT model by incorporating contrastive learning and knowledge distillation into our auxiliary pre-training task. This model can be effectively fine-tuned for Chinese medical punctuation restoration.
- Our experiments reveal that compared with SOTA Chinese RoBERTa, our model achieves its 95% performance while only having its 10% model size.
- The ablation study reveals that our auxiliary task and supervised contrastive learning can enhance the performance of Chinese medical punctuation restoration.

2. Background

Punctuation Restoration. In the early years, punctuation restoration models primarily relied on lexical and prosodic features, but the acoustic feature is often noisy and error-prone [18]. Until the advent of transformer-based models, most of the punctuation restoration models are based on recurrent neural networks [19]. Recently, pre-trained language models can enhance the performance of downstream tasks (e.g., token classification), and punctuation restoration is often formulated as a token classification task [20].

Distillation. Knowledge distillation is a common method for compressing models, which aims to optimize a small student model to emulate a larger teacher model. As different layers of the transformer capture distinct information, the student model must acquire knowledge from both the final and intermediate layers of the teacher model [21]. Knowledge distillation has shown to be effective for natural language processing and speech processing [22, 23].

Contrastive Learning. Contrastive learning focuses on training models by highlighting the difference between pairs of data instances [24]. The core idea is to learn representations that make similar instances more alike while making dissimilar in-
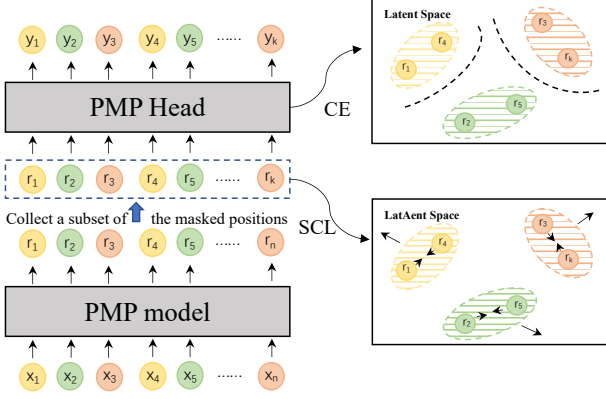
---

* corresponding author

Figure 1: *Overview of pre-training.*

stances more distinct [15].

3. Approach

3.1. Punctuation Mark Prediction

Masked Language Model (MLM) [11] is the main pre-training task in BERT and its variants, which learns contextual information that could be applied to downstream tasks (e.g., punctuation restoration). To boost the performance of punctuation prediction, we propose an auxiliary pre-training task called Punctuation Mark Prediction (PMP).

Specifically, as shown in Figure 1, given the input tokens $X = (x_1, ..., x_n)$, PMP model computes the contextual representations $H^{(L)} = (r_1, ..., r_n)$, $H^{(L)} \in \mathbb{R}^{N \times d}$ by an embedding layer and a consecutive L-layer transformer, where d is the dimension of hidden layer and N is the maximum number of tokens.

$$H^{(0)} = Embedding(X) \qquad (1)$$

$$H^{(i)} = Transformer(H^{(i-1)}), i \in \{1, ..., L\} \qquad (2)$$

PMP is similar to the MLM task, which aims to predict the token's type at the [MASK] position. After getting the contextual representations $H^{(L)}$, we collect a subset that contains each [MASK] token representation $H^{(m)} \in \mathbb{R}^{k \times d}$, where k is the number of [MASK] tokens. Following original MLM, k is set to $[N \times 15\%]$. Because the [MASK] token doesn't exist in downstream tasks, therefore 80% of the replaced tokens are [MASK], 10% are random tokens, and 10% are original tokens in the original MLM. Nevertheless, we utilize [MASK] tokens in punctuation restoration and the masking probability is set to 100%.

To conduct punctuation mark prediction, PMP-Head is introduced to compute the probability p over the punctuation mark set Y, such as $\{O, COMMA(,), PERIOD(.), QUESTION(?)\}$, O denotes no-punctuation token. PMP-Head contains a weight matrix $W \in \mathbb{R}^{|Y| \times d}$ and a bias term $B \in \mathbb{R}^{|Y|}$.

$$p = H^{(m)} W^{\top} + B \qquad (3)$$

Considering the data imbalance of punctuation classes, we incorporate a supervised contrastive learning (SCL) loss with the standard cross-entropy (CE) loss. The core of contrastive learning is to compare one class with other classes in a batch, which represents that the same classes are pulled together in the latent space while simultaneously pushing apart from different classes, as shown in Fig 1. Since SCL is a supervised method,
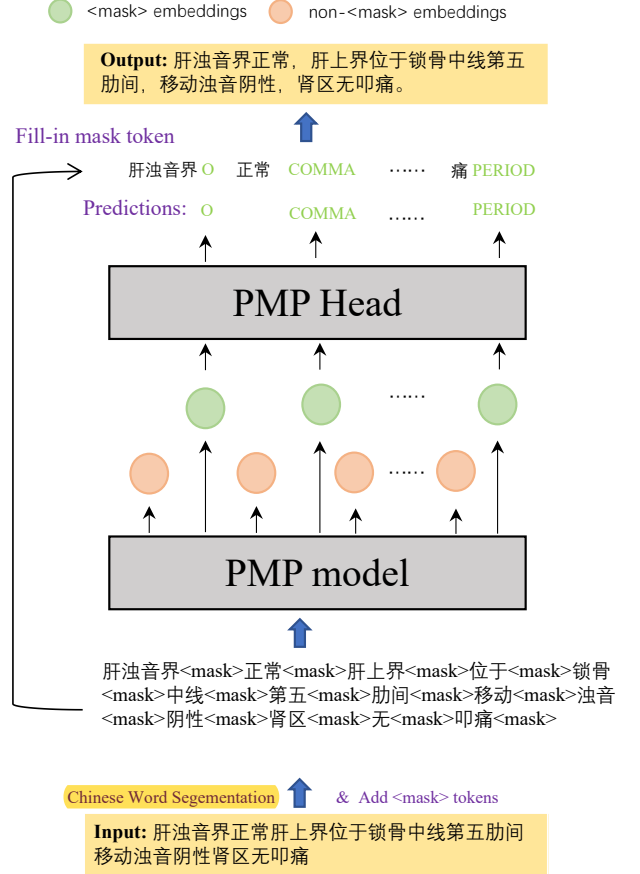

Figure 2: *Overview of fine-tuning.*

we pre-set a sample as anchor in each batch, the positive samples are defined as the ones in the sample class as the anchor, while the negative samples are at the opposite side. The CE and SCL loss are defined as:

$$\mathcal{L}_{CE} = -\frac{1}{M} \sum_{i=1}^{M} y_i \log p_i \qquad (4)$$

$$\mathcal{L}_{SCL} = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp\left(\Phi(r_i) \cdot \Phi(r_p)/\tau\right)}{\sum_{k \in A(i)} \exp\left(\Phi(r_i) \cdot \Phi(r_k)/\tau\right)} \qquad (5)$$

where i is the anchor sample, $P(i)$ represents the set of positive samples and $A(i)$ represents the set of all samples, regardless of the sample type. Φ is ℓ_2 normalization and τ is a scalar to stabilize the calculation. The contrastive loss is computed by similarity scores according to dot product. Finally, we add a hyper-parameter λ to combine two losses.

$$\mathcal{L}_{PMP} = (1 - \lambda)\mathcal{L}_{CE} + \lambda\mathcal{L}_{SCL} \qquad (6)$$

3.2. Knowledge Distillation

During the stage of pre-training, we distill the PMP model that follows conventional knowledge distillation method (Teacher-Student). First, we utilize Chinese RoBERTa-wwm to initialize PMP model as our teacher model, and select several transformer layers from the teacher model to initialize our student model.

Table 1: *Comparison of model structures.*

Model	Layers	Hidden Size	FFN Size	MAHs	#Params (Millions)
RoBERTa-wwm (teacher)	12	768	3072	12	102.3 (100%)
ALBERT-base-Chinese	12	768	3072	12	10.5 (10.3%)
RBT-6	6	768	3072	12	59.7 (58.4%)
MiniRBT	6	256	1024	8	10.3 (10.1%)
TinyBERT	4	312	1200	12	11.4 (11.2%)
PMP-H768	6	768	3072	12	59.7 (58.4%)
PMP-H256	6	256	1024	8	10.3 (10.1%)
PMP-H312	4	312	1200	12	11.4 (11.2%)

Table 2: *Data distribution of CMRPT.*

Dataset	Split	O (No punctuation)	Comma	Period	Colon
CMRPT	Train	4,332,870	434,981	83,248	33,304
	Dev	442,505	44,477	8,415	3,518
	Test	44,834	4,283	786	334

Then we distill the hidden layer from teacher model to student model through the Mean Square Error (MSE) function. The objective is defined as:

$$\mathcal{L}_{distill} = \frac{1}{M} \sum_{i=1}^{M} (H_s^{j'} W_h - H_t^j)^2 \tag{7}$$

where $H_s^{j'}$ is the j'-th layer of student model and H_t^j is the j-th layer of teacher model. The weight matrix $W_h \in \mathbb{R}^{d' \times d}$ matches the hidden layer of student model and the hidden layer of teacher model. d' and d represent the dimension of hidden layer of student model and teacher model respectively.

3.3. Slot Tagging for [MASK]

After distilling PMP model, vanilla MLM suffers the gap between pre-training stage and fine-tuning stage due to the absence of the artificial tokens (e.g., [MASK]) in the real downstream fine-tuning task.

To address this issue, we re-formalize punctuation restoration as Slot Tagging problem in fine-tuning stage and show an illustration in Figure 2. Specifically, given a non-punctuation input tokens $x = (x_1, x_2, x_3, x_4, ..., x_n)$ and punctuation tags $y = (y_1, y_2, y_3, y_4, ..., y_n)$, where n denotes the length of the input tokens. We utilize Jieba[1] as our Chinese word segmentation (CWS) tool and convert the output tokens to original Chinese words. Then, the special token [M] is inserted into each word.

$$x_t = (x_1 x_2, [M], x_3 x_4, [M], ..., x_n, [M]) \tag{8}$$

where [M] is the abbreviated form of [MASK], $x_1 x_2$ and $x_3 x_4$ are Chinese words that combined with Chinese characters x_1, x_2 and x_3, x_4, respectively.

Then, we utilize a regular classifier to fill the [M] tokens, which consists of a feed-forward layer that is initialized by PMP Head. This classifier need to output a sequence $(\hat{y_1}, \hat{y_2}, ..., \hat{y_K})$ in punctuation mark set Y, where K is the number of the [M] tokens, and project final outputs as $\hat{y} = (O, \hat{y_1}, O, \hat{y_2}, ..., \hat{y_K})$. The predicted punctuation marks of x_1 and x_3 are O, in order not to insert punctuation marks into Chinese whole words.

4. Corpus and Experimental Setup

4.1. Corpus

Aiming at training our PMP model on Chinese medical domain, we crawl up a huge amount of patient reports on medical website[2]. Then, we remove all html tags, duplicate data and patient privacy. We split our corpus into pre-training dataset and fine-tuning dataset, and the fine-tuning dataset is further divided into train, dev, test set. Pre-training dataset consists of 50 million tokens without labels. For fine-tuning dataset, we focused on three punctuation marks: {*COMMA*, *PERIOD*, *COLON*}. The reason why exclamation and question marks are not used is that these two punctuation marks appear less frequently in Chinese clinical report and are not needed in practical application scenarios. We remove punctuation marks in each sequence, marking the label according to whether there is punctuation mark after each token. The format of fine-tuning dataset is similar to Named Entity Recognition (NER) dataset, which contains many non-punctuation sequences with punctuation marks as labels. Our corpus is named as CMRPT (Chinese Medical Reports Punctuation Transcription). Detailed statistics are given in Table 2.

4.2. Experimental Setup

We employ RoBERTa-wwm [13] as the teacher model. To compare with various distilled version of BERT, including RBT-6[3] (re-trained 6-layer RoBERTa-wwm model), MiniRBT [12], TinyBERT [25], we present three students with similar hidden size and feed-forward size. The model structures are shown in Table 1. In the pre-training stage, we set the maximum sequence length to 512 with a batch size of 256, and learning rate is 4e-4. For supervised contrastive learning, λ and τ are 0.1 and 0.07 respectively. For knowledge distillation, the temperature is set to 8 and the number of training steps is 100K. In the fine-tuning stage, we use the AdamW [26] as optimizer with linear scheduler, a learning rate of 3e-5 and 0.1 warm-up step. We set weight-decay coefficient as 1e-5 and maximum gradient norms as 1.0. We train the model for 10 epochs on the training set, and the development set is used to save the best checkpoint by epoch. The experimental results are evaluated on the test set. We do pre-training tasks on 4 NVIDIA 80GB A100 GPUs and fine-tune on single NVIDIA RTX 4090 GPU. We run experiments in the fine-tuning stage 10 times with 10 different seeds and report average results. All experiments are evaluated using precision, recall and F1-score over the 4 punctuation marks.

---

[1] https://github.com/fxsjy/jieba

[2] Medical Reports: https://bingli.iiyi.com/. Chinese medicine books: https://www.zysj.com.cn/. Chinese medical encyclopedia: http://www.a-hospital.com/

[3] https://huggingface.co/hfl/rbt6

Table 3: *Results in terms of precision (P %), recall (R %), and F1-score (F1 %) on CMRPT.*

Model	Comma			Period			Colon			Overall		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
RoBERTa-wwm	94.35	95.71	95.03	89.88	92.15	91.00	88.38	83.77	86.00	90.87	90.54	90.68
RBT-6	93.45	86.58	89.88	88.94	84.41	86.62	82.57	81.44	81.99	88.32	84.14	86.16
MiniRBT	88.78	88.54	88.66	86.03	79.23	82.49	87.58	76.04	81.41	87.46	81.27	84.18
TinyBERT	88.79	89.01	88.9	86.77	79.45	82.95	86.52	74.42	79.99	87.36	80.96	83.95
PMP-H768	93.90	95.41	94.65	91.67	90.82	91.24	87.03	80.15	83.43	90.87	88.79	89.77
PMP-H256	93.28	87.07	90.07	88.68	85.06	86.83	84.16	77.04	80.32	88.71	83.06	85.74
PMP-H312	93.35	86.52	89.81	89.52	85.11	87.26	82.81	77.94	80.29	88.56	83.19	85.78

Table 4: *Comparison of PMP with ALBERT.*

Model	P	R	F1
ALBERT	89.44	85.87	87.54
PMP-H768	90.87	88.79	89.77

Table 5: *Ablation Study. The model is PMP-H256 pre-trained with 30K steps*

Model	P	R	F1
PMP-H256	89.01	84.56	86.72
w/o CE	88.09	80.79	84.15
w/o SCL	88.09	83.19	85.56
w/o KD	87.91	89.05	88.44

Our code and data are available at `https://github.com/rickltt/punctuation_restoration`.

5. Experiment

5.1. Main Results

Table 3 presents the results of fine-tuning different distilled models for punctuation restoration. Compared with RoBERTa-wwm, PMP-256 and PMP-312 achieve its 95% performance with almost 10% of the model size. With the same number of parameters (PMP-H256 v.s. MiniRBT, PMP-H312 v.s. Tiny-BERT, PMP-H768 v.s. RBT-6), PMP models achieve better performance, demonstrating that the overall feasibility and effectiveness of our pre-training stage. Then, PMP-H312 and PMP-H256 achieve similar performance of RBT-6 even though it has 5-6 times (10.3M v.s. 59.7M, 11.4 v.s. 59.7M) more parameters, proving that narrow and deep models yield higher performance than wide and shallow models. These results suggest that the pre-training stage effectively reduces the gap in size between teacher (RoBERTa-wwm) and student models (PMP), thus allowing PMP models to maintain excellent performance even with small sizes. Overall, our PMP models achieve the best performance among all the other distilled models, and remain almost 95% performance of the teacher model.

To compare with the current state-of-the-art light version of BERT, Table 4 shows the results of ALBERT [27] and PMP-H768. Our model achieves improvements of 4.3% precision, 4.3% recall and 4.3% F1. The superiority of PMP with the same number of parameters of ALBERT proves that our proposed model is effective. Our experiments demonstrate that PMP has notable performance advantages over other models with the same number of parameters.

The competitive performance of our proposed method can be attributed to several key factors. Firstly, we distill pre-trained models, leveraging the power of transfer learning to capture general linguistic knowledge. Secondly, we incorporate supervised contrastive learning, which enhances the model's ability to discern fine-grained differences between inputs, crucial for accurate punctuation restoration. Finally, we introduce a novel auxiliary pre-training task: Punctuation Mark Prediction. This task specifically targets punctuation restoration, allowing the model to gain a deeper understanding of punctuation patterns and their contextual usage. The combination of these three elements results in a model that is well-suited for punctuation restoration, explaining the superior performance of our method.

5.2. Ablation Study

To explore which part plays a key role, we perform the ablation study in Table 5. First, the declining result of eliminating CE verifies that punctuation prediction is effective to pre-training. Second, we remove SCL in the pre-training and the precision, recall and F1-score decrease 0.9, 1.3 and 1.1 points, respectively. Then, we directly remove knowledge distillation but keep CE and SCL. The performance of F1 score have a little boost from 86.72% to 88.44%, but with large model size and longer inference time. The metrics on F1-score decreases 1.16-2.57 points when we drop out SCL and CE, respectively. The studies confirm that the design of PMP can boost the effectiveness of punctuation restoration.

6. Conclusion

This paper focuses on Chinese medical punctuation restoration and leverages supervised contrastive learning and a novel auxiliary pre-training task to enhance this task. To train a fast and light model, we employ knowledge distillation in pre-training stage. Extensive experiments and analyses show that PMP has promising performance on punctuation restoration.

7. Acknowledgements

8. References

[1] T. Hodgson and E. Coiera, "Risks and benefits of speech recognition for clinical documentation: a systematic review," *Journal of*

the american medical informatics association, vol. 23, no. e1, pp. e169–e179, 2016.

[2] W. Salloum, G. Finley, E. Edwards, M. Miller, and D. Suendermann-Oeft, "Deep learning for punctuation restoration in medical reports," in *BioNLP 2017*, 2017, pp. 159–164.

[3] X.-Y. Fu, C. Chen, M. T. R. Laskar, S. Bhushan, and S. Corston-Oliver, "Improving punctuation restoration for speech transcripts via external data," in *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, 2021, pp. 168–174.

[4] B. Lin and L. Wang, "Joint prediction of punctuation and disfluency in speech transcripts." in *Interspeech*, 2020, pp. 716–720.

[5] Q. Huang, T. Ko, H. L. Tang, X. Liu, and B. Wu, "Token-level supervised contrastive learning for punctuation restoration," in *Annual Conference of the International Speech Communication Association*. International Speech Communication Association, 2021.

[6] N. Shi, W. Wang, B. Wang, J. Li, X. Liu, and Z. Lin, "Incorporating external pos tagger for punctuation restoration," *Proc. Interspeech 2021*, pp. 1987–1991, 2021.

[7] O. Klejch, P. Bell, and S. Renals, "Sequence-to-sequence models for punctuated transcription combining lexical and acoustic features," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5700–5704.

[8] F. Wang, W. Chen, Z. Yang, and B. Xu, "Self-attention based network for punctuation restoration," in *2018 24th International Conference on Pattern Recognition (ICPR)*, 2018, pp. 2803–2808.

[9] Y. Zhu, L. Wu, S. Cheng, and M. Wang, "Unified multimodal punctuation restoration framework for mixed-modality corpus," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7272–7276.

[10] M. Sunkara, S. Ronanki, D. Bekal, S. Bodapati, and K. Kirchhoff, "Multimodal semi-supervised learning framework for punctuation prediction in conversational speech," *Proc. Interspeech 2020*, pp. 4911–4915, 2020.

[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *NAACL-HLT*, 2019, pp. 4171–4186.

[12] X. Yao, Z. Yang, Y. Cui, and S. Wang, "Minirbt: A two-stage distilled small chinese pre-trained model," *arXiv preprint arXiv:2304.00717*, 2023.

[13] Y. Cui, W. Che, T. Liu, B. Qin, and Z. Yang, "Pre-training with whole word masking for chinese bert," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3504–3514, 2021.

[14] M. Federico, M. Cettolo, L. Bentivogli, P. Michael, and S. Sebastian, "Overview of the iwslt 2012 evaluation campaign," in *Proceedings of the international workshop on spoken language translation (IWSLT)*, 2012, pp. 12–33.

[15] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," *Advances in neural information processing systems*, vol. 33, pp. 18 661–18 673, 2020.

[16] Z. Yang, Y. Cui, Z. Chen, W. Che, T. Liu, S. Wang, and G. Hu, "Textbrewer: An open-source knowledge distillation toolkit for natural language processing," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 2020, pp. 9–16.

[17] Y.-B. Kim, M. Jeong, K. Stratos, and R. Sarikaya, "Weakly supervised slot tagging with partially labeled sequences from web search click logs," in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2015, pp. 84–92.

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

[19] K. Xu, L. Xie, and K. Yao, "Investigating lstm for punctuation prediction," in *2016 10th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. IEEE, 2016, pp. 1–5.

[20] V. Lai, A. Salinas, H. Tan, T. Bui, Q. Tran, S. Yoon, H. Deilamsalehy, F. Dernoncourt, and T. Nguyen, "Boosting punctuation restoration with data generation and reinforcement learning," in *Proceedings of INTERSPEECH*. International Speech Communication Association, 2023.

[21] T. Ashihara, T. Moriya, K. Matsuura, and T. Tanaka, "Deep versus Wide: An Analysis of Student Architectures for Task-Agnostic Knowledge Distillation of Self-Supervised Speech Models," in *Proc. Interspeech 2022*, 2022, pp. 411–415.

[22] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.

[23] H.-J. Chang, S.-w. Yang, and H.-y. Lee, "Distilhubert: Speech representation learning by layer-wise distillation of hidden-unit bert," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7087–7091.

[24] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[25] X. Jiao, Y. Yin, L. Shang, X. Jiang, X. Chen, L. Li, F. Wang, and Q. Liu, "Tinybert: Distilling bert for natural language understanding," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020, pp. 4163–4174.

[26] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2018.

[27] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "Albert: A lite bert for self-supervised learning of language representations," in *International Conference on Learning Representations*, 2019.