Reducing Spurious Correlations for Answer Selection by Feature Decorrelation and Language Debiasing

Zeyi Zhong [1,2], Min Yang [1]*, Ruifeng Xu [3]
[1] Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences
[2] University of Science and Technology of China
[3] Harbin Institute of Technology (Shenzhen)
{zhongzy,min.yang}@siat.ac.cn, xuruifeng@hit.edu.cn

Abstract

Deep neural models have become the mainstream in answer selection, yielding state-of-the-art performance. However, these models tend to rely on spurious correlations between prediction labels and input features, which in general suffer from robustness and generalization. In this paper, we propose a novel **S**purious **C**orrelation reduction method to improve the robustness of the neural **AN**swer selection models (SCAN) from the sample and feature perspectives by removing the feature dependencies and language biases in answer selection. First, from the sample perspective, we propose a feature decorrelation module by learning a weight for each instance at the training phase to remove the feature dependencies and reduce the spurious correlations without prior knowledge of such correlations. Second, from the feature perspective, we propose a feature debiasing module with contrastive learning to alleviate the negative language biases (spurious correlations) and further improve the robustness of the AS models. Experimental results on three benchmark datasets show that SCAN achieves substantial improvements over strong baselines. For reproducibility, we will release our code and data at https://github.com/xish9/SCAN.

1 Introduction

Answer selection, which aims to select the most applicable answers from an answer candidate pool, has broad applications in information retrieval (IR) and natural language processing (NLP). Conventional answer selection methods primarily focus on designing various features, such as syntactic features (Li, 2003), dependency trees (Wang et al., 2007), and translation features (Surdeanu et al., 2008). However, the remarkable success of these methods relies heavily on feature engineering, which is a labor-intensive and time-consuming process.

Subsequently, deep neural models (Qiu and Huang, 2015; Guo et al., 2017; Tay et al., 2017; Zhou et al., 2018) have been widely employed for answer selection and become the mainstream techniques for answer selection by automatically learning the contextual representations of questions and answers. To capture the relationships between the question-answer pairs, different attention mechanisms (Zhang et al., 2017; Tay et al., 2018a; Shen et al., 2018; Yang et al., 2019a; Xie et al., 2020) have been proposed to learn the interactive features of the questions and the answers. Recently, pre-trained language models, such as BERT (Devlin et al., 2018) and RoBERTa (Liu et al., 2019), have been proposed and applied to answer selection (Garg et al., 2020; MacAvaney et al., 2020; Zhang et al., 2021a), obtaining the state-of-the-art results.

Despite the remarkable progress of previous works, these deep neural models are prone to rely on spurious correlations between input features and prediction labels, which capture the prediction correlations that hold for most training samples but do not hold in general. The spurious correlations limit the robustness and generalization ability of the neural AS models to the out-of-distribution and challenging datasets. In particular, for answer selection, the word-overlap between the question and the answers is highly correlated with the relevance prediction label. Thus, the deep AS models perform poorly on the out-of-distribution or challenging corpora that cannot be tackled with these superficial correlations (e.g., word overlap). This issue is also referred to as dataset bias (Clark et al., 2019) and data distribution shift (Sagawa et al., 2020).

In this paper, we propose a novel **S**purious **C**orrelation reduction method to improve the robustness of the neural **AN**swer selection models (SCAN) from the sample and feature perspectives by removing the feature dependencies and language biases in answer selection. First, from the sample perspective, we employ the feature decorre-

*Min Yang is corresponding author.

Proceedings of the 29th International Conference on Computational Linguistics, pages 1753–1764
October 12–17, 2022.

lation module based on Random Fourier Features (Rahimi and Recht, 2007) to decorrelate the relevant and irrelevant features by learning a weight for each sample during the training phase, which facilitate the deep AS models to reduce spurious correlations and concentrate on the true discriminative features (relevant features) for label prediction. Second, from the feature perspective, we propose a feature debiasing module with contrastive learning to weaken the negative biases in language and improve the robustness of the AS models. Concretely, the feature debiasing module aims to make the base contextual representation of input sample close to the debiased features and away from the negative bias features.

Our main contributions are three-fold:

- We propose a feature decorrelation module by learning a weight for each training instance to remove the feature dependencies and reduce the spurious correlations without prior knowledge of such correlations.
- We propose a feature debiasing module with contrastive learning to alleviate the negative language biases (spurious correlations) and improve the robustness of the AS models.
- Experimental results show that our SCAN method achieves substantial improvements over the state-of-the-art baseline methods for answer selection.

2 Related Work

2.1 Deep Learning for Answer Selection

Answer selection has received remarkable attention in various tasks, such as dialogue systems (Yuan et al., 2019; He et al., 2022b,a), knowledge base question answering (Niu et al., 2021; Saxena et al., 2020), and information retrieval (Li et al., 2021). So far, deep learning approaches have become the mainstream in answer selection (AS) due to their impressive improvement. Severyn and Moschitti (2015) was an early representative neural AS model, which utilized convolutional neural network (CNN) to learn question and answer representations separately followed by a similarity function to compute the relevance score. Tay et al. (2017) extended the long short-term memory (LSTM) network with holographic composition for sentence modeling and semantic matching. Several works (Yin et al., 2016; Tan et al., 2016) explored different attention mechanisms to capture the relations between sentences. For example, Tay et al. (2018c) proposed a casted attention for feature augmentation to improve the representation learning process. Shen et al. (2017b) proposed an inter-weighted alignment network, which utilized the word-level similarity matrix to explore the fine-grained alignment of two sentences. Tay et al. (2018b) presented HyperQA which leveraged a parameter efficient network to model the relations between the question and answer representations with PLMs in the Hyperbolic space instead of the Euclidean space.

Recently, the pre-trained language models (PLMs), such as BERT (Devlin et al., 2018) and RoBERTa (Liu et al., 2019), have been applied to the answer selection task and yielded state-of-the-art results by capturing rich linguistic knowledge from large textual corpora. Yoon et al. (2019) employed ELMo (Peters et al., 2018) to a compare aggregate architecture, which leveraged the latent-cluster information to enhance the AS model. Lai et al. (2019) combined a gated self-attention memory network and the pre-trained language models for answer selection. Garg et al. (2020) proposed a two-step transfer-adapt (TANDA) method, which fine-tuned the pre-trained language models by using a large QA dataset ASNQ. Recently, Zhang et al. (2021b, 2022) focused on exploiting the interrelated information between candidate answers and obtained the best results for answer selection.

2.2 Spurious Correlation Reduction in NLP

Despite the remarkable progress made by deep neural networks, some studies (Gururangan et al., 2018; McCoy et al., 2019; Zhang et al., 2021a) have revealed that the deep models often relied on spurious correlations between the learned features and the prediction labels, making the deep models unstable and not generalize well to the data with different distributions. For example, previous studies (Gururangan et al., 2018; McCoy et al., 2019) observed that specific linguistic phenomena or syntactic heuristics are highly correlated with certain inference classes in natural language inference (NLI). Jia and Liang (2017) revealed that the question-answering (QA) models trained on SQuAD were not robust to perturbations with modified semantics since the QA models cannot possess the true text understanding. Recently, there are several efforts made to reduce the spurious correlations

by removing the data bias explicitly. For example, CoQA (Reddy et al., 2019) limited the question annotation process by avoiding using exact words in the passage. SWAG (Zellers et al., 2018) utilized an adversarial filter methodology to construct the debiased dataset. In addition, several studies focused on recognizing these spurious correlations and then removing them implicitly. Clark et al. (2019) proposed a two-stage training procedure, which built a bias-only strategy to train a robust model through the ensembling approach. Sagawa et al. (2020) coupled the distributionally robust optimization with regularization to improve the worst group generalization. To the best of our knowledge, we are the first to reduce the spurious correlations for answer selection, leveraging feature decorrelation and language debiasing.

3 Methodology

We assume there are N instances (question-answer pairs) in the training set. Given a question q_i and a set of K candidate answers $A_i = \{a_1, a_2, ..., a_K\}$, the answer selection task aims to find the best answer by ranking the candidate answers based on their relevance to the given question. Benefiting from the pairwise ranking, we can reformalize the answer selection as a classification problem by predicting the relevance label y_i of each question-answer pair (q_i, a_i). We represent each question q_i and answer a_i as $q = [w_1^{q_i}, \ldots, w_n^{q_i}]$ and $a_i = [w_1^{a_i}, \ldots, w_m^{a_i}]$, where n and m are the lengths of question q_i and answer a_i, respectively.

In this paper, we propose a novel SCAN method for answer selection. As illustrated in Figure 1, the proposed SCAN consists of two primary modules: the feature decorrelation module based on sample weighting and the feature debiasing module based on contrastive learning (Chuang et al., 2020; Liu et al., 2021). Next, we will introduce the base context encoder and two key spurious correlation reduction components in detail.

3.1 Base Context Encoder

Inspired by the remarkable success of pre-trained language models (PLMs) on most NLP tasks, we employ RoBERTa (Liu et al., 2019) as our base context encoder to obtain the contextual representations of each question-answer pair.

We take the concatenation of the question q_i and each candidate answer a_i as input, and use RoBERTa to generate the contextual representation of the i-th question-answer pair as:

$$\mathbf{E}_i = \text{RoBERTa}([\text{cls}, q_i, \text{sep}, \text{sep}, a_i, \text{sep}]) \quad (1)$$

where $\mathbf{E}_i \in \mathbb{R}^{(n+m+4)\times d_h}$ denotes the hidden states of the question-answer pair (q_i, a_i) and d_h is the dimension of each hidden state. The special tokens [cls] and [sep] represent the classification token and the separation token respectively. We denote the hidden vector of the special [cls] token as $\mathbf{H}_i \in \mathbb{R}^{d_h}$, which can be treated as the base contextual representation of the question-answer pair (q_i, a_i) for prediction.

3.2 Feature Decorrelation with Sample Weighting from Sample Perspective

Spurious correlations are very common in deep models, especially when the answer selection model is overparameterized. Spurious correlations could hurt the stability and generality of the model when deployed in practice. In this paper, we employ the feature decorrelation method with sample weighting to decorrelate the relevant and irrelevant features, and make the model focus on discriminative features (relevant features) that are truly related to the label prediction.

Given the training data with N question-answer pairs, the representations learned by the base context encoder can be denoted as $\mathbf{H} \in \mathbb{R}^{N\times d_h}$. We input the representation $\mathbf{H}$ into the feature decorrelation module based on Random Fourier Features (Rahimi and Recht, 2007), which learns a weight for each instance such that features are decorrelated on the weighted training data.

Formally, we use $\mathbf{w} \in \mathbb{R}^N$ to denote the local weights of individual samples, which are initialized with all-ones vector at the beginning of each training iteration. During the optimization process with stochastic gradient descent (SGD), there are merely part of samples being observed in each batch, while the global weights of all samples would be ignored. Thus, we leverage global weights $\mathbf{w}^G \in \mathbb{R}^N$ and global features $\mathbf{H}^G \in \mathbb{R}^{N\times d_h}$ to exploit the global information of the training data. By concatenating the global and local information, we can obtain the combined features $\mathbf{H}^{\text{com}}$ and weights $\mathbf{w}^{\text{com}}$ as:

$$\begin{aligned} \mathbf{H}^{\text{com}} &= \text{Concat}(\mathbf{H}^G, \mathbf{H}) \\ \mathbf{w}^{\text{com}} &= \text{Concat}(\mathbf{w}^G, \mathbf{w}) \end{aligned} \quad (2)$$

We denote the combined features of the i-th sample as $\mathbf{H}_i^{\text{com}}$. The j-th feature in the combined representation space is denoted as $\mathbf{H}_{:,j}^{\text{com}}$.

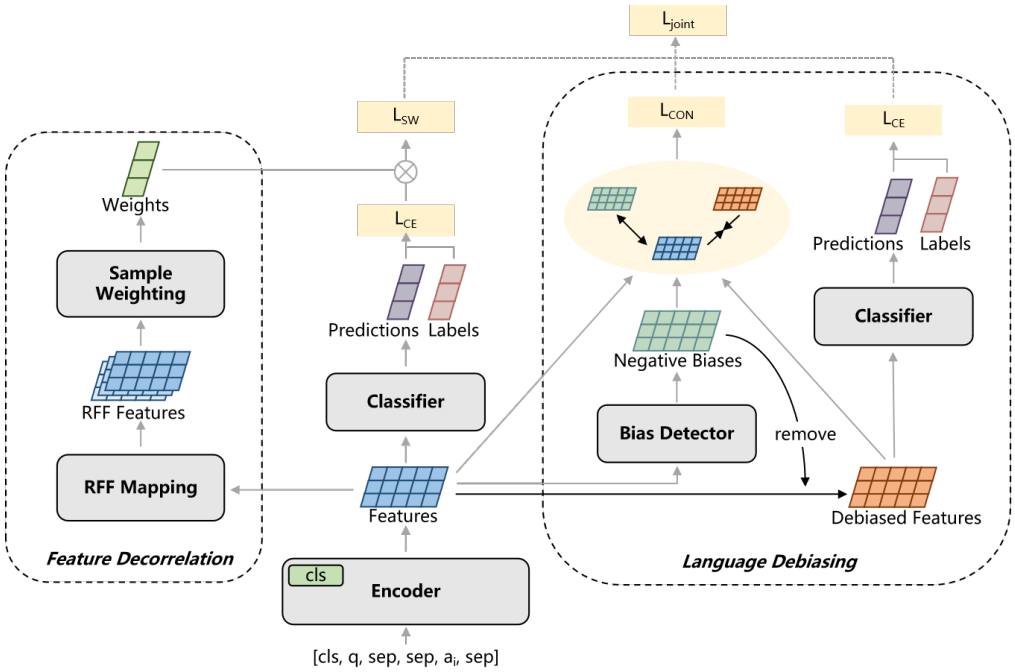

Figure 1: The overall architecture of our SCAN method, which reduces the spurious correlations with a feature decorrelation module module and a feature debiasing module.

To eliminate the correlation between features, we measure their independence via Hilbert-Schmidt Independence Criterion (HSIC) which is a kernel statistical test of independence (Gretton et al., 2007), inspired by (Zhang et al., 2021a). To reduce the computational complexity, we approximate the test statistical independence by Frobenius norm. In particular, we sample r Random Fourier Features(RFF) mapping functions from the function sapce $\mathcal{G}$ respectively, and then convert the combined representations $\mathbf{H}^{\text{com}}$ into $\widetilde{\mathbf{H}} \in \mathbb{R}^{N \times d_h \times r}$:

$$\widetilde{\mathbf{H}}_{i,j} = \left(g_1(\mathbf{H}^{\text{com}}_{i,j}), \ldots, g_r(\mathbf{H}^{\text{com}}_{i,j})\right) \quad (3)$$

$$\text{where} \quad g_k(\mathbf{H}_{i,j}) \in \mathcal{G}, \forall k, \quad (4)$$

$$\mathcal{G} = \{g : x \rightarrow \sqrt{2}\cos(\omega x + \phi) | \quad (5)$$
$$\omega \sim \mathcal{N}(0,1), \phi \sim \text{Uniform}(0, 2\pi)\}$$

where ω is sampled from the standard Normal distribution and ϕ is sampled from the Uniform distribution to approximate continuous shift-invariant kernels*. With the sample weights $\mathbf{w}^{\text{com}}$, we can calculate the weighted partial cross-covariance matrix $\hat{\sum}_{j_1,j_2}$ of two features $\mathbf{H}^{\text{com}}_{:,j_1}$ and $\mathbf{H}^{\text{com}}_{:,j_2}$ by:

$$\hat{\sum}_{j_1,j_2} = \frac{1}{N-1} \cdot \sum_{i=1}^{N} \left[\left(w_i^{\text{com}} \widetilde{\mathbf{H}}_{i,j_1} - \mathbb{E}(\widetilde{\mathbf{H}}_{:,j_1})\right)^T \cdot \left(w_i^{\text{com}} \widetilde{\mathbf{H}}_{i,j_2} - \mathbb{E}(\widetilde{\mathbf{H}}_{:,j_2})\right)\right] \quad (6)$$

$$\text{where} \quad \mathbb{E}(\widetilde{\mathbf{H}}_{:,j}) = \frac{1}{N} \sum_{i=1}^{N} w_i^{\text{com}} \widetilde{\mathbf{H}}_{i,j} \quad (7)$$

where w_i^{com} is the weight of the i-th question-answer pair (q_i, a_i).

We use the squared Frobenius norm of the partial cross-covariance matrix to estimate the independence between any pair of features. Thus, we optimize the sample weights $\mathbf{w}^{\text{com}}$ by minimizing the squared Frobenius norm between any pair of features, which can be defined as follows:

$$\mathbf{w}^{\text{com}*} = \underset{\mathbf{w}^{\text{com}}}{\text{argmin}} \sum_{1 \le j_1 < j_2 \le d_h} \left\| \hat{\sum}_{j_1,j_2} \right\|_F^2 \quad (8)$$

During the procedure of learning the weights $\mathbf{w}^{\text{com}}$, we keep the model parameters fixed.

With the learned weights $\mathbf{w}^{\text{com}}$ of question-answer pairs during the training phase, we can optimize the model parameters by minimizing the weighted cross-entropy loss function as:

$$\mathcal{L}_{\text{SW}} = -\sum_{i=1}^{N} w_i^{\text{com}} \mathbf{y}_i \log \hat{\mathbf{y}}_i \quad (9)$$

*Similar to (Zhang et al., 2021a), we adopt both sin and cosine functions to learn better features.

where $\mathbf{y}_i$ denotes the one-hot vector of the ground-truth relevance label y_i of the i-th question-answer pair (q_i, a_i). $\hat{\mathbf{y}}_i$ is the predicted relevance label of (q_i, a_i), which is defined as:

$$\hat{\mathbf{y}}_i = softmax(\mathbf{H}_i) \tag{10}$$

During the procedure of updating the model parameters via back propagation, we keep the weights of training samples fixed.

Note that for efficient optimization the weights of training samples and the model parameters are learned iteratively, and the training procedure is repeated until convergence. At the end of each training iteration, we update the global features $\mathbf{H}^G$ and the corresponding weights $\mathbf{w}^G$ as:

$$\begin{aligned} \mathbf{w}^G &= \alpha \mathbf{w}^G + (1-\alpha)\mathbf{w} \\ \mathbf{H}^G &= \alpha \mathbf{H}^G + (1-\alpha)\mathbf{H} \end{aligned} \tag{11}$$

where α denotes the hyperparameter for controlling the impact of global information.

3.3 Language Debiasing with Contrastive Learning from Feature Perspective

Most previous AS models frequently follow the superficial correlations (i.e., language bias) induced by the training data, which is another kind of the spurious correlation. The language biases makes the neural AS models brittle to linguistic variations in questions/answers. However, not all the language biases are harmful in answer selection, and some language biases may contain commonsense knowledge that is beneficial for answer selection. For example, when a question begins with "*When*", the corresponding answer should contain words that indicate time or period. In this section, we propose a feature debiasing module with contrastive learning, which weakens the negative biases in language and improves the robustness of AS models.

First, we utilize a bias detection method to recognize the negative biases that existed in the base contextual representation $\mathbf{H}_i$ of the i-th question-answer pair learned by the base context encoder. The detection function $\sigma(\cdot)$ consists of dense layer followed by a sigmoid activation function. Formally, we learn the bias weight vector $\mathbf{b}_i$ as follows:

$$\mathbf{b}_i = \sigma(\mathbf{H}_i^{\text{trans}}), \text{ where } \mathbf{H}_i^{\text{trans}} = \rho_b(\mathbf{H}_i) \tag{12}$$

where ρ_b denotes a multi-layer perceptron (MLP). $\mathbf{H}_i^{\text{trans}}$ represents the transformed feature containing language biases.

Second, we can learn the negative bias representation $\mathbf{H}_i^{\text{bias}}$ based on the base contextual representation $\mathbf{H}_i$ by the product of the bias weight vector $\mathbf{b}_i$ and transformed feature $\mathbf{H}_i^{\text{trans}}$ as:

$$\mathbf{H}_i^{\text{bias}} = \mathbf{b}_i \cdot \mathbf{H}_i^{\text{trans}} \tag{13}$$

Then, we learn the debiased representation $\mathbf{H}_i^d \in \mathbb{R}^{d_h}$ by removing the negative bias representation $\mathbf{H}_i^{\text{bias}}$ from the original base contextual representation $\mathbf{H}_i$. We compute the debiased representation $\mathbf{H}_i^{\text{debias}}$ as follows:

$$\mathbf{H}_i^{\text{d}} = \rho_d(\mathbf{H}_i - \mathbf{H}_i^{\text{bias}}) \tag{14}$$

where ρ_d denotes another MLP layer.

Cross-Entropy Loss The learned debiased representation $\mathbf{H}_i^{\text{d}}$ of the i-th question-answer pair is fed into a classifier with a softmax layer as:

$$\hat{\mathbf{y}}_i = softmax(\mathbf{H}_i^{\text{d}}) \tag{15}$$

where $\hat{\mathbf{y}}_i$ represents the predicted relevance label of (q_i, a_i). We can optimize the answer selection model by minimizing the cross-entropy loss as:

$$\mathcal{L}_{\text{CE}} = -\sum_{i=1}^{N} \mathbf{y}_i \log \hat{\mathbf{y}}_i \tag{16}$$

where $\mathbf{y}_i$ denotes the one-hot vector of the ground-truth relevance label y_i of the question-answer pair (q_i, a_i). N is the number of training instances.

Contrastive Loss To avoid using additional parameters in inference phase, we attempt to discard the language debiasing module in for inference and make the base context features $\mathbf{H}$ and debiased features $\mathbf{H}^{\text{d}}$ learned by the language debaising module as similar as possible. In this paper, we leverage the contrastive learning to learn robust representations by incorporating instance-level semantic discriminativeness into the representation learning. Concretely, we leverage a contrastive loss function $\mathcal{L}_{\text{CL}}$ to make each base context representation $\mathbf{H}_i$ close to the corresponding debiased feature $\mathbf{H}_i^{\text{d}}$ and away from the negative bias feature $\mathbf{H}_i^{\text{bias}}$. Formally, we define the contrastive loss as:

$$\mathcal{L}_{\text{CL}} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\mu(\mathbf{H}_i, \mathbf{H}_i^{\text{d}})}{\mu(\mathbf{H}_i, \mathbf{H}_i^{\text{d}}) + \mu(\mathbf{H}_i, \mathbf{H}_i^{\text{bias}})} \tag{17}$$

$$\mu(\mathbf{H}_i, \mathbf{H}_i^{\text{bias}}) = \exp(sim(\mathbf{H}_i, \mathbf{H}_i^{\text{bias}})/\tau) \tag{18}$$

where $sim()$ denotes a cosine similarity function. τ is a temperature value.

3.4 Joint Training Objective

Overall, our method consists of three training objectives, including the sample weighting loss $\mathcal{L}_{\text{SW}}$, cross-entropy loss $\mathcal{L}_{\text{CE}}$, and the contrastive loss $\mathcal{L}_{\text{CL}}$. We minimize the joint loss function $\mathcal{L}_{\text{joint}}$ by summing up the three training objectives as:

$$\mathcal{L}_{\text{joint}} = \mathcal{L}_{\text{SW}} + \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{CL}} \quad (19)$$

Although the sample weights are optimized according to Eq. (8) during the training process, we do not include the weight optimization function during the overall training objective for optimizing the model parameters.

3.5 Inference Stage

In the inference phase, given the back propagation is disabled, we escape the sample weighting phase without any calculation of sample weights and discard the language debiasing phase without introducing additional parameters. Instead, we conduct the prediction directly via Eq. (10) by merely leveraging the optimized base RoBERTa encoder.

4 Experimental Setup

4.1 Datasets

To evaluate the effectiveness of our method, we conduct comprehensive experiments on three publicly available corpora. The statistics of the three datasets are shown in Table 1.

WikiQA The WikiQA dataset (Yang et al., 2015) is an open-domain question answering dataset. The original WikiQA contains 3047 questions originally sampled from Bing query logs and 29258 answer sentences from Wikipedia. We denote the questions that have no correct answer sentences as "All-" and the questions that have only correct answer sentences as "*All+*". The remaining data set without both "*All-*" and "*All+*" questions is denoted as "*Clean*". Following the previous works (Garg et al., 2020), we train the AS models on the no "*All-*" questions, and then test the models on the "*Clean*" questions. The statistics of WikiQA are shown in Table 1.

SelQA The SelQA (Jurczyk et al., 2016) dataset is similar to WikiQA but covers more diverse topics drawn from Wikipedia. It consists of a larger number of questions, which is about 6 times larger than WikiQA. We adopt the original data split as in (Jurczyk et al., 2016) to verify the AS models. The statistics of SelQA are shown in Table 1.

Dataset		Train	Dev	Test
WikiQA	#Q	873	122	237
	#A	8672	1126	2341
SelQA	#Q	5529	785	1590
	#A	66438	9377	19435
ANTIQUE	#Q	2226	200	200
	#A	25229	2193	6589

Table 1: Statistics of the three experimental datasets.

ANTIQUE The ANTIQUE dataset (Hashemi et al., 2020) is an open-domain non-factoid QA dataset collected from a community question answering service, Yahoo!Answers. Different from WikiQA and SelQA, ANTIQUE has four-level relevance labels between 1 to 4. Following previous work (MacAvaney et al., 2020), we regard scores 3 and 4 as relevant, while scores 1 and 2 are treated as irrelevant. Since the original dataset has no validation set, we choose 200 questions from the training set as a held-out set for validation, similar to MacAvaney et al. (2020). The statistics of ANTIQUE are shown in Table 1.

4.2 Baselines

For WikiQA and SelQA which are widely used in previous works, we compare the proposed SCAN with several advanced baselines, including CNN-DAN (Santos et al., 2017), CNN-hinge (Santos et al., 2017), ACNN (Shen et al., 2017a), AdaQA (Shen et al., 2017a), HyperQA (Tay et al., 2018b), DRCN (Kim et al., 2019), RE2 (Yang et al., 2019b), a compare aggregate model (CompAgg) (Yoon et al., 2019), BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019), TANDA (Garg et al., 2020), answer support-based reranker (ASR) (Zhang et al., 2021b), DAR and DAR-DPR Zhang et al. (2022). For ANTIQUE, we compare SCAN with four benchmark baselines including aNMM (Yang et al., 2016), BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019), a curriculum learning method with BERT (BERT-CL) (MacAvaney et al., 2020), a bilateral generation method (BERT-BiG) (Deng et al., 2021), and TANDA (Garg et al., 2020).

4.3 Implementation Details

We adopt the RoBERTa-base (Liu et al., 2019) that is pre-trained on large-scale English corpus and fine-tuned on ASNQ corpus (Garg et al., 2020) as the sentence encoder. In the experiments, we apply the grid search algorithm (Huang et al., 2012)

on the validation set to tune the hyper-parameters. Concretely, we set the maximum sequence length to 128. The training batch size is set to 140. The dimension of hidden state (i.e., d_h) is set to 768. We adopt the Adam optimizer to optimize the whole SCAN method. The temperature value τ, learning rate and α for decorrelation module are 1/0.3/0.5, 1e-4/1e-4/1e-3 and 0.7/0.9/0.9 for WikiQA, SelQA and ANTIQUE, respectively. For reproducibility, we will release our code and data upon the publication of this paper.

4.4 Evaluation Metrics

For WikiQA and SelQA, we measure our method on test set with three official metrics: Mean Average Precision (MAP), Mean Reciprocal Rank (MRR), Precision at 1 of ranked candidates (P@1). For ANTIQUE, we also measure the MAP, MRR and P@1. In addition, we compute Normalized Discounted Cumulative Gain (i.e., nDCG@1, nDCG@3 and nDCG@10) with the original four-level relevance labels.

5 Experimental results

5.1 Overall Performance

Tables 2-4 summarize the experimental results on WikiQA, SelQA and ANTIQUE, respectively. SCAN performs significantly and consistently better than the compared baselines on all the three datasets, verifying the effectiveness of our SCAN method. From Table 2 and Table 3, we can observe that the CNN- or RNN-based methods perform poorly because they do not take advantage of the pre-trained language models (PLMs). TANDA outperforms BERT and CNN-based methods by adopting the RoBERTa-base that is pre-trained on large-scale general corpus and fine-tuned on ASNQ corpus. ASR, DAR and DAR-DRP, which are based on TANDA, improve the performance of TANDA by exploiting the interrelated information between the target answer and the other candidate answers. SCAN takes a further step towards reducing the spurious correlations for answer selection by feature decorrelation and language debiasing.

Table 4 reports the experimental results on the ANTIQUE dataset, demonstrating that the proposed SCAN method is also effective on the non-factoid QA. Specifically, SCAN exceeds the TANDA model (the base model of SCAN) by 3.97% on MRR and 6% on P@1. This verifies that it is necessary to remove the spurious correlations between the text representations and the prediction relevance labels.

Method	MAP	MRR	P@1
HyperQA ♮	0.7120	0.7270	-
RE2 ♯	0.7452	0.7618	-
Comp-Agg †	0.7640	0.7840	-
Comp-Agg (QNLI) †	0.8340	0.8480	-
BERT †	0.8130	0.8280	-
RoBERTa	0.8441	0.8551	0.7553
TANDA †	0.8890	0.9010	-
TANDA-re ‡	0.8860	0.8983	0.8189
ASR ‡	0.9014	0.9123	0.8436
DAR ‡	0.9011	0.9136	0.8519
DAR-DRP ‡	0.9051	0.9164	0.8560
SCAN	**0.9164**	**0.9281**	**0.8776**

Table 2: Experimental Results on WikiQA. The results with ♮ are retrieved from (Tay et al., 2018b), with ♯ are retrieved from (Yang et al., 2019b), with † are retrieved from (Garg et al., 2020), with ‡ are retrieved from (Zhang et al., 2022). TANDA-re denotes a reimplementation of TANDA. The best scores are in bold.

Method	MAP	MRR	P@1
CNN-DAN ♮	0.8660	0.8730	-
CNN-hinge ♮	0.8760	0.8810	-
ACNN ♮	0.8740	0.8800	-
AdaQA ♮	0.8910	0.8980	-
DRCN ♮	0.9250	0.9300	-
TANDA-re ‡	0.9512	0.9587	0.9302
ASR ‡	0.9519	0.9592	0.9314
DAR ‡	0.9592	0.9653	0.9415
SCAN	**0.9641**	**0.9701**	**0.9484**

Table 3: Experimental Results on SelQA. The results with ♮ are retrieved from (Kim et al., 2019), with ‡ are retrieved from (Zhang et al., 2022).

5.2 Ablation study

To verify the effectiveness of feature decorrelation and language debiasing in SCAN, we perform ablation test of SCAN on two types of QA corpora (WikiQA and ANTIQUE) in terms of removing the feature decorrelation module (denoted as w/o FD) and language debiasing (denoted as w/o LD), respectively. In particular, for the w/o FD model, the weighted cross-entropy loss is replaced with a normal cross-entropy loss without considering sample weights. We also report the results of removing both feature decorrelation and language debiasing (w/o FD+LD).

The ablation test results are reported in Table 5. Generally, both feature decorrelation and language

Method	MAP	MRR	P@1	nDCG@1	nDCG@3	nDCG@10
aNMM ♮	0.2563	0.6250	0.4847	0.5289	0.5127	0.4904
BERT ♮	0.3771	0.7968	0.7092	0.7126	0.6570	0.6423
RoBERTa	0.6137	0.7763	0.6550	0.6683	0.6525	0.6765
BERT-CL ♯	-	0.7335	0.6450	-	-	-
BERT-BiG †	-	0.8470	0.7650	0.7500	0.7100	0.7200
TANDA	0.6511	0.8258	0.7250	0.7167	0.6969	0.7091
SCAN	**0.6722**	**0.8637**	**0.7850**	**0.7550**	**0.7186**	**0.7297**

Table 4: Experimental Results on ANTIQUE. The results with ♮ are retrieved from (Hashemi et al., 2020), with ♯ are retrieved from (MacAvaney et al., 2020), with † are retrieved from (Deng et al., 2021).

Method	WikiQA			ANTIQUE		
	MAP	MRR	P@1	MAP	MRR	P@1
SCAN	0.9164	0.9281	0.8776	0.6722	0.8637	0.7850
w/o FD	0.9004	0.9148	0.8523	0.6667	0.8399	0.7400
w/o LD	0.9011	0.9144	0.8523	0.6603	0.8424	0.7500
w/o FD+LD	0.8943	0.9063	0.8354	0.6511	0.8258	0.7250

Table 5: Experimental Results of the ablation study on WikiQA and ANTIQUE.

debiasing contribute noticeable improvement to the proposed SCAN method. Concretely, the performances decrease sharply, especially in terms of MRR and $P@1$, when removing either the FD model or the LD module. This is within our expectation since both feature decorrelation and language debiasing can reduce the spurious correlations for answer selection.

5.3 Robustness to Noise and Perturbation

To further analyze the robustness of our method, we conduct experiments on the WikiQA dataset with injected noise and adversarial perturbations. Following previous work (Gokhale et al., 2022), we create adversarial samples by adding character-level perturbations such as swapping, inserting or deleting characters to 30% of samples. In addition, similar to (Garg et al., 2020), we inject noise into the training samples in WikiQA by randomly sampling 20% of question-answer pairs from the training set and switching their labels. The experimental results are shown in Table 6. SCAN achieves consistently better performance than TANDA on these settings, verifying the robustness of our method to noise and adversarial perturbations.

5.4 Case Study

We use a representative exemplary case that is selected from the WikiQA test set to further investigate the effectiveness of SCAN. This chosen question is incorrectly predicted by TANDA while being correctly predicted by SCAN. From Table 7, We observe that TANDA simply picks up the answer that contains the matching words *cricket wireless* without understanding the deep semantics. On the contrary, SCAN obtains the correct answer since it can recognize the real intention of the question. Another example shown in Tabel 8 is from ANTIQUE, where the topic of answers is more diverse than factoid QA dataset. While TANDA predicts an answer with superficial relation with the question, our model make a more precise prediction without the disturbance of spurious relation. These examples demonstrates that our model can focus on the true correlation between the question and the answer, which is critical when the candidate answers contain misleading information.

Method	MAP	MRR	P@1
TANDA	0.8943	0.9063	0.8354
TANDA-Perturb	0.8898	0.9019	0.8270
TANDA-Noise	0.8740	0.8871	0.8059
SCAN	0.9164	0.9281	0.8776
SCAN-Perturb	0.9111	0.9228	0.8692
SCAN-Noise	0.8907	0.9055	0.8439

Table 6: Performance comparison when noise and perturbations are injected into WikiQA.

5.5 Error Analysis

Although our SCAN model achieves better performance than previous models, it still fails to handle

Question: "what company is cricket wireless by?"
Predicted by TANDA: "Cricket Communications, Inc., (d.b.a. Cricket Wireless) founded in 1999, provides wireless services to over 7 million customers in the United States." (***incorrect answer***)
Predicted by SCAN: "The company is a subsidiary of Leap Wireless, utilizing its CDMA 1X, 1xEV-DO and LTE networks." (***correct answer***)

Table 7: A question from WikiQA with the answers predicted by TANDA and SCAN, respectively.

Question: "what are some easy ways to get a toddler to go to sleep without being mean?"
Predicted by TANDA: "There are plenty of ways.... - The most obvious is try to sleep. - Take some Pepto-bismol. - Have a piece of peppermint (that's good too even if you still get sick)." (***incorrect answer***)
Predicted by SCAN: "Let them play for a while. Also, play with them. That way they'll feel like you care about them. Also, try laying down with them. That used to help my son, who now is 3 years old." (***correct answer***)

Table 8: A question from ANTIQUE with the answers predicted by TANDA and SCAN, respectively.

some cases. To investigate the limitations of SCAN, we analyze the bad cases produced by SCAN. We summarize the several reasons for obtaining the incorrect predictions. First, SCAN fails to tackle some questions that require commonsense knowledge to reason correct answers. In particular, some questions and the corresponding answers have different expressions for the same entities, thus our method struggles to capture the relations of the question-answer pairs based on the contextual representations only. One possible solution is to leverage knowledge bases to facilitate the reasoning process. Second, there are some noises (confused candidates) existing in the datasets. For example, the question *"Where was the first ski flying hill built?"* has two candidate answers *"Nevertheless the first-ever ski flying hill was built in Planica, Slovenia"* and *"The first ski flying hill was built in Planica in Slovenia"* with the former one labeled as incorrect and the latter one labeled as correct. However, both answers convey the same meaning. We may update the datasets by carefully examining the relevance labels of candidate answers.

6 Conclusion

In this paper, we proposed a novel spurious correlation reduction method to improve the robustness of the answer selection models from the sample and feature perspectives. First, we devised a feature decorrelation module by learning a weight for each training instance to remove the feature dependencies and reduced the spurious correlations without prior knowledge of such correlations. Second, we introduced a feature debiasing module with contrastive learning to alleviate the negative language biases and improved the robustness of the AS models. We conducted extensive experiments on three benchmark datasets and the experimental results showed the effectiveness of SCAN.

Acknowledgements

This work was partially supported by National Natural Science Foundation of China (61906185, 61876053), Youth Innovation Promotion Association of CAS China (No. 2020357), Shenzhen Science and Technology Innovation Program (Grant No. KQTD20190929172835662), Shenzhen Basic Research Foundation (No. JCYJ20210324115614039 and No. JCYJ20200109113441941).

References

Ching-Yao Chuang, Joshua Robinson, Yen-Chen Lin, Antonio Torralba, and Stefanie Jegelka. 2020. Debiased contrastive learning. *Advances in neural information processing systems*, 33:8765–8775.

Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. 2019. Don't take the easy way out: ensemble based methods for avoiding known dataset biases. *EMNLP*.

Yang Deng, Wenxuan Zhang, and Wai Lam. 2021. Learning to rank question answer pairs with bilateral contrastive data augmentation. *arXiv preprint arXiv:2106.11096*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Siddhant Garg, Thuy Vu, and Alessandro Moschitti. 2020. Tanda: Transfer and adapt pre-trained transformer models for answer sentence selection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7780–7788.

Tejas Gokhale, Swaroop Mishra, Man Luo, Bhavdeep Sachdeva, and Chitta Baral. 2022. *Generalized but not Robust?* comparing the effects of data modification methods on out-of-domain generalization and adversarial robustness. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2705–2718, Dublin, Ireland. Association for Computational Linguistics.

Arthur Gretton, Kenji Fukumizu, Choon Teo, Le Song, Bernhard Schölkopf, and Alex Smola. 2007. A kernel statistical test of independence. *Advances in neural information processing systems*, 20.

Jiahui Guo, Bin Yue, Guandong Xu, Zhenglu Yang, and Jin-Mao Wei. 2017. An enhanced convolutional neural network model for answer selection. In *WWW*, pages 789–790.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R Bowman, and Noah A Smith. 2018. Annotation artifacts in natural language inference data. *arXiv preprint arXiv:1803.02324*.

Helia Hashemi, Mohammad Aliannejadi, Hamed Zamani, and W Bruce Croft. 2020. Antique: A non-factoid question answering benchmark. In *European Conference on Information Retrieval*, pages 166–173. Springer.

Wanwei He, Yinpei Dai, Min Yang, Jian Sun, Fei Huang, Luo Si, and Yongbin Li. 2022a. Unified dialog model pre-training for task-oriented dialog understanding and generation. In *SIGIR*, pages 187–200.

Wanwei He, Yinpei Dai, Yinhe Zheng, Yuchuan Wu, Zheng Cao, Dermot Liu, Peng Jiang, Min Yang, Fei Huang, Luo Si, et al. 2022b. Galaxy: A generative pre-trained model for task-oriented dialog with semi-supervised learning and explicit policy injection. In *AAAI*, volume 36, pages 10749–10757.

Qiujun Huang, Jingli Mao, and Yong Liu. 2012. An improved grid search algorithm of svr parameters optimization. In *2012 IEEE 14th International Conference on Communication Technology*, pages 1022–1026. IEEE.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. *arXiv preprint arXiv:1707.07328*.

Tomasz Jurczyk, Michael Zhai, and Jinho D Choi. 2016. Selqa: A new benchmark for selection-based question answering. In *2016 IEEE 28th International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 820–827. IEEE.

Seonhoon Kim, Inho Kang, and Nojun Kwak. 2019. Semantic sentence matching with densely-connected recurrent and co-attentive information. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 6586–6593.

Tuan Lai, Quan Hung Tran, Trung Bui, and Daisuke Kihara. 2019. A gated self-attention memory network for answer selection. *arXiv preprint arXiv:1909.09696*.

Jieke Li, Min Yang, and Chengming Li. 2021. Clc-rs: A chinese legal case retrieval system with masked language ranking. In *CIKM*, pages 4734–4738.

Xiaoyan Li. 2003. Syntactic features in question answering. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 383–384.

Che Liu, Rui Wang, Jinghua Liu, Jian Sun, Fei Huang, and Luo Si. 2021. Dialoguecse: Dialogue-based contrastive learning of sentence embeddings. *arXiv preprint arXiv:2109.12599*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Sean MacAvaney, Franco Maria Nardini, Raffaele Perego, Nicola Tonellotto, Nazli Goharian, and Ophir Frieder. 2020. Training curricula for open domain answer re-ranking. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 529–538.

R Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. *arXiv preprint arXiv:1902.01007*.

Guanglin Niu, Yang Li, Chengguang Tang, Ruiying Geng, Jian Dai, Qiao Liu, Hao Wang, Jian Sun, Fei Huang, and Luo Si. 2021. Relational learning with gated and attentive neighbor aggregator for few-shot knowledge graph completion. In *SIGIR*, pages 213–222.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Xipeng Qiu and Xuanjing Huang. 2015. Convolutional neural tensor network architecture for community-based question answering. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*.

Ali Rahimi and Benjamin Recht. 2007. Random features for large-scale kernel machines. *Advances in neural information processing systems*, 20.

Siva Reddy, Danqi Chen, and Christopher D Manning. 2019. Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. 2020. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *ICLR*.

Cicero Nogueira dos Santos, Kahini Wadhawan, and Bowen Zhou. 2017. Learning loss functions for semi-supervised learning via discriminative adversarial networks. *arXiv preprint arXiv:1707.02198*.

Apoorv Saxena, Aditay Tripathi, and Partha Talukdar. 2020. Improving multi-hop question answering over knowledge graphs using knowledge base embeddings. In *ACL*, pages 4498–4507.

Aliaksei Severyn and Alessandro Moschitti. 2015. Learning to rank short text pairs with convolutional deep neural networks. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, pages 373–382.

Dinghan Shen, Martin Min, Yitong Li, and Lawrence Carin. 2017a. Adaptive convolutional filter generation for natural language understanding.

Gehui Shen, Yunlun Yang, and Zhi-Hong Deng. 2017b. Inter-weighted alignment network for sentence pair modeling. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 1179–1189.

Ying Shen, Yang Deng, Min Yang, Yaliang Li, Nan Du, Wei Fan, and Kai Lei. 2018. Knowledge-aware attentive neural network for ranking question answer pairs. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 901–904.

Mihai Surdeanu, Massimiliano Ciaramita, and Hugo Zaragoza. 2008. Learning to rank answers on large online qa collections. In *proceedings of ACL*, pages 719–727.

Ming Tan, Cicero Dos Santos, Bing Xiang, and Bowen Zhou. 2016. Improved representation learning for question answer matching. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 464–473.

Yi Tay, Minh C Phan, Luu Anh Tuan, and Siu Cheung Hui. 2017. Learning to rank question answer pairs with holographic dual lstm architecture. In *SIGIR*, pages 695–704. ACM.

Yi Tay, Luu Anh Tuan, and Siu Cheung Hui. 2018a. Cross temporal recurrent networks for ranking question answer pairs. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Yi Tay, Luu Anh Tuan, and Siu Cheung Hui. 2018b. Hyperbolic representation learning for fast and efficient neural question answering. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 583–591.

Yi Tay, Luu Anh Tuan, and Siu Cheung Hui. 2018c. Multi-cast attention networks for retrieval-based question answering and response prediction. *arXiv preprint arXiv:1806.00778*.

Mengqiu Wang, Noah A Smith, and Teruko Mitamura. 2007. What is the jeopardy model? a quasi-synchronous grammar for qa. In *EMNLP-CoNLL*, pages 22–32.

Yuexiang Xie, Ying Shen, Yaliang Li, Min Yang, and Kai Lei. 2020. Attentive user-engaged adversarial neural network for community question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9322–9329.

Liu Yang, Qingyao Ai, Jiafeng Guo, and W Bruce Croft. 2016. anmm: Ranking short answer texts with attention-based neural matching model. In *Proceedings of the 25th ACM international on conference on information and knowledge management*, pages 287–296.

Min Yang, Lei Chen, Xiaojun Chen, Qingyao Wu, Wei Zhou, and Ying Shen. 2019a. Knowledge-enhanced hierarchical attention for community question answering with multi-task and adaptive learning. In *IJCAI*, pages 5349–5355.

Runqi Yang, Jianhai Zhang, Xing Gao, Feng Ji, and Haiqing Chen. 2019b. Simple and effective text matching with richer alignment features. *arXiv preprint arXiv:1908.00300*.

Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. WikiQA: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2013–2018, Lisbon, Portugal. Association for Computational Linguistics.

Wenpeng Yin, Hinrich Schütze, Bing Xiang, and Bowen Zhou. 2016. Abcnn: Attention-based convolutional neural network for modeling sentence pairs. *Transactions of the Association for Computational Linguistics*, 4:259–272.

Seunghyun Yoon, Franck Dernoncourt, Doo Soon Kim, Trung Bui, and Kyomin Jung. 2019. A compare-aggregate model with latent clustering for answer selection. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 2093–2096.

Chunyuan Yuan, Wei Zhou, Mingming Li, Shangwen Lv, Fuqing Zhu, Jizhong Han, and Songlin Hu. 2019. Multi-hop selector network for multi-turn response selection in retrieval-based chatbots. In *EMNLP-IJCNLP*, pages 111–120.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. Swag: A large-scale adversarial dataset for grounded commonsense inference. *arXiv preprint arXiv:1808.05326*.

Xiaodong Zhang, Sujian Li, Lei Sha, and Houfeng Wang. 2017. Attentive interactive neural networks for answer selection in community question answering. In *AAAI*, pages 3525–3531.

Xingxuan Zhang, Peng Cui, Renzhe Xu, Linjun Zhou, Yue He, and Zheyan Shen. 2021a. Deep stable learning for out-of-distribution generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5372–5382.

Zeyu Zhang, Thuy Vu, and Alessandro Moschitti. 2021b. Joint models for answer verification in question answering systems. *arXiv preprint arXiv:2107.04217*.

Zeyu Zhang, Thuy Vu, and Alessandro Moschitti. 2022. Double retrieval and ranking for accurate question answering. *arXiv preprint arXiv:2201.05981*.

Xiaoqiang Zhou, Baotian Hu, Qingcai Chen, and Xiaolong Wang. 2018. Recurrent convolutional neural network for answer selection in community question answering. *Neurocomputing*, 274:8–18.