PINGAN Omini-Sinitic at SemEval-2021 Task 4: Reading Comprehension of Abstract Meaning

Ye Wang **Yanmeng Wang** **Haijun Zhu** **Bo Zeng**
Zhenghong Hao **Shaojun Wang** **Jing Xiao**
Ping An Technology, Beijing 100191, China
{wangye430,wangyanmeng219,zhuhaijun416,zengbo345,
haozhenghong145,wangshaojun851,xiaojing661}@pingan.com.cn

Abstract

This paper describes the winning system for subtask 2 and the second-placed system for subtask 1 in SemEval 2021 Task 4: Reading Comprehension of Abstract Meaning. We propose to use pre-trained ELECTRA discriminator to choose the best abstract word from five candidates. An upper attention and auto denoising mechanism is introduced to process the long sequences. The experiment results demonstrate that this contribution greatly facilitates the contextual language modeling in reading comprehension task. The ablation study is also conducted to show the validity of our proposed methods.

1 Introduction

Reading Comprehension of Abstract Meaning (ReCAM) (Zheng et al., 2021) is a cloze-style task, which takes a document and related human written abstract with one word replaced by a placeholder as input. The model is required to choose the best word from five candidates. The ReCAM consists of three subtasks. In subtask 1 and 2, participating systems are required to choose the best imperceptible concept and hyper-nyms concepts word respectively. Subtask 3 aims to evaluate performance of a system trained on one definition and test on the other.

Traditional cloze-style reading comprehension model (SA reader) (Kadlec et al., 2016) uses attention to directly pick the answer from the context, which makes model incapable to answer the question where the answer does not appear in passage. Furthermore, GA reader (Dhingra et al., 2017) adopts multi-hop attention mechanism to build query-specific representation of answer for ranking the candidates which is not part of passage.

Pre-trained language models (Radford et al., 2018; Devlin et al., 2019; Yang et al., 2019; Lan et al., 2020) have been widely adopted for context modeling in many Natural Language Processing tasks. These models are pre-trained on huge corpora with plain texts and can better model contextual dependencies of tokens, thus enhance the performance of downstream approaches. As described in (Lai et al., 2017; Zhang et al., 2020; Chen et al., 2019; Zhu et al., 2018), they improved the performance of single-choice reading comprehension tasks by introducing pre-trained model, but they takes excessive memory for concatenating each option with the question and the passage.

GPT2 (Radford et al., 2018) has outperformed the SOTA result on cloze-style task CBT (Hill et al., 2016). As stated in (Radford et al., 2018), GPT2 computes the probability of each choice and the rest of the sentence conditioned on this choice according to the pre-trained model, the answer is the choice with highest probability. GPT2 outperforms RNN-based models without fine-tuning on CBT task, we assume that pre-trained language model has better potential to address the cloze-style problem than fine-tuning the pre-trained model with an additional ranking network.

The most relevant work to our model is Pattern-Exploiting Training (PET) (Schick and Schütze, 2020a,b), which proposed to reformulate the sentence classification task to cloze-style task with defined golden answer word as supervising signal. The comparison between PET and our proposed method is reported in section 5.

Different with PET, We propose a novel Auto Denoising Discriminator for Abstract Concept in reading comprehension (ADDAC) by fine-tuning the pre-trained discriminator of ELECTRA (Clark et al., 2020). Auto Denoising is introduced while processing long sequences. By fine-tuning the pre-trained model on its own structure with the original pre-training loss , the tasks results is significantly improved even with small train dataset, we suppose the representations stored in the pre-trained model has been maximum reserved in this way.

Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), pages 820–826
Bangkok, Thailand (online), August 5–6, 2021.

2 Background

2.1 Task Description

The task intends to answer a cloze-style question, the answer to which depends on the understanding of a context document provided with the question. The model is also provided with a set of possible answers from which the correct one is to be selected. This can be formalized as follows:

The training data consists of tuples (q, d, a, A), where q is an abstract sentence of document d and one word in q is replaced with a placeholder. A is a set of possible options and $a \in A$ is the golden answer. Both q and d are sequences of words and golden answer a does not appear in the article d.

2.2 ELECTRA vs ALBERT

Since the success of BERT (Devlin et al., 2019), pre-trained language models have adopted a large amount of parameters to achieve better modeling performance. ALBERT (Lan et al., 2020) uses factorized embedding parameterization and cross-layer parameter sharing to greatly reduce the amount of model parameters and achieved SOTA in multiple natural language understanding task. ALBERT outperforms other pre-trained language models which is trained by MLM(masked language model) in combination with PET method for this task.

ELECTRA (Clark et al., 2020) proposed the RTD (replaced token detection) task with adversarial learning as an alternative to the MLM task. A smaller generator is used to replace the special token [MASK] in training samples, and then a discriminator is trained to predict each word in the input is real or generated by generator. In section 3.1, we will elaborate the details about our proposed discriminator mechanisms based on ELECTRA. The performance comparison between ALBERT with PET and our proposed discriminator model on ReCAM task is reported in Section 5.

3 System overview

3.1 Pre-trained Discriminator

We approach the competition tasks as cloze-style task, which can be reformulated to masked language modeling (Devlin et al., 2019) problems. As shown in Figure 1, we replace placeholder with golden and negative options in question q denoted as q_A, which is further concatenated with corresponding document d in pre-trained model input

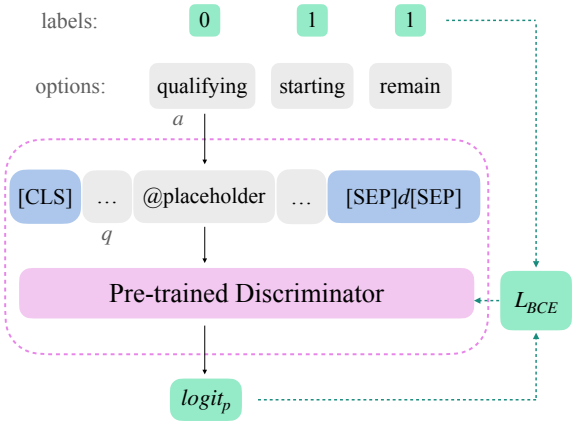

Figure 1: Discriminator overview, the placeholder in q is replaced by a which is one of candidate options, the label of golden option is 0 followed the ELECTRA pre-training setting.

style ($[CLS]q_A[SEP]d[SEP]$). We ignore the part of sequence which exceed the maximum length of input sequence . The input sequence is forward to ELECTRA discriminator and the hidden states are calculated by Equation 1

$$H_{q_A} = F([q_A; d]) \tag{1}$$

$$logit_p = D(H^p) \tag{2}$$

where F is the pre-trained 24-layer transformers and D is a linear layer which classify the hidden states of replaced word from ELECTRA Discriminator. $H_{q_A} \in \mathbb{R}^{N \times d}$ are the hidden states of input sequence, in which N and d are the maximum length of input sequence and the dimension of hidden states. Then we only use the hidden states of placeholder $H^p \in \mathbb{R}^d$ selected from H_{q_A} as the input to D. BCE (Binary Cross Entropy) loss, which measures the Binary Cross Entropy between the golden label and the output, is used for binary classification as Equation 3.

$$L_{BCE} = BCE(Sigmoid(logit_p), l_p) \tag{3}$$

where l_p is the binary label of option word (replacing placeholder in input sequence). The label is set to 0 for golden option, 1 for negative options, which is the same as ELECTRA pre-training setting (Clark et al., 2020). While in inference, the option with lowest $logit_p$ is regard as the right answer to the question. The experiment results show that discriminator outperforms the ranking and PET implementation on this task (see section 5).

We also implement a ranking model based on ELECTRA for single-choice reading comprehension task to compare with the Discriminator approach. The question and document is combined with each option as the input of ELECTRA encoder, which is denoted as $S \in \mathbb{R}^{N \times m}$, where N is the length of sequences and m is the number of options. A linear layer and a softmax layer is added after ELECTRA encoder as the ranking network, which use the hidden states of [CLS] in S as input.

3.2 Processing Long Input Sequence

Most of the pre-trained models have a limitation in processing long sequences. The maximum sequence length of ELECTRA is 512, which is much shorter than the maximum length of input sequence (i.e. concatenation of question and document). We propose two methods for processing long input sequences: 1) document is segmented to shorter passages, which leads to the problem of mislabel samples. We introduce an auto denoising mechanism to address the problem. 2) We adopt an upper attention upon transformers.

Auto Denoising The whole document d is segmented into a set of fragments $\{s_1^d, s_2^d, \cdots, s_k^d\}$ with a fixed window size. Then, these fragments combine q_a to form model input sequence. In the inference phase, the lowest predicted logit is selected from all results of fragments as Equation 4.

$$logit_p = \min_{i=1}^{k} D(F([q_A; s_i^d])) \tag{4}$$

where q with placeholder replaced by a specific option from A denote as q_A. However, this method causes the noisy-label problem. Supposed that the answer a just finds proof from fragment s_1^d, which results in other samples $(0, [q_a; s_{i \neq 1}^d])$ to be mislabeled samples, which significantly impact training of model and decrease the prediction accuracy. Therefore, we take the advantage of a noise-tolerant loss (bi-tempered logistic loss, BT (Amid et al., 2019)) and a noise detection method (over-fitting to under-fitting, O2U (Huang et al., 2019)) in this work. The BT logistic loss lowers gradient on noisy samples which relieve the negative effects on model training via adjusting bi-temperature. The O2u makes full use of the property that model is easier to forget the mislabeled samples than the clean samples, to identify and filter the mislabeled samples.

Upper Attention The long sequence of input is segmented into small segments with the length of 512 tokens and each segments are concatenated with same question to form the input sequences, which are encoded into hidden states $H_i \in R^{d \times 512}$, where i is the index of segments of passage, the d is the dimension of hidden states and 512 is the sequence length. The hidden states of placeholder is denoted as $H_i^p \in R^d$. We use a 1-layer multi-head self-attention block to fuse the hidden states of placeholder from multiple segments output.

$$H_{fuse}^p = A_g(H_1^p, H_2^p ... H_k^p) \tag{5}$$

where k is the number of segments, A_g is 1-layer multi-head-attention, without residual connection. H_{fuse}^p is applied in Equation 2 and Equation 3 for training.

3.3 Optimizer

The ELECTRA-large which has large amount of parameters tends to over-fit on small training dataset. We utilize RecAdam (Chen et al., 2020) to fine-tune the pre-trained model to address the over-fitting problem. RecAdam optimizer is proposed to address the catastrophic forgetting problem of sequential transfer learning paradigm by introducing a recall and learn mechanism into Adam optimizer, which maintain the learned knowledge in pre-trained model while learning a new task.

3.4 Data augmentation

To further boost the performance of our proposed model, we conduct data augmentation.We random pick 3000 articles in CNN/DailyMail (Hermann et al., 2015), and craw 824 latest articles from BBC news website. The CNN news is much longer than the training samples, while the length of BBC news is approximately same. The extra training samples are generated in following steps: 1) The title of news article is used as abstract. 2) We pick one most meaningful word from abstract by TF-IDF scores and the origin word is used as golden option. 3) We use words with same category of POS (Part Of Speech) from other documents as negative options. We train models on the extra to dataset as warming up and further train the models with training dataset. Unfortunately, extra training data did not effectively improve the performance of our model on this task. We just use extra data in models ensemble phase.

4 Experimental setup

4.1 Data

We use the official released dataset of SemEval2021 Task4 for experiments. The dataset of subtask 1 contains 3227/837/2025 samples for train, dev and test data. The subtask 2 dataset contains 3318/851/2017 samples for train, dev and test data. The maximum/mean length of subtask 1 and subtask 2 training data are 2275/374.34 and 2274/578.31, respectively. The statistics of sample length shows that length of article in subtask 2 is much longer than the length in subtask 1. The rate of the sequences exceeding 512 tokens is 13.95% in subtask 1, 42.46% in subtask 2, which may cause that the methods for processing long sequences are more effective in subtask 2.

Since the provided dataset is small, we apply data augmentation in model ensemble to further improve generalization of the model (see Section 3.4).

4.2 Parameter settings

Our implementation is based on the Pytorch framework for transformer-based models(Wolf et al., 2020). We trained our model based on the pre-trained ELECTRA-Large discriminator, and adopt the same model structure for subtask 1 and subtask 2. We use Adam optimizer with a learning rate of 1e-5, batch-size of 32 to train the baseline model, which is actually the ELECTRA discriminator. The max sequence length is 512 and the epoch of training is set to 5. To address the overfitting problem, we apply RecAdam with sigmoid annealing function, where the annealing rate is 0.01 and the annealing time-step is 500. The coefficient of the quadratic penalty is set to be 5,0000. For Auto Denoising, the two temperatures and label smoothing of BT are equal to 0.9, 1.5 and 0.1 respectively. The maximum learning rate, minimum learning rate and epochs in cyclical round about O2U are set to be 5e-5, 1e-6 and 5.0 respectively.

Since only 5 submissions are permitted in submitting phase, we trained multiple models under different settings for model ensemble. We also adopt 8-fold cross-validation training to improve the model generalization.

4.3 Ensemble

Two strategies are used for our final submissions on test data: 1) we ensemble all 8 models from 8-fold cross-validation training by averaging their outputs, which is trained on the train data of each subtask; 2) we trained multiple models on the train data and the augmented data with different model structures. 7 top different models are selected based on the dev accuracy for models ensemble, then average their outputs as the final output. Moreover, the model trained in cross-validation is Discriminator with Auto Denoising and RecAdam optimizer for subtask 2, and Discriminator with RecAdam for subtask 1. While in top ensemble, the techniques of Discriminator, Auto Denoising, RecAdam and Upper Attention are applied in different models.

5 Results and Analysis

5.1 Single Model Performance

We implement two baseline models for comparison with our proposed method. The ALBERT PET is the combination of PET method (Schick and Schütze, 2020a,b) and ALBERT-xxlarge model, which is trained by the MLM. The golden answer of training dataset is used as the target for MLM, but the negative ones are omitted in training. The ELECTRA Rank reformulates the cloze-style task to single-choice task, which is described in section 3.1.

Label accuracy is the official metric of the tasks and Table 1 shows the results on development dataset. Task3(1-2) is the subtask 3 which is trained on subtask 1 dataset and test on subtask 2, Task3(2-1) means train on subtask 2 and test on subtask 1. It is obvious that our proposed ELECTRA Discriminator significantly improve the performance of all tasks and outperform the best baseline by 4.7%/1.16%/3.76%/8.85% in subtask1/subtask2/subtask3(1-2)/subtask3(2-1) respectively. This confirms our hypothesis that pre-trained language model has more potential for cloze-style task. The PET with ALBERT is not suitable for this task because it can not utilize the negative options. The ELECTRA Rank performance is unsatisfactory, suggesting that fine-tuning on ranking network damage the knowledge stored in the pre-trained model.

The results of ablation study are also reported in Table 1. Disc (Discriminator) with RecAdam achieves further improvement in all tasks by 0.25% / 0.53% / 0.25% / 0.26% in subtask1 / subtask2 / subtask3(1-2) / subtask3(2-1) respectively, which prove the RecAdam optimizer is more effective for pre-trained model and also promotes the model generalization. Disc+Upper,

Table 1: Single Model Performance on dev dataset, the ablation study is demonstrated below.

Method	Task1	Task3(1-2)	Task2	Task3(2-1)
ALBERT PET	89.25	83.31	90.48	82.80
ELECTRA Rank	88.53	88.13	90.24	83.75
ELECTRA Discriminator	93.42	90.71	93.88	91.16
+RecAdam	**93.66**	91.19	94.12	91.40
+Upper	93.54	89.54	**94.35**	91.39
+Upper+RecAdam	93.31	**91.42**	94.12	91.51
+AutoDenoise+RecAdam	93.55	**91.42**	94.01	**91.88**

Table 2: Ensemble Performance on dev and test dataset, where 8-fold is the models from 8-fold cross-validation training, top ensemble means that ensemble the models with top dev accuracy.

	Method	Task1	Task3(1-2)	Task2	Task3(2-1)
Dev set	8-fold ensemble	92.47	-	92.94	-
	top ensemble	94.50	-	95.53	-
Test set	8-fold ensemble	**93.04**	93.90	94.99	91.35
	top ensemble	92.74	94.19	**95.29**	91.65

Disc+AutoDenoise+RecAdam achieve the best results in task2 and task3(2-1) respectively. This prove the validity of the two methods we proposed for long sequences. The Upper Attention achieve the best result on task2, while AutoDenoise achieve the best result on task3. We ensemble them to produce a stronger system. The Upper and AutoDenoise do not effectively improve the baseline in subtask 1, since the mean length of subtask 1 is 374.34 which is much shorter than the max sequence length of original pre-trained model.

5.2 Ensemble Performance

The performances of ensemble models are shown on Table 2, which is obtained from the competition leader-board. Our system got the first place in subtask 2 and the second place in subtask 1. For the subtask 3, our ranking is first place in subtask3(2-1), and second place in subtask3(1-2), that indicates our system has strong transfer capability in abstractive reading comprehension tasks. Ensemble results on dev dataset are exhibited for comparison, and we do not experiment model ensemble on dev data of subtask 3.

5.3 Memory-Efficiency

We trained all compared models on 16GB Tesla-V100 GPU, except for ELECTRA Rank which takes 17GB with batch size set to 1. Our proposed Discriminator only take 9GB, since one option with article and question is considered as single training sample for binary classification task. In contrast, ELECTRA Rank is required to encode the input sequence with different options at once to learn the ranking function between positive and negative options. ELECTRA Discriminator is much faster than ELECTRA Rank to converge. The epoch of training ELECTRA Discriminator is less than 5 and ELECTRA Rank needs at least 10 train epochs to converge.

Conclusion

In this paper, we propose an effective framework of combining ELECTRA discriminator with denoising learning method to boost the performance of cloze-style reading comprehension task. Our proposed model outperforms all others participating system on subtask 2 and gets the second-placed on subtask 1. We have conducted an ablation study, demonstrating the validity of Discriminator, Upper attention and Auto denoising. Pre-trained models have made great performance gain compared to traditional neural network models in many natural language tasks and is able to build comprehensive hidden representation of input text. The above experiment results may suggest that the current pre-trained model mechanism still has room for improvement.

References

Ehsan Amid, Manfred K. Warmuth, Rohan Anil, and Tomer Koren. 2019. Robust bi-tempered logistic loss based on bregman divergences. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 14987–14996.

Sanyuan Chen, Yutai Hou, Yiming Cui, Wanxiang Che, Ting Liu, and Xiangzhan Yu. 2020. Recall and learn: Fine-tuning deep pretrained language models with less forgetting. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 7870–7881. Association for Computational Linguistics.

Zhipeng Chen, Yiming Cui, Wentao Ma, Shijin Wang, and Guoping Hu. 2019. Convolutional spatial attention model for reading comprehension with multiple-choice questions. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 6276–6283.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: pre-training text encoders as discriminators rather than generators. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*.

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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Bhuwan Dhingra, Hanxiao Liu, Zhilin Yang, William W. Cohen, and Ruslan Salakhutdinov. 2017. Gated-attention readers for text comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1832–1846.

Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 1693–1701.

Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. 2016. The goldilocks principle: Reading children's books with explicit memory representations. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

Jinchi Huang, Lie Qu, Rongfei Jia, and Binqiang Zhao. 2019. O2u-net: A simple noisy label detection approach for deep neural networks. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 3325–3333. IEEE.

Rudolf Kadlec, Martin Schmid, Ondrej Bajgar, and Jan Kleindienst. 2016. Text understanding with the attention sum reader network. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard H. Hovy. 2017. RACE: large-scale reading comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 785–794.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2018. Language models are unsupervised multitask learners.

Timo Schick and Hinrich Schütze. 2020a. Exploiting cloze questions for few-shot text classification and natural language inference. *CoRR*, abs/2001.07676.

Timo Schick and Hinrich Schütze. 2020b. It's not just size that matters: Small language models are also few-shot learners. *CoRR*, abs/2009.07118.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems*

2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada, pages 5754–5764.

Shuailiang Zhang, Hai Zhao, Yuwei Wu, Zhuosheng Zhang, Xi Zhou, and Xiang Zhou. 2020. DCMN+: dual co-matching network for multi-choice reading comprehension. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 9563–9570.

Boyuan Zheng, Xiaoyu Yang, Yuping Ruan, Quan Liu, Zhen-Hua Ling, Si Wei, and Xiaodan Zhu. 2021. SemEval-2021 task 4: Reading comprehension of abstract meaning. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*.

Haichao Zhu, Furu Wei, Bing Qin, and Ting Liu. 2018. Hierarchical attention flow for multiple-choice reading comprehension. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 6077–6085.