Professional Text Review Under Limited Sampling Constraints

Leiwen Yang[1], Tao Yang[1], Feng Yuan[2], and Yuqing Sun[1(✉)]

[1] School of Software, Shandong University, Jinan, China
sun_yuqing@sdu.edu.cn
[2] Shandong Shanda Oumasoft Co., Ltd., Jinan, China
sdyuanf@sina.com

Abstract. Text review is a task that determines whether the knowledge expression in a student answer is consistent with a given reference answer. In the professional scenarios, the number of labeled samples is limited, usually ranging from dozens to hundreds, which makes the text review task more challenging. This paper proposes a text review method based on data augmentation, which is performed by the combination of different positive and negative labeled samples. The review model infers the unlabeled samples, where the pseudo-labeled samples with the high confidences are selected for the subsequent training rounds. Experimental results in real national qualification exam datasets show that our method has improvement compared with the traditional method on the text review task under the limited sampling constraints.

Keywords: Text review · Limited sampling constraints · Data augmentation · Self-Training

1 Introduction

The subjective questions are commonly used in the professional examinations. Experts review the student answers by comparing their expressions with the reference answer from the point of knowledge. Expert reviewing is time-consuming and labor-intensive. Typically, we can get a limited quantity of expert reviewed samples so as to train the review model. Essentially student answers objectively reflect the different cognitive viewpoints and the text review task need infer the professional knowledge contained within the student answers. Thus this task is more complex than the traditional text inference.

In this paper, we propose a professional text review method based on data augmentation. We combine various labeled samples with different ratio for data

This work was supported by the National Nature Science Foundation of China, NSFC (62376138) and the Innovative Development Joint Fund Key Projects of Shandong NSF (ZR2022LZH007).

L. Yang and T. Yang—Equal contribution.

Y. Sun et al. (Eds.): ChineseCSCW 2023, CCIS 2013, pp. 287–295, 2024.
https://doi.org/10.1007/978-981-99-9640-7_21

augmentation. We adopt multiple rounds of self-training and obtain the pseudo-labeled samples with the confidences of predication by the review model infers unlabeled samples, where the selected pseudo-samples for the subsequent training rounds.

Experiments are conducted on the real national qualification exam datasets. The experimental results shown that the effectiveness of our method. By comparing with the traditional self-training methods, our proposed self-training model has the capability to mitigate the problem of error accumulation.

2 Related Work

2.1 Text Inference

Text inference involves deducing semantics and categorizing text into the predefined classes. The representative researches primarily based on the semantic similarity between text pair. For example, two texts are encoded separately and combined by the operations like Hadamard product to measure the similarity. Then the resulting semantic vector is fed into a classifier [1] to obtain the inference results [2,3]. The drawback of such methods lies in their independent encoding of two texts, which lacks the deep semantic interactions. Another way is to employ the attention mechanisms at the phrases, sentences, or syntax to obtain the alignment features of each text pair, which are fed into a classifier for inference [4]. For example, the pre-trained language model BERT [5] is used for text inference. Some methods introduce the external knowledge to assist inference. For example, zhang et al. combine the text semantic information with text rule information to enhance the performance [6], while Li et al. [7] introduce the supervised contrastive loss [8] on the interaction vectors of text pair to assist inference.

2.2 Low-Resource Learning

Low-resource learning refers to training model by a limited number of labeled samples. Prompt learning inserts the prompts and masks into the input text, it transforms the downstream tasks to predict the vocabulary in the masked spaces [9]. The MAML algorithm [10] uses a series of learning tasks to comprise the support and query sets for train the model. By addressing the problem of non-overlapping potential reasoning logic between texts [11], high-quality meta-learning tasks can be constructed. Self-training augments the training dataset with the pseudo-samples obtained by unlabeled samples [12]. Data augmentation is an important method to addressing low-resource issues. For example, the UDA [13] employs reverse translation on the unlabeled texts and synonym replacement techniques to perform data augmentation. By iteratively performing reverse translation on the unlabeled texts and using weighted averaging with sharpening [14], the soft labels are generated to optimize consistency between the unlabeled text and its copies. Utilize the idea of the virtual adversarial training [15], the KL divergence between the probability distributions of the original

text and the perturbed text output is computed as a measure to evaluate the quality of augmented samples. Above methods solely focus on individual text and do not consider the complexity of semantic relationships between text pair.

3 Method

The overall framework of our method is illustrated in Fig. 1. It consists of three main components: data augmentation, model training, and pseudo-sample acquisition. Data augmentation is responsible for generating more samples for model training, while pseudo-sample acquisition process provides the high-quality samples for the self-training process. Each of these components are detailed below, separately.

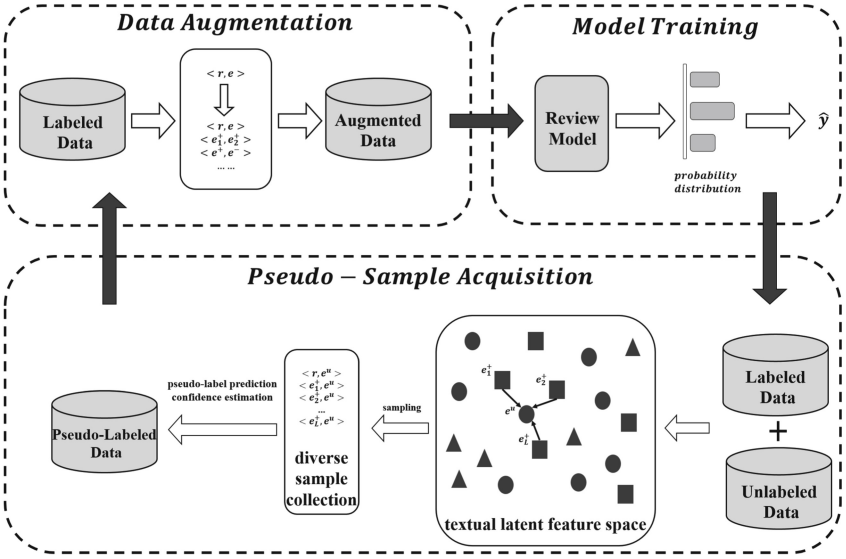

Fig. 1. The framework of professional text review based on data augmentation

3.1 Data Augmentation

The form of review sample is defined as a triplet $< r, e, y >$, where r, e, y represents the reference answer, student answer, label, respectively. Furthermore, we use e^+, e^-, and e^u represents the positive, negative, and unlabeled student answers, respectively.

Data Augmentation Based on Text Pair Construction. The traditional review methods involve the semantic reasoning only between r and e. Since the semantics of e^+ and r are similar, we construct the new meaningful text pairs for data augmentation by the following patterns, as shown in Table 1, where $\alpha, \beta, \gamma, \eta$ are the hyper-parameters and $\alpha + \beta + \gamma = 1$.

Table 1. Construction patterns for data augmentation.

Label	Description	Pattern	Proportion
positive	The text pair has consistent semantics.	$< r, e^+ >$	$\alpha * \eta$
		$< e_1^+, e_2^+ >$	$\alpha * (1 - \eta)$
negative	The text pair has inconsistent semantics.	$< r, e^- >$	$\beta * \eta$
		$< e^+, e^- >$	$\beta * (1 - \eta)$
neutral	The text pair has no inferential relationship.	$< r, e^* >$	$\gamma * \eta$
		$< e, e^* >$	$\gamma * (1 - \eta)$

3.2 Model Training

The text review model adopts an 'encoder + reasoner' architecture. The encoder employs BERT [5], and the reasoner utilizes a multi-layer perceptron (MLP) and the softmax function.

The texts r and e are concatenated with $[SEP]$ as the input for the encoder. After encoding, the $[CLS]$ token embedding, denoted as $\mathbf{x}$ (Eq. 1), is fed to the reasoner.It generates the probabilities for different classes (Eq. 2), where the classes $Y = \{1, 2, 3\}$ represent positive, negative, and neutral, respectively. We take the class with the highest probability as the review result (Eq. 3). The cross-entropy loss function is adopted as the objective function.

$$\mathbf{x} = \text{BERT}\,(r[SEP]e)_{[CLS]} \tag{1}$$

$$[p_1, p_2, p_3] = softmax(\text{MLP}(\mathbf{x})) \tag{2}$$

$$\hat{y} = \arg\max_{i \in Y} p_i \tag{3}$$

3.3 Self-training Process

We adopt τ-round self-training mode. M_t represents the model trained in the t-th round. The following section describes the details on how to acquire the pseudo-samples after training the model in each round.

Pseudo-sample Acquisition based on Multi-sample joint Reasoning. First, we compute the TF-IDF vectors for all texts in the dataset. Then we randomly choose an unlabeled student answer e^u, and select the kNN positive samples for e^u forming a diverse sample collection E, which contains the $L-1$ e^+ and e^u. Next, we inference the sample in E using M_t, and vote inference results to obtain the pseudo-label y^u for e^u. For the pseudo-sample (e^u, r, y^u). we calculate the confidence b^u for it (Eq. 4,5), where $p_{l,i}$ represents the inference probability being i in the l-th e^+. A larger value of b^u indicates higher confidence. Lastly, we sample pseudo-samples of different class without replacement, following the descending order of b^u. Then adds the sampled pseudo-samples to the training set of the next round, resulting in the dataset D_{t+1}.

$$\delta_l = \min_{i \in C \wedge i \neq y^u} p_{l,y^u} - p_{l,i} \tag{4}$$

$$b^u = \sum_{l=1}^{L} \delta_l \tag{5}$$

4 Experiment

4.1 Datasets

To prevent data leakage, we uses a dataset different from the NQE (Section 4.2) for professional text review. The NQE-30 dataset includes 15 questions from the years 2018 to 2021, where each question set includes two sub-questions, denoted by *sq* in the following discussion. Table 2 shows the statistics. Here, the naming patten is *LetterNumber* like A1, A2, and etc., where the letters represent the question number, and the digits represent the sub-question number. During training, only a limited quantity of samples are extracted to simulate limited sample constraints, and the labels are removed from the samples that are not extracted to serve as unlabeled samples.

Table 2. Number of samples in the NQE-30 dataset (thousand)

Datasets	Label		Total	Datasets	Label		Total	Datasets	Label		Total
	Pos	Neg			Pos	Neg			Pos	Neg	
A1	5.5	23.7	29.2	F1	4.1	5.6	9.7	K1	29.5	17.8	47.3
A2	13.4	15.8	29.2	F2	8.2	1.5	9.7	K2	35.2	12.1	47.3
B1	4.0	25.0	29.0	G1	2.3	3.6	5.9	L1	20.7	27.4	48.1
B2	19.5	9.5	29.0	G2	2.9	3.0	5.9	L2	40.1	8.0	48.1
C1	3.8	49.5	53.3	H1	2.0	10.9	12.9	M1	5.4	41.9	47.3
C2	24.0	29.3	53.3	H2	7.8	5.1	12.9	M2	33.6	13.7	47.3
D1	14.0	34.0	48.0	I1	4.7	43.4	48.1	N1	7.8	42.9	50.7
D2	20.9	27.1	48.0	I2	42.7	5.4	48.1	N2	11.2	39.5	50.7
E1	1.0	12.7	13.7	J1	8.3	39.7	48.0	O1	10.0	12.8	22.8
E2	1.5	12.2	13.7	J2	34.8	13.2	48.0	O2	19.4	3.4	22.8

4.2 Comparative Methods

We mainly change the encoder (Section 3.2) for comparative experiments. Baseline employs BERT_{base} [5] as the encoder. For comparison, RoBERTa_{base} [16] and SBERT_{base} [3] are also employed as encoder. To adapt the review model to the professional domain, we utilize the labeled data that covers a few years of the National Qualification Examination (NQE) for pre-training the model (NQEPT), which is employed as the encoder.

Below are variant models that utilize our proposed method. Models with the suffixes *-aug, *-ST, and *-MST represent the review model trained using data augmentation, traditional self-training and multi-sample joint reasoning for self-training, respectively.

4.3 Analysis of Experiments in Professional Text Review

To verify the performance of our method, we extract 50, 100, and 200 training samples for each sub-question in the NQE-30 dataset for experiments. The results are shown in Table 3.

Table 3. Method comparison on the NQE-30 dataset (Acc and Ma-F1 represents accuracy and macro F1, respectively)

Method		Number					
		$50/sq$		$100/sq$		$200/sq$	
		Acc	Ma-F1	Acc	Ma-F1	Acc	Ma-F1
BERT_{base} (baseline)		93.30%	93.16	94.59%	94.48	96.00%	95.92
SBERT_{base}		89.39%	88.10	93.75%	93.61	95.76%	95.67
RoBERTa_{base}		93.16%	92.96	94.73%	94.60	95.72%	95.62
BERT_{base}-aug		94.19%	94.06	95.18%	95.07	96.10%	96.02
BERT_{base}-aug-MST	$\tau = 1$	94.40%	94.26	95.47%	95.38	**96.26%**	**96.17**
	$\tau = 2$	94.62%	94.49	**95.57%**	**95.47**	–	–
	$\tau = 3$	**94.65%**	**94.53**	–	–	–	–
NQEPT		94.60%	94.48	95.01%	94.91	96.03%	95.94
NQEPT-aug		95.03%	94.92	95.60%	95.51	96.34%	96.25
NQEPT-aug-MST	$\tau = 1$	**95.17%**	**95.08**	95.70%	95.60	**96.42%**	**96.33**
	$\tau = 2$	95.11%	95.02	**95.82%**	**95.73**	–	–
	$\tau = 3$	95.06%	94.97	–	–	–	–

The performances for all methods improve with the increasing size of training dataset. Comparing the results of BERT_{base} with BERT_{base}-aug, we can see that the data augmentation method can achieve an accuracy improvement of $0.1\% \sim 0.89\%$. It is similar with the case of NQEPT with NQEPT-aug. Comparing

the results of BERT$_{base}$-aug with BERT$_{base}$-aug-MST and NQEPT-aug with NQEPT-aug-MST, it can be seen that the self-training method based on multi-sample joint inference can achieve an accuracy improvement of 0.08% $\sim$ 0.46%. Almost every round of self-training is better than the results of the previous round of self-training. In addition, the review model using NQEPT as the encoder is better than the review model using BERT$_{base}$ as the encoder under the same conditions, proving that the professional knowledge learned from pre-training the review model has good transferability.

4.4 Analysis of Self-training Experiments

To verify the performance of our improved self-training method, we randomly extract 100 samples for each sub-question from the NQE-30 dataset. We also conduct the experiments to assess the impact of different inference sample numbers $L = \{3, 5, 7\}$. The results are shown in Table 4.

Table 4. The accuracy(Acc) of each self-training and error(Err) ratio of pseudo-label under the sample size of $100/sq$

Method		Round			
		1		2	
		Acc	Err	Acc	Err
BERT$_{base}$-ST		95.02%	2.82%	95.20%	3.32%
BERT$_{base}$-aug-MST	$L = 3$	95.50%	1.56%	94.49%	2.30%
	$L = 5$	95.48%	1.68%	95.54%	2.46%
	$L = 7$	95.47%	1.65%	95.57%	2.51%

In each round of self-training, both BERT$_{base}$-aug-MST and BERT$_{base}$-ST show the improvement in accuracy and exceed the 94.59% accuracy of BERT$_{base}$ under the same conditions. The accuracy of BERT$_{base}$-aug-MST always higher than BERT$_{base}$-ST. This proves that our self-training method is more effective than the traditional self-training methods, and indicates that single inference in traditional self-training methods is more likely to introduce bias.

In the two rounds of self-training, the error rate of pseudo-sample sampling of BERT$_{base}$-aug-MST is significantly lower than BERT$_{base}$-ST. Our method is better than traditional method, and achieves the best at $L = 3$. Although the inference sample number $L = \{5, 7\}$ has a relatively high error rate, its accuracy is still slightly higher than $L = 3$, which shows that the quality of pseudo-samples increases with the number of inference samples.

5 Conclusion

The paper proposes a text review method based on data augmentation, designed for the task of professional text review under limited sampling constraints.

We also propose the self-training strategy for training the model, where the high-confidence pseudo-samples are obtained for the next round of self-training. Experimental results demonstrate that the proposed data augmentation method and self-training strategy exhibit excellent performance on real professional datasets, which also satisfy the constraints of limited samples. For the future research, we would explore the relevant domain knowledge to enhance the understanding and reasoning the professional semantics of text.

References

1. Tai, K.S., Socher, R., Manning, C.D.: Improved semantic representations from tree-structured long short-term memory networks. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 1556–1566 (2015)
2. Conneau, A., Kiela, D., Schwenk, H., Barrault, L., Bordes, A.: Supervised learning of universal sentence representations from natural language inference data. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 670–680. Association for Computational Linguistics (2017)
3. Reimers, N., Gurevych, I.: Sentence-BERT: sentence embeddings using Siamese BERT-networks. 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). Association for Computational Linguistics (2019)
4. Li, D., Liu, T., Pan, W., Liu, X., Sun, Y., Yuan, F.: Grading Chinese answers on specialty subjective questions. In: Sun, Y., Lu, T., Yu, Z., Fan, H., Gao, L. (eds.) ChineseCSCW 2019. CCIS, vol. 1042, pp. 670–682. Springer, Singapore (2019). https://doi.org/10.1007/978-981-15-1377-0_52
5. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In Proceedings of NAACL-HLT, pp. 4171–4186 (2019)
6. Zhang, Z., et al.: Semantics-aware BERT for language understanding. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 9628–9635 (2020)
7. Li, S., Hu, X., Lin, L., Wen, L.: Pair-level supervised contrastive learning for natural language inference. In: ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 8237–8241. IEEE (2022)
8. Khosla, P., et al.: Supervised contrastive learning. Adv. Neural. Inf. Process. Syst. **33**, 18661–18673 (2020)
9. Schick, T., Schütze, H.: Exploiting cloze-questions for few-shot text classification and natural language inference. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pp. 255–269 (2021)
10. Finn, C., Abbeel, P., Levine, S.: Model-agnostic meta-learning for fast adaptation of deep networks. In: International Conference on Machine Learning, pp. 1126–1135. PMLR (2017)
11. Murty, S., Hashimoto, T.B., Manning, C.D.: DReCa: a general task augmentation strategy for few-shot natural language inference. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 1113–1125 (2021)

12. Pseudo-Label, D.-H.L.: The simple and efficient semi-supervised learning method for deep neural networks. In: ICML 2013 Workshop: Challenges in Representation Learning, pp. 1–6 (2013)
13. Xie, Q., Dai, Z., Hovy, E., Luong, T., Le, Q.: Unsupervised data augmentation for consistency training. Adv. Neural. Inf. Process. Syst. **33**, 6256–6268 (2020)
14. Chen, J., Yang, Z., Yang, D.: MixText: linguistically-informed interpolation of hidden space for semi-supervised text classification. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 2147–2157 (2020)
15. Miyato, T., Dai, A.M., Goodfellow, I.: Adversarial training methods for semi-supervised text classification. In: International Conference on Learning Representations (2016)
16. Liu, Y., et al.: Roberta: a robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692 (2019)