Self-Supervised Curriculum Learning for Spelling Error Correction

Zifa Gan[1] **Hongfei Xu**[2] **Hongying Zan**[1*]
[1]Zhengzhou University, Henan, China
[2]DFKI and Saarland University, Informatics Campus, Saarland, Germany
{zfganlp, hfxunlp}@foxmail.com, iehyzan@zzu.edu.cn

Abstract

Spelling Error Correction (SEC) that requires high-level language understanding is a challenging but useful task. Current SEC approaches normally leverage a pre-training then fine-tuning procedure that treats data equally. By contrast, Curriculum Learning (CL) utilizes training data differently during training and has shown its effectiveness in improving both performance and training efficiency in many other NLP tasks. In NMT, a model's performance has been shown sensitive to the difficulty of training examples, and CL has been shown effective to address this. In SEC, the data from different language learners are naturally distributed at different difficulty levels (some errors made by beginners are obvious to correct while some made by fluent speakers are hard), and we expect that designing a curriculum correspondingly for model learning may also help its training and bring about better performance. In this paper, we study how to further improve the performance of the state-of-the-art SEC method with CL, and propose a Self-Supervised Curriculum Learning (SSCL) approach. Specifically, we directly use the cross-entropy loss as criteria for: 1) scoring the difficulty of training data, and 2) evaluating the competence of the model. In our approach, CL improves the model training, which in return improves the CL measurement. In our experiments on the SIGHAN 2015 Chinese spelling check task, we show that SSCL is superior to previous norm-based and uncertainty-aware approaches, and establish a new state of the art (74.38% F1).

1 Introduction

Spelling Error Correction (SEC) aims to automatically correct the spelling errors in written text either at word-level or character-level (Yu and Li, 2014; Yu et al., 2014; Zhang et al., 2015; Wang et al., 2018; Hong et al., 2019; Wang et al., 2019a). Although being a very valuable natural language application, SEC is a challenging task and needs high-level language understanding.

Curriculum Learning (CL) (Bengio et al., 2009) facilitates model training in an easy-to-hard order. Previous studies (Kocmi and Bojar, 2017; Platanios et al., 2019; Zhang et al., 2019; Zhou et al., 2020) use sentence length or word rarity for CL, but merely consider features over sentences, which is not capable to fully reflect the data challenge for a model. SEC data difficulty is influenced by many factors, such as sentence length, word rarity and a great diversity of errors. In addition, previous CL approaches require careful design for data difficulty and training curricula. Ruiter et al. (2020) show that self-supervised learning is a curriculum learner, which might be useful to avoid such efforts. In this paper, we propose a novel Self-Supervised CL (SSCL) approach to evaluating data difficulty from the model's perspective and automatically arranging curricula for the model. Specifically, we use the training loss as the measurement of data difficulty (i.e., data of higher loss are harder to learn), and evaluate the model competence based on the loss reduction during training (i.e., a model checkpoint of lower loss is of higher performance). We expect CL to improve the model training, which in return improves the CL measurements in a virtuous circle.

Our main contributions are as follows:

- We propose a novel SSCL approach which avoids human design of CL measurements to improve the SOTA SEC model;
- We empirically show that our SSCL approach is better than the previous norm-based and uncertainty-aware CL approaches, and establish a new SOTA (74.38% F1) on the SIGHAN 2015 spelling error check task.

* Corresponding author.

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 3487–3494
November 7–11, 2021.

Algorithm 1 Self-Supervised Curriculum Learning Strategy.
Input: Training set $D = \{\langle x^n, y^n \rangle\}_{n=1}^{N}$.
Output: Spelling error correction model θ.

Train the initial SEC model θ on the synthetic training set for one epoch.
Compute data difficulty $\left\{\hat{d}\left(\langle x^n, y^n \rangle\right)\right\}_{n=1}^{N}$ using the pre-trained system θ, Eq. 1 and Eq. 2.
while θ is not converged **do**
  Compute model competence $\check{c}\left(t\right)$ using Eq. 7.
  Generate training subset $D_t = \left\{\langle x^n, y^n \rangle \;\middle|\; \hat{d}\left(\langle x^n, y^n \rangle\right) < \check{c}\left(t\right),\ \langle x^n, y^n \rangle \in D\right\}$.
  Compute instance-level data weight $W_d = \{w_d\left(\langle x^n, y^n \rangle,\ t\right) | \langle x^n, y^n \rangle \in D_t\}$.
  Compute token-level data weight $W_t = \{w_t\left(\langle x_i^n, y_i^n \rangle,\ t\right) | \langle x_i^n, y_i^n \rangle \in \langle x^n, y^n \rangle, \langle x^n, y^n \rangle \in D_t\}$.
  Update θ with the loss of examples $E_{\langle x^n, y^n \rangle \sim D_t}$ calculated by W_d, W_t and Eq. 6.
end while
return θ

2 Self-Supervised Curriculum Learning

Curriculum learning requires to evaluate data difficulty and model competence during training, so as to selectively feed data of similar competence as the model's ability to the model. The algorithm is shown in Algorithm 1. We use the SEC model trained on the 5M synthetic data for one epoch to compute the data difficulty. For every epoch, we first compute model competence, and then select instances whose data difficulties are no more than the model competence to train model. In every training step, we compute data weights for back-propagation.

2.1 Data Difficulty

We use the training loss of each data instance as the measurement of data difficulty. Intuitively, the data with a lower loss are easier for the model. For a dataset with N instances $\langle \mathbf{X}, \mathbf{Y} \rangle = \{\langle x^n, y^n \rangle\}_{n=1}^{N}$, where x^n and y^n are the input and the reference respectively, SSCL measures the data difficulty by the training loss.

$$d\left(\langle x^n, y^n \rangle\right) = -\log \mathrm{P}\left(y^n | x^n\right) \tag{1}$$

We use the Cumulative Density Function (CDF) to transfer the distribution of data difficulty into $(0, 1]$, following Liu et al. (2020):

$$\hat{d}\left(\langle x^n, y^n \rangle\right) \in (0, 1] = \mathrm{CDF}\left(\{d\left(\langle x^n, y^n \rangle\right)\}_{n=1}^{N}\right)^n \tag{2}$$

The score of more difficult data tends to be 1, while that of easier data tends to be 0.

Rather than using the random initialized model directly for the data difficulty evaluation, the SEC model is first pre-trained for one epoch on the full synthetic training set to ensure evaluation quality of the start point.

Compared to previous approaches, SSCL has the following advantages:

- It does not require manually designed data difficulty evaluation metrics;
- The evaluation quality of data difficulty can be improved together with the training of the model.

2.2 Data Weight

In the training process of competence-based CL (Platanios et al., 2019), the model treats all the selected data equally, which may overuse the easy data with low difficulty. It is however counter-intuitive and wastes computational resources (Liu et al., 2020). To address this issue, we additionally introduce a weight to the loss function at instance-level or token-level or both levels.

Following Liu et al. (2020), the instance-level weight is defined as:

$$w_d\left(\langle x^n, y^n \rangle,\ t\right) = \left(\frac{\hat{d}\left(\langle x^n, y^n \rangle\right)}{\check{c}\left(t\right)}\right)^{\lambda_w} \tag{3}$$

where λ_w is the scaling hyperparameter smoothing the data weight, $\hat{d}\left(\langle x^n, y^n \rangle\right)$ is the loss-based data difficulty, and $\check{c}\left(t\right)$ is the model competence (described in Section 2.3).

For training step t and the corresponding model competence $\check{c}\left(t\right)$, the weighted training loss of the instance $w_d\left(\langle x^n, y^n \rangle,\ t\right)$ is:

$$\hat{l}\left(\left\langle x^n, y^n\right\rangle, t\right) = -\log \mathrm{P}\left(y^n|x^n\right) w_d\left(\left\langle x^n, y^n\right\rangle, t\right) \tag{4}$$

where $w_d\left(\left\langle x^n, y^n\right\rangle, t\right)$ encourages the training to pay more attention to more difficult data with higher data weights than to easier data.

Inspired by the token-level confidence (Wan et al., 2020), we also weigh different tokens of a data instance differently, and present the token-level weight based on the squared token-level cross-entropy loss normalized at the sentence-level:

$$w_t\left(\left\langle x_i^n, y_i^n\right\rangle, t\right) = 1 + \frac{l\left(\left\langle x_i^n, y_i^n\right\rangle, t\right)^2}{\sum_{j=1}^{I} l\left(\left\langle x_j^n, y_j^n\right\rangle, t\right)^2} \tag{5}$$

where $l\left(\left\langle x_i^n, y_i^n\right\rangle, t\right)$ stands for the cross-entropy loss of the ith token of the example $\left\langle x^n, y^n\right\rangle$ of the tth training step. We ensure all weights to be larger than 1 to ensure the gradient norm during backpropagation (Gu et al., 2020).

The token-level weight unties tokens from training instances and encourages the model to pay more attention to the difficult tokens in the sentence.

We consider the combination of both instance-level and token-level as:

$$\begin{aligned} &\check{l}\left(\left\langle x^n, y^n\right\rangle, t\right) = w_d\left(\left\langle x^n, y^n\right\rangle, t\right) * \\ &\sum_{i=1}^{I} -logP\left(y_i^n|x_i^n\right) w_t\left(\left\langle x_i^n, y_i^n\right\rangle, t\right) \end{aligned} \tag{6}$$

2.3 Model Competence

To evaluate the model competence during training, Platanios et al. (2019) use the training step to determine the model competence. Liu et al. (2020) utilize the norm of the model's source embedding to compute the model competence. Based on the design of Liu et al. (2020), but using the loss reduction during training instead of the embedding norm, we define the model competence as:

$$\check{c}(t) = \min\left(1, \sqrt{l_t \frac{1 - c_0^2}{\lambda_s l_0} + c_0^2}\right) \tag{7}$$

where $c_0 = 0.01$, l_t denotes the loss reduction in the training, l_0 is the total initial loss, and λ_s is a task-independent hyperparameter to control the length of the curriculum.

With l_t increasing from low to high, the model's training gradually includes increasingly more difficult training data.

3 Experiments

3.1 Settings

We apply CL approaches to the SOTA Soft-Masked BERT model (Zhang et al., 2020) to test their effectiveness.

Soft-Masked BERT (Zhang et al., 2020) is a model architecture for SEC. It employs a Bi-GRU as the detection network and the pre-trained BERT (Devlin et al., 2019) as the correction network. The detection network predicts the probabilities of errors and the correction network predicts the probabilities of error corrections, while the former passes its prediction results to the latter.

Experiments were conducted on the SIGHAN 2015 Chinese spelling check task, we followed Zhang et al. (2020) for experiment settings. Models were first pre-trained on 5M synthetic data, and then fine-tuned on the SIGHAN data. Parameters were initialized under the Lipschitz constraint (Xu et al., 2020).

We also compared our SSCL approach with the Norm-Based CL (NBCL) (Liu et al., 2020) and the Uncertainty-Aware CL (UACL) approaches (Zhou et al., 2020). NBCL uses the norm of word embeddings to measure the difficulty of the sentence, the competence of the model and the weight of the sentence. UACL utilizes the average cross-entropy of words in an example as its data difficulty, and exploits the variance of distributions over the Monte Carlo Dropout (Gal and Ghahramani, 2016) results of the model's output probabilities to present the model uncertainty.

Performance of different approaches was evaluated by the sentence-level accuracy, precision, recall, and F1 score.

3.2 Main Results

The results of our approach and baselines are shown in Table 1.

Table 1 shows that: 1) CL methods are able to significantly further improve the performance of the SOTA Soft-Masked BERT model (66.4% F1). Specifically, NBCL and UACL are able to further improve the Soft-Masked BERT model by +6.81% and +7.33% F1 respectively; 2) our SSCL

Method	Detection				Correction			
	Acc.	Prec.	Rec.	F1.	Acc.	Prec.	Rec.	F1.
NTOU (2015)	42.2	42.2	41.8	42.0	39.0	38.1	35.2	36.6
NCTU-NTUT (2015)	60.1	71.7	33.6	45.7	56.4	66.3	26.1	37.5
HanSpeller++ (2015)	70.1	80.3	53.3	64.0	69.2	79.7	51.5	62.5
Hybird (2018)	-	56.6	69.4	62.3	-	-	-	57.1
FASPell (2019)	74.2	67.6	60.0	63.5	73.7	66.6	59.1	62.6
Confusionset (2019a)	-	66.8	73.1	69.8	-	71.5	59.5	64.9
BERT-Pretrain (2020)	6.8	3.6	7.0	4.7	5.2	2.0	3.8	2.6
BERT-Finetune (2020)	80.0	73.0	70.8	71.9	76.6	65.9	64.0	64.9
Soft-Masked BERT (2020)	**80.9**	73.3	**73.2**	73.5	77.4	66.7	**66.2**	66.4
Soft-Masked BERT$_{NBCL}$	80.27	**86.49**	70.98	77.97	76.91	**85.26**	64.14	73.21
Soft-Masked BERT$_{UACL}$	80.09	85.31	71.90	78.03	77.00	84.12	65.62	73.73
Soft-Masked BERT$_{SSCL}$	80.82	86.34	72.46	**78.79**	**77.64**	85.20	65.99	**74.38**

Table 1: Performances of different methods on the SIGHAN 2015 Chinese spelling check task.

Weight	Detection				Correction			
	Acc.	Prec.	Rec.	F1.	Acc.	Prec.	Rec.	F1.
w_d	80.00	85.27	71.72	77.91	76.64	83.97	64.88	73.20
w_t	79.55	85.11	70.79	77.30	76.09	83.74	63.77	72.40
both	**80.82**	**86.34**	**72.46**	**78.79**	**77.64**	**85.20**	**65.99**	**74.38**

Table 2: Ablation study on the instance-level and token-level weight.

λ_s	Detection				Correction			
	Acc.	Prec.	Rec.	F1.	Acc.	Prec.	Rec.	F1.
0.85	80.36	85.87	71.90	78.27	77.36	84.76	65.80	74.09
0.90	**80.82**	86.34	**72.46**	**78.79**	77.64	85.20	**65.99**	**74.38**
0.95	80.73	**86.97**	71.53	78.50	**77.73**	**85.92**	65.43	74.29

Table 3: Impact of different values of λ_s.

brings about more improvements over both NBCL (+1.17% F1) and UACL (+0.65% F1), indicating that our automatic SSCL is superior to the previous approaches that require careful design for data difficulty and training curricula; and 3) our SSCL approach establishes a new SOTA (74.38% F1).

3.3 Effects of Instance-Level Weight and Token-Level Weight

We carried out an ablation study for the instance-level weight and token-level weight mechanisms. The results are shown in Table 2.

Table 2 depicts that the instance-level weight brings more improvements (+0.80% F1) than the token weight. But they are complementary and their combination leads to the best performance.

3.4 Effects of Hyperparameter λ_s

We study the effects of the hyperparameter λ_s (in Equation 7), and the results are shown in Table 3.

A larger λ_s value means a more elaborate CL process for the model. Table 3 shows that the highest F1 score was obtained with 0.90 as λ_s, which indicates that 0.9 is a proper value for the learning with the curriculum.

4 Related Work

Spelling Error Correction. SEC is helpful for many applications, such as essay scoring (Burstein and Chodorow, 1999), search (Martins and Silva, 2004; Gao et al., 2010), Optical Character Recognition (OCR) (Afli et al., 2016), machine translation and tagging (Heigold et al., 2018), and many studies have been conducted on the SEC task. Unsupervised approaches using language models and rules (Yu and Li, 2014; Tseng et al., 2015) are widely adopted. SEC is treated as a sequential labeling problem in machine learning approaches, and conditional random fields or hidden Markov models (Tseng et al., 2015; Zhang et al., 2015) are previously employed. Recently, Guo et al. (2019); Wang et al. (2019a) apply deep learning approaches to spelling error correction, and based on the BERT encoder, Hong et al. (2019) build a seq2seq model for SEC.

Curriculum Learning. CL (Bengio et al., 2009) aims to facilitate the model training in an easy-to-hard order, which leads to improved model performance (Tsvetkov et al., 2016; Sachan and Xing,

2016; Amiri et al., 2017). Many studies adopt CL to reinforce learning to optimize the model parameters (Saito, 2018; Kumar et al., 2019). CL also has shown to be useful for data processing to improve the quality of the training data (Huang and Du, 2019). Recently, CL has been widely employed in the machine learning for NLP. It improves the performance and the training efficiency of the NMT models based on linguistic features (Liu et al., 2020; Zhou et al., 2020; Wang et al., 2020a), enhances the multi-domain correlation, and addresses the domain imbalance issue (Wang et al., 2020b). It also has been explored in other tasks, such as response generation (Shen and Feng, 2020) and reading comprehension (Tay et al., 2019).

Self-Supervised Learning. The basic idea of self-supervised learning (SSL) is to **automatically** generate or find supervision signals to solve tasks. For instance, it is used to learn representations from unlabeled data (Raina et al., 2007; Bengio et al., 2013). Tang et al. (2019) use SSL to mine useful attention supervision information from the training corpus to refine attention mechanisms. Kedia and Chinthakindi (2021) combine the SSL with pseudo-labels and meta-learning during inference to improve generalization. Ruiter et al. (2019) use an emergent NMT system to simultaneously select training data and learn internal NMT representations in a SSL way without parallel data. SSL is also adopted to solve many other problems, such as document-level context or sentence summarization (West et al., 2019; Wang et al., 2019b), dialogue learning (Wu et al., 2019), improving data scarcity or labeling costs (Fu et al., 2020; Yuan et al., 2020) and generating meta-learning tasks from unlabeled text (Bansal et al., 2020).

Comparison to Previous Work. Compared to previous CL studies, we apply SSL to CL and propose SSCL that uses the model to measure data difficulty for training instance selection in an easy-to-hard order. Compared to previous SEC approaches, we employ SSCL for the training of SEC, which establishes a new SOTA (74.38% F1) on the SIGHAN 2015 Chinese spelling check task.

5 Conclusion

In this paper, we applied curriculum learning to spelling error correction and present a novel Self-Supervised Curriculum Learning method.

We verify the effectiveness of the SSCL approach on the SIGHAN 2015 Chinese spelling check task. Experiment results show that SSCL is able to significantly improve the performance of the state-of-the-art Soft-Masked BERT model and establishes a new state-of-the-art performance (74.38% F1). The fact that SSCL brings about more improvements than the previous norm-based and uncertainty-aware CL approaches also supports its effectiveness as a CL approach.

Acknowledgements

We thank our anonymous reviewers for their insightful comments. We thank Yue Zhang and Josef van Genabith for constructive suggestions. Zifa Gan and Hongyin Zan acknowledge the support of the National Social Science Fund of China (Grant No. 17ZDA138 and Grant No. 14BYY096). Hongfei Xu is supported by the German Federal Ministry of Education and Research (BMBF) under funding code 01IW20010 (CORA4NLP) and China Scholarship Council ([2018]3101, 201807040056).

References

Haithem Afli, Zhengwei Qiu, Andy Way, and Páraic Sheridan. 2016. Using SMT for OCR error correction of historical texts. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 962–966, Portorož, Slovenia. European Language Resources Association (ELRA).

Hadi Amiri, Timothy Miller, and Guergana Savova. 2017. Repeat before forgetting: Spaced repetition for efficient and effective training of neural networks. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2401–2410, Copenhagen, Denmark. Association for Computational Linguistics.

Trapit Bansal, Rishikesh Jha, Tsendsuren Munkhdalai, and Andrew McCallum. 2020. Self-supervised meta-learning for few-shot natural language classification tasks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 522–534, Online. Association for Computational Linguistics.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48.

Jill Burstein and Martin Chodorow. 1999. Automated essay scoring for nonnative English speakers. In *Computer Mediated Language Assessment and Evaluation in Natural Language Processing*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Tsu-Jui Fu, Xin Wang, Scott Grafton, Miguel Eckstein, and William Yang Wang. 2020. SSCR: Iterative language-based image editing via self-supervised counterfactual reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4413–4422, Online. Association for Computational Linguistics.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR.

Jianfeng Gao, Xiaolong Li, Daniel Micol, Chris Quirk, and Xu Sun. 2010. A large scale ranker-based system for search query spelling correction. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 358–366, Beijing, China. Coling 2010 Organizing Committee.

Shuhao Gu, Jinchao Zhang, Fandong Meng, Yang Feng, Wanying Xie, Jie Zhou, and Dong Yu. 2020. Token-level adaptive training for neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1035–1046, Online. Association for Computational Linguistics.

Jinxi Guo, Tara N Sainath, and Ron J Weiss. 2019. A spelling correction model for end-to-end speech recognition. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5651–5655. IEEE.

Georg Heigold, Stalin Varanasi, Günter Neumann, and Josef van Genabith. 2018. How robust are character-based word embeddings in tagging and MT against wrod scramlbing or randdm nouse? In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 68–80, Boston, MA. Association for Machine Translation in the Americas.

Yuzhong Hong, Xianguo Yu, Neng He, Nan Liu, and Junhui Liu. 2019. FASPell: A fast, adaptable, simple, powerful Chinese spell checker based on DAE-decoder paradigm. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 160–169, Hong Kong, China. Association for Computational Linguistics.

Yuyun Huang and Jinhua Du. 2019. Self-attention enhanced CNNs and collaborative curriculum learning for distantly supervised relation extraction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 389–398, Hong Kong, China. Association for Computational Linguistics.

Akhil Kedia and Sai Chetan Chinthakindi. 2021. Keep learning: Self-supervised meta-learning for learning from inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 63–77, Online. Association for Computational Linguistics.

Tom Kocmi and Ondřej Bojar. 2017. Curriculum learning and minibatch bucketing in neural machine translation. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 379–386, Varna, Bulgaria. INCOMA Ltd.

Gaurav Kumar, George Foster, Colin Cherry, and Maxim Krikun. 2019. Reinforcement learning based curriculum optimization for neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2054–2061, Minneapolis, Minnesota. Association for Computational Linguistics.

Xuebo Liu, Houtim Lai, Derek F. Wong, and Lidia S. Chao. 2020. Norm-based curriculum learning for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 427–436, Online. Association for Computational Linguistics.

Bruno Martins and Mário J Silva. 2004. Spelling correction for search engine queries. In *International Conference on Natural Language Processing (in Spain)*, pages 372–383. Springer.

Emmanouil Antonios Platanios, Otilia Stretcu, Graham Neubig, Barnabas Poczos, and Tom Mitchell. 2019. Competence-based curriculum learning for neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1162–1172, Minneapolis, Minnesota. Association for Computational Linguistics.

Rajat Raina, Alexis Battle, Honglak Lee, Benjamin Packer, and Andrew Y Ng. 2007. Self-taught learning: transfer learning from unlabeled data. In *Proceedings of the 24th international conference on Machine learning*, pages 759–766.

Dana Ruiter, Cristina España-Bonet, and Josef van Genabith. 2019. Self-supervised neural machine

translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1828–1834, Florence, Italy. Association for Computational Linguistics.

Dana Ruiter, Josef van Genabith, and Cristina España-Bonet. 2020. Self-induced curriculum learning in self-supervised neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2560–2571, Online. Association for Computational Linguistics.

Mrinmaya Sachan and Eric Xing. 2016. Easy questions first? a case study on curriculum learning for question answering. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 453–463, Berlin, Germany. Association for Computational Linguistics.

Atsushi Saito. 2018. Curriculum learning based on reward sparseness for deep reinforcement learning of task completion dialogue management. In *Proceedings of the 2018 EMNLP Workshop SCAI: The 2nd International Workshop on Search-Oriented Conversational AI*, pages 46–51, Brussels, Belgium. Association for Computational Linguistics.

Lei Shen and Yang Feng. 2020. CDL: Curriculum dual learning for emotion-controllable response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 556–566, Online. Association for Computational Linguistics.

Jialong Tang, Ziyao Lu, Jinsong Su, Yubin Ge, Linfeng Song, Le Sun, and Jiebo Luo. 2019. Progressive self-supervised attention learning for aspect-level sentiment analysis. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 557–566, Florence, Italy. Association for Computational Linguistics.

Yi Tay, Shuohang Wang, Anh Tuan Luu, Jie Fu, Minh C. Phan, Xingdi Yuan, Jinfeng Rao, Siu Cheung Hui, and Aston Zhang. 2019. Simple and effective curriculum pointer-generator networks for reading comprehension over long narratives. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4922–4931, Florence, Italy. Association for Computational Linguistics.

Yuen-Hsien Tseng, Lung-Hao Lee, Li-Ping Chang, and Hsin-Hsi Chen. 2015. Introduction to SIGHAN 2015 bake-off for Chinese spelling check. In *Proceedings of the Eighth SIGHAN Workshop on Chinese Language Processing*, pages 32–37, Beijing, China. Association for Computational Linguistics.

Yulia Tsvetkov, Manaal Faruqui, Wang Ling, Brian MacWhinney, and Chris Dyer. 2016. Learning the curriculum with Bayesian optimization for task-specific word representation learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 130–139, Berlin, Germany. Association for Computational Linguistics.

Yu Wan, Baosong Yang, Derek F. Wong, Yikai Zhou, Lidia S. Chao, Haibo Zhang, and Boxing Chen. 2020. Self-paced learning for neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1074–1080, Online. Association for Computational Linguistics.

Chengyi Wang, Yu Wu, Shujie Liu, Ming Zhou, and Zhenglu Yang. 2020a. Curriculum pre-training for end-to-end speech translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3728–3738, Online. Association for Computational Linguistics.

Dingmin Wang, Yan Song, Jing Li, Jialong Han, and Haisong Zhang. 2018. A hybrid approach to automatic corpus generation for Chinese spelling check. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2517–2527, Brussels, Belgium. Association for Computational Linguistics.

Dingmin Wang, Yi Tay, and Li Zhong. 2019a. Confusionset-guided pointer networks for Chinese spelling check. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5780–5785, Florence, Italy. Association for Computational Linguistics.

Hong Wang, Xin Wang, Wenhan Xiong, Mo Yu, Xiaoxiao Guo, Shiyu Chang, and William Yang Wang. 2019b. Self-supervised learning for contextualized extractive summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2221–2227, Florence, Italy. Association for Computational Linguistics.

Wei Wang, Ye Tian, Jiquan Ngiam, Yinfei Yang, Isaac Caswell, and Zarana Parekh. 2020b. Learning a multi-domain curriculum for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7711–7723, Online. Association for Computational Linguistics.

Peter West, Ari Holtzman, Jan Buys, and Yejin Choi. 2019. BottleSum: Unsupervised and self-supervised sentence summarization using the information bottleneck principle. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3752–3761, Hong Kong, China. Association for Computational Linguistics.

Jiawei Wu, Xin Wang, and William Yang Wang. 2019. Self-supervised dialogue learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3857–3867, Florence, Italy. Association for Computational Linguistics.

Hongfei Xu, Qiuhui Liu, Josef van Genabith, Deyi Xiong, and Jingyi Zhang. 2020. Lipschitz constrained parameter initialization for deep transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 397–402, Online. Association for Computational Linguistics.

Junjie Yu and Zhenghua Li. 2014. Chinese spelling error detection and correction based on language model, pronunciation, and shape. In *Proceedings of The Third CIPS-SIGHAN Joint Conference on Chinese Language Processing*, pages 220–223, Wuhan, China. Association for Computational Linguistics.

Liang-Chih Yu, Lung-Hao Lee, Yuen-Hsien Tseng, and Hsin-Hsi Chen. 2014. Overview of SIGHAN 2014 bake-off for Chinese spelling check. In *Proceedings of The Third CIPS-SIGHAN Joint Conference on Chinese Language Processing*, pages 126–132, Wuhan, China. Association for Computational Linguistics.

Michelle Yuan, Hsuan-Tien Lin, and Jordan Boyd-Graber. 2020. Cold-start active learning through self-supervised language modeling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7935–7948, Online. Association for Computational Linguistics.

Shaohua Zhang, Haoran Huang, Jicong Liu, and Hang Li. 2020. Spelling error correction with soft-masked BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 882–890, Online. Association for Computational Linguistics.

Shuiyuan Zhang, Jinhua Xiong, Jianpeng Hou, Qiao Zhang, and Xueqi Cheng. 2015. HANSpeller++: A unified framework for Chinese spelling correction. In *Proceedings of the Eighth SIGHAN Workshop on Chinese Language Processing*, pages 38–45, Beijing, China. Association for Computational Linguistics.

Xuan Zhang, Pamela Shapiro, Gaurav Kumar, Paul McNamee, Marine Carpuat, and Kevin Duh. 2019. Curriculum learning for domain adaptation in neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1903–1915, Minneapolis, Minnesota. Association for Computational Linguistics.

Yikai Zhou, Baosong Yang, Derek F. Wong, Yu Wan, and Lidia S. Chao. 2020. Uncertainty-aware curriculum learning for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6934–6944, Online. Association for Computational Linguistics.