Enhancing Multiple-choice Machine Reading Comprehension by Punishing Illogical Interpretations

Yiming Ju[1,2], **Yuanzhe Zhang**[1,2], **Zhixing Tian**[1,2], **Kang Liu**[1,2],
Xiaohuan Cao[3], **Wenting Zhao**[3], **Jinlong li**[3], **Jun Zhao**[1,2]
[1] National Laboratory of Pattern Recognition, Institute of Automation, CAS, Beijing, China
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
[3] AI Lab, China Merchant Bank, ShenZhen, 518057, China
{yiming.ju, yzzhang, zhixing.tian, kliu}@nlpr.ia.ac.cn
{xhcao, wtzhao, lucida}@cmbchina.com, jzhao@nlpr.ia.ac.cn

Abstract

Machine Reading Comprehension (MRC), which requires a machine to answer questions given the relevant documents, is an important way to test machines' ability to understand human language. Multiple-choice MRC is one of the most studied tasks in MRC due to the convenience of evaluation and the flexibility of answer format. Post-hoc interpretation aims to explain a trained model and reveal how the model arrives at the prediction. One of the most important interpretation forms is to attribute model decisions to input features. Based on post-hoc interpretation methods, we assess attributions of paragraphs in multiple-choice MRC and improve the model by punishing the illogical attributions. Our method can improve model performance without any external information and model structure change. Furthermore, we also analyze how and why such a self-training method works.

1 Introduction

Machine reading comprehension (MRC), which requires a machine to answer questions according to given documents, is an important way to test the ability of intelligence systems to understand human language (Hermann et al., 2015; Chen, 2018). As with other tasks in Natural Language Processing (NLP), deep models have achieved great success on MRC. At the same time, deep models' opaqueness grows in tandem with their power (Doshi-Velez and Kim, 2017), which has motivated efforts to interpret how these black-box models work. Post-hoc interpretation aims to explain a trained model and reveal how the model arrives at the prediction (Jacovi and Goldberg, 2020; Molnar, 2020), as shown in Figure 1. This goal is usually approached with attribution method, which assesses attributions of inputs to model predictions (Bach et al., 2015; Sundararajan et al., 2017; Shrikumar et al., 2017). In NLP, interpretations are usually given by assessing attributions of words, phrases, sentences and paragraphs (Ribeiro et al., 2016; Lundberg and Lee, 2017; Plumb et al., 2018; Chen et al., 2020; De Cao et al., 2020; Jacovi and Goldberg, 2020), in which positive attributions mean support to the prediction and negative ones mean opposition.

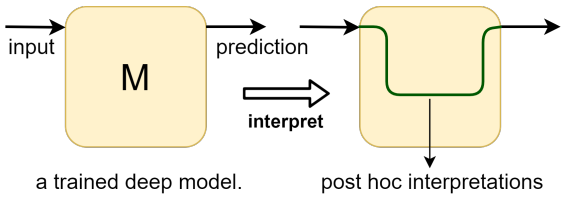


Figure 1: Post-hoc interpretation aims to explain a trained model and reveal how the model arrives at the prediction.

It is well known that the strong fit ability of deep models can cause incredibly high performance on the training set. The correct prediction of an MRC model on the training set can't reflect the model has understood the sample and used a suitable way to predict. Since post-hoc interpretations can provide insights into how the model arrives at the prediction, we argue that we can use these insights to explore problems which predictions can't reflect and to improve model performance. In this work, we interpret multiple-choice MRC models by assessing attributions of paragraphs and improve model performance by punishing the illogical parts of these attributions. The illogical attributions here mean the positive ones to the wrong choices and negative ones to the right, reflecting paragraphs' support to the wrong choices and opposition to the right in the model reasoning process.

Figure 2 shows two specific examples, both of which are from the training set. Numbers on the right are model predictions and numbers on the left are attributions of paragraphs. In example1, if we only observe model predictions, the model makes the right choice: B and does not appear to be

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 3641–3652
November 7–11, 2021.

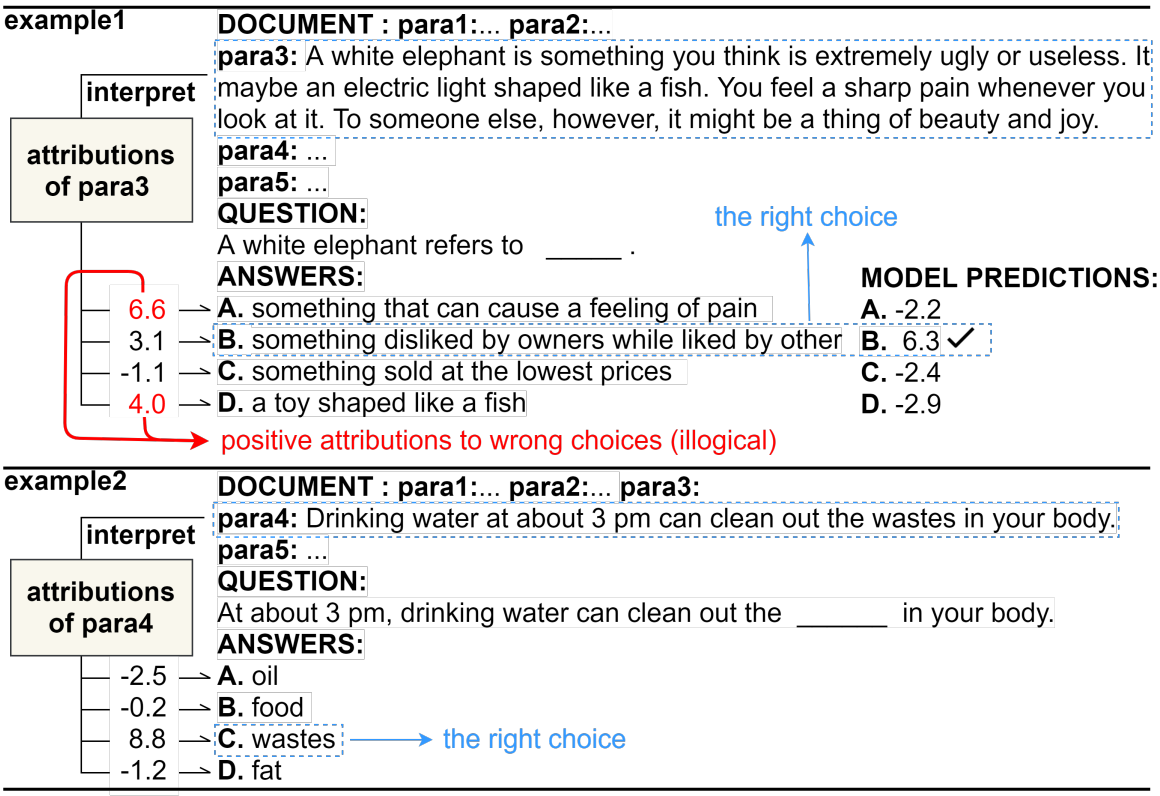


Figure 2: Examples of attribution interpretations for multiple-choice MRC, consisting of attributions of paragraphs to each answer choice. Model predictions are the unnormalized model outputs, in which the largest one corresponds to the predicted choice. The examples are from the RACE training set and trained on BERT$_{base}$.

distracted. However, the attributions show strong support of $para3$ to distractors A and D, which overlap words in it. These attributions show the model's strong dependency on word-overlap form, which is not suitable for answering this question. We constrain the model from such dependency by punishing these attributions. Attributions in this example reflect problems predictions fail to, and we take advantage of this to improve the model.

It is worth noting that we don't simply constrain the model from certain forms. On the contrary, we let the model learn differences between different circumstances. For example, in example2, attributions show strong support of $para4$ to choice C, reflecting the model's dependency on the word-overlap form. However, we will not constrain the model from such dependency as in example1 because the choice C is the right choice and the attribution here is logical. This way, we let the model learn which circumstance is suitable for using such forms and which one isn't.

Compared to existing work (Niu et al., 2020; Jin et al., 2020; Zhu et al., 2020), our method does not need any external information and model structure change. We simply train a new model after getting attributions of the original model. We demonstrate the effectiveness of our method through experiments on three representative datasets: RACE (Lai et al., 2017), MULTIRC (Khashabi et al., 2018) and DREAM (Sun et al., 2019). The main attributions of this paper are summarized as follows:

- We innovatively explore the illogical attributions of the multiple-choice MRC model, and improve the model by punishing them. To the best of our knowledge, we are the first to improve MRC models resorting to post-hoc interpretations.
- We conduct extensive experiments and the results demonstrate that our method can improve multiple-choice MRC consistently on three datasets. Our method can improve both trivial and strong baselines (BERT$_{base}$ and ALBERT$_{xxlarge}$). Furthermore, our method can be applied to the most advanced model.
- We conduct an in-depth analysis of the experimental results and analyze why our method works.

2 Related Work

2.1 Attribution Interpretation Methods

In the post-hoc interpretation research field, methods to get attributions can be classified as erasure-based methods, gradient-based methods, and attention-based methods. In erasure-based methods, attributions of inputs are measured by the change of output when these inputs are removed (Li et al., 2016; Feng et al., 2018). In gradient-based and attention-based methods, the magnitudes of the gradients and attention weights serve as feature importance scores, respectively (Serrano and Smith, 2019; Vashishth et al., 2019; Sundararajan et al., 2017; Shrikumar et al., 2017). Erasure-based methods are model-agnostic. Gradient-based and attention-based methods are applicable for differentiable models and models with attention structures, respectively. The advantage of erasure-based methods is that it is conceptually simple and can optimize well-defined goals (De Cao et al., 2020). The advantage of gradient-based and attention-based methods is they are computationally efficient, However, attention-based and gradient-based methods have received much scrutiny (Sixt et al., 2019; Nie et al., 2018; Jain and Wallace, 2019), arguing that they cannot theoretically prove that the network ignores low-scored features.

2.2 Multiple-choice Machine Reading Comprehension

Multiple-choice MRC requires the machine to decide the correct choice from a set of answer choices given the relevant documents and questions.The question and choice types of multiple-choice MRC are flexible, such as arithmetic, abstract, common sense, logical reasoning, language inference, and sentiment analysis (Lai et al., 2017; Sun et al., 2018; Jin et al., 2020). It requires many advanced reading skills for the machine to perform well on the multiple-choice MRC task.

3 Methods

3.1 Task Description

In multiple-choice MRC, given a relevant document D containing n paragraphs $\{p_1,\ p_2, ..., p_n\}$, a question Q and an choice set with m choices $C = \{c_1, c_2, ..., c_m\}$, the model should determine which choice is correct. The task can be formalized as:

$$\hat{c} = \arg\max_{c' \in C} P(c'|Q, D).$$

3.2 Method

The overview of our method is shown in Figure 3, which contains three steps:
1. Training and Interpreting: train a model and obtain attributions $Attr$.
2. Processing attributions: find the illogical attributions and record the corresponding paragraph indexes I.
3. Retraining: train a new model with I. I is used to normalize the model during training.

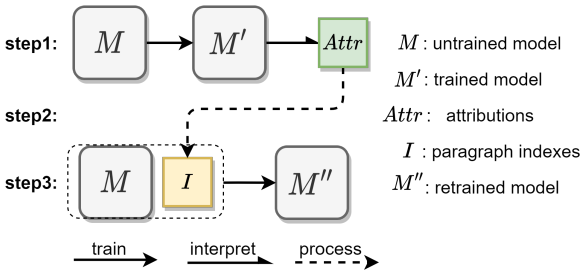


Figure 3: Pipeline of our method.

3.2.1 Training and Interpreting

The commonly used framework for multiple-choice MRC is shown in Figure 4: The document, question, and one of the choices are concatenated together, resulting in m sequences for one question. The model takes these sequences as input separately, and outputs logits $L = \{l_1,\ l_2,\ , ..., l_m\}$ for m choices. choice with the largest logit is the predicted choice. If the softmax function is used to normalize the logits, $P(c_i|Q, D) = \frac{e^{l_i}}{\sum_{j=1}^{m} e^{l_j}}$, and the corresponding cross-entropy loss is:

$$loss_{mc} = -log\left(P(c_r|Q, D)\right),$$

where c_r denotes the correct choice.

We train a multiple-choice MRC model and use erasure-based method to obtain attributions of the trained model. Following previous work (Chen et al., 2020; Feng et al., 2018; Ribeiro et al., 2016; Li et al., 2016), an input subset's attributions are obtained by calculating the output change when erasing this subset. In this work, we use **leave-one-out** (Li et al., 2016) method to perform erasure and get attributions of paragraphs.

As shown in Figure 4, given a document D containing n paragraphs $\{p_1,\ p_2, ..., p_n\}$, we use D^{-i} to represent D with p_i erased. For D^{-i}, Q, C, the model will output logits $L^{-i} = \left\{l_1^{-i},\ l_2^{-i}, ...,\ l_m^{-i}\right\}$, which means model's output with p_i erased. Thus, the attributions of p_i can be

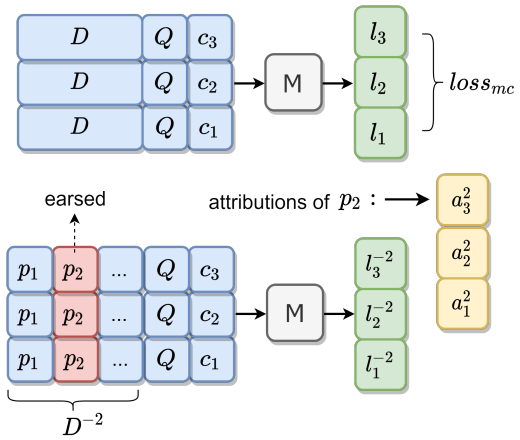


Figure 4: Commonly used framework for multiple-choice MRC and the leave-one-out method to get attributions. a_j^i represents the attributions of paragraph p_i to choice c_j and can be calculated by subtracting l_j^{-i} from l_j (e.g., $a_1^2 = l_1 - l_1^{-2}$).

calculated by subtracting L^{-i} from L. For example, $a_j^i = l_j - l_j^{-i}$ is the attributions of p_i to c_j . For p_i, we get attributions $Attr^i = \{a_1^i, a_2^i, ..., a_m^i\}$ for all choices $C = \{c_1, c_2, ..., c_m\}$. Since erasure-based method is model-agnostic, we don't need to make any changes to the structure of the MRC model.

3.2.2 Processing Interpretations

The illogical attributions mean positive ones to the wrong choices and negative ones to the right, reflecting paragraphs' support to the wrong choices and opposition to the right, formally as:

a_j^i is illogical $\Leftrightarrow a_j^i \neq 0 \wedge (a_j^i > 0 \; xor \; c_j = c_r)$.

The absolute value of a_j^i reflects the degree of support and opposition, which can be used to measure the illogical degree. For each choice, if a_j^i is illogical, we record the corresponding paragraph index i and calculate a_j^i during retraining. To shorten retraining time, if we find more than one illogical attribution for one choice, we only use the one with the largest illogical degree. For each sample, we obtain a paragraph index set $I = \{i_1, i_2, ..., i_m\}$ corresponding to m choices, where i_j is a number or None. (If there is no illogical attribution for c_j, we record $i_j =$ None.)

3.2.3 Retraining

We train a new model and punish the model for generating illogical attributions corresponding to I during the training process. As shown in Figure 5, we calculate attributions $Attr^I = \left\{a_1^{i_1}, a_2^{i_2}, ..., a_m^{i_m}\right\}$

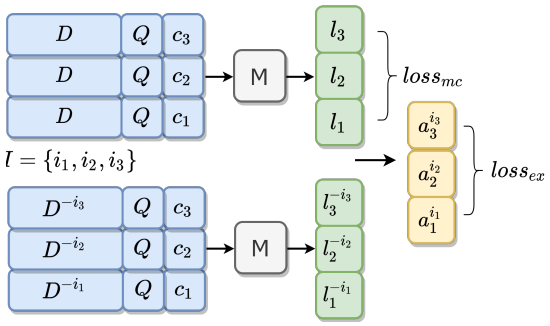


Figure 5: Overview of the retraining process.

and add a extra loss:

$$loss_{ex} = \sum_{j=1}^{m} (a_j^{i_j})^2,$$

to perform punishment. If $i_j =$ None, $D^{i_j} = D$, which means $a_j^{i_j} = 0$ and $loss_{ex} = 0$. The extra loss is used to punish the model for generating illogical attributions. The total loss of retraining is the combination of the task-specific loss and the extra loss:

$$loss = loss_{mc} + \alpha \; loss_{ex},$$

where α is a factor to balance the two loss terms. Though we need to calculate a_j^{ij} at each step, this only requires one additional subtraction operation. The main complexity introduced is the amount of input is doubled. It takes about twice as long to retrain compared to train the initial model.

Recently, a new task form has emerged in multiple-choice MRC, which has an uncertain number of correct choices for each question (Khashabi et al., 2018). It requires the model to determine the correctness of each choice respectively and can be formalized as a binary classification task as shown in Figure 6. We use the same method to solve such tasks, in which the task is seen as a single-choice task with two choices: right and wrong.

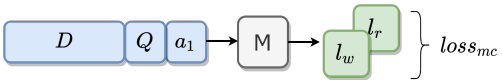


Figure 6: Framework of tasks with an uncertain number of correct choices. The task can be formalized as a binary classification task in which the model should determine whether an choice is correct or wrong.

Model / Dataset	RACE-M		RACE-H		RACE		MultiRC			DREAM	
	Dev	Test	Dev	Test	Dev	Test	Dev			Dev	Test
	acc	acc	acc	acc	acc	acc	$F1_m$	$F1_a$	$EM0$	acc	acc
BERT$_{base}$	72.3	71.7	64.1	62.6	66.5	65.2	71.8	69.1	21.2	63.4	63.2
+ retraining	**74.2**	**72.6**	**66.1**	**65.2**	**68.5**	**67.4**	**73.5**	**70.7**	**22.9**	**64.1**	**63.7**
ALBERT$_{xxlarge}$	90.2	89.3	86.2	85.7	87.3	86.7	88.3	86.5	59.5	89.2	88.5
+ retraining	**91.3**	**91.2**	**87.2**	**86.7**	**88.4**	**88.0**	**89.4**	**87.8**	**61.6**	**90.2**	**90.0**

Table 1: Results on three multiple-choice MRC datasets. (F1a: F1 score on all choices; F1m: macro-average F1 score of all questions; EM0: exact match.) Results of baseline models are copied from the corresponding leaderboards and papers (Niu et al., 2020). Additionally, our reproduced results of ALBERT$_{xxlarge}$ and BERT$_{base}$ on the RACE dataset are higher than those on the leaderboards, so we report our results.

4 Experiments

4.1 Datasets

We evaluate our method on three representative multiple-choice MRC datasets:

RACE (Lai et al., 2017) is a large-scale dataset collected from English examinations. RACE has a wide variety of question types such as summarization, inference, deduction and context matching, and most of the questions need reasoning more than lexical-level matching.

MULTIRC (Khashabi et al., 2018) requires gathering information from multiple sentences to choice a question. MULTIRC requires evaluating the correctness of each choice individually. Following previous work (Yadav et al., 2020; Niu et al., 2020), we use the original MULTIRC dataset[1], not the version on SuperGLUE (Wang et al., 2019).

DREAM (Sun et al., 2019) is a dialogue-based dataset collected from English examinations. Most of the questions are non-extractive and need reasoning from more than one sentence.

4.2 Baselines and Implement Details

Since the wide use of pre-trained language models in NLP, we choose two pre-trained language models **BERT**$_{base}$ (Devlin et al., 2018) and **ALBERT**$_{xxlarge}$ (Lan et al., 2019) as the trivial and strong baselines respectively. We use the same model architecture as that in Transformer[2], which is commonly used in multiple-choice MRC: a pre-trained language model as the encoder and a single-layer linear network connected to [CLS] as the matching network. In addition to the commonly used model architecture, we also experimented on **DUMA** (Zhu et al., 2020), which is the state-of-the-art model architecture on the DREAM leaderboards.

Our implementation is based on Transformer[2]. We use default model settings in Transformer[2] and follow basic experimental settings in the leaderboards and corresponding papers. We directly adopted the same learning rate and batch size as the baseline models for retraining. We search the coefficient α among 0.1, 0.5, and 1. For MULTIRC and DREAM, we use the original paragraph divisions of the datasets. For RACE with a noisy paragraph division, we limit the length of paragraphs based on the original paragraph division.

4.3 Main Results

We evaluate our method on three multiple-choice MRC datasets and adopt the metrics from the referred papers. The results are summarized in Table 1. Our method can improve model performance remarkably: **1.33%** and **1.63%** average performance improvement for BERT$_{base}$ and ALBERT$_{xxlarge}$, which demonstrates that our method can help both a trivial baseline as well as a competitive baseline. Furthermore, ALBERT$_{xxlarge}$ + retraining produces competitive results: On the RACE leaderboard, our result only lags behind super-large pre-trained language model Megatron-LM (Shoeybi et al., 2019). On the DREAM leaderboard, our results only lags behind DUMA (Zhu et al., 2020). Note that we only compare single-task and non-ensemble models. In addition, since our method is model-agnostic, we also experiment with DUMA as model architecture on the DREAM dataset, which is the state-of-the-art model architecture on the DREAM leaderboards.

Because Zhu et al. (2020) did not provide some important details such as the number of DUMA attention heads and the head size, we use the settings in another re-implementation Wan (2020) and

[1] https://cogcomp.seas.upenn.edu/multirc/

[2] https://github.com/huggingface/transformers

Model	Dev	Test
basic model architecture	89.2	88.5
DUMA (Zhu et al., 2020)	89.3	90.4
DUMA (implementation by (Wan, 2020))	90.7	88.6
DUMA(our implementation)	90.5	88.9
+ retraining	**91.3**	**90.2**

Table 2: Experimental results with DUMA as model architecture on the DREAM dataset. All models use ALBERT$_{xxlarge}$ as the encoder.

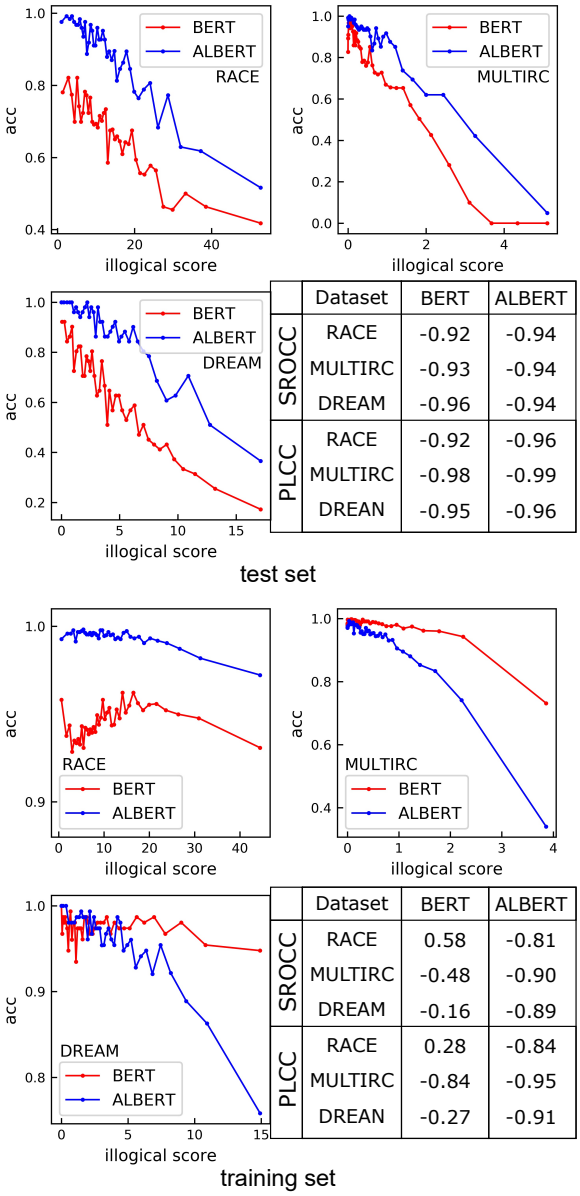
gain similar results: higher accuracy on the dev set and lower accuracy on the test set compared to Zhu et al. (2020). As shown in Table 2, the DUMA architecture outperforms the basic model architecture on the DREAM dataset, and our method can further improve model performance on this basis. The results further demonstrate the effectiveness of our method.

4.4 The Relationship Between Illogical Attributions and Model Performance

In this section, we explore the relationship between illogical attributions and model performance. We use the maximum value of illogical attributions as the illogical score of a choice and sum the scores of all choices as the illogical score of a sample. According to illogical scores, we sort samples and divide them into 20 subsets of the same number of samples. We evaluate model performance on these subsets and investigate the relationship between illogical score and model performance. We use two widely used correlation coefficients: Spearman rank-order correlation coefficient (SROCC) (Spearman, 1961) and Pearson correlation coefficient (PLCC) (Benesty et al., 2009) to evaluate the correlation between them.

Test Set Results We first experiment on the test set. As shown in Figure 7, the SROCC and PLCC values on the test are close to -1. The results show that there is a strong correlation between illogical score and model performance, where a higher illogical score corresponds to poorer model performance. Moreover, since MRC models' understanding ability is evaluated via test set performance, the results also demonstrate that we can utilize interpretations to evaluate MRC models' understanding ability from another perspective.

Training Set Results The correlation on the training set is weaker than the test set. This is because the model has fitted training samples during training, which causes the training set performance cannot reflect MRC models' understanding ability. However, we find a interesting phenomenon: the correlation of strong model is stronger consistently in all datasets. We hypothesize that the stronger model can fit more linguistic features of the training samples while the trivial model needs to fit more unique features of the training samples. Since these unique features are hard to interpret and generalize



	Dataset	BERT	ALBERT
SROCC	RACE	-0.92	-0.94
	MULTIRC	-0.93	-0.94
	DREAM	-0.96	-0.94
PLCC	RACE	-0.92	-0.96
	MULTIRC	-0.98	-0.99
	DREAN	-0.95	-0.96

	Dataset	BERT	ALBERT
SROCC	RACE	0.58	-0.81
	MULTIRC	-0.48	-0.90
	DREAM	-0.16	-0.89
PLCC	RACE	0.28	-0.84
	MULTIRC	-0.84	-0.95
	DREAN	-0.27	-0.91

Figure 7: Relationship between illogical score and model performance. SROCC, KROCC are correlation coefficients in which the closer the absolute value is to 1, the stronger the correlation. We scaled the illogical score of BERT proportionally to draw on the same graph with ALBERT.

to test set samples, the correlation between interpretation (illogical score) and performance is weak, and the test set performance is poor.

4.5 Effectiveness of Retraining the Illogical Interpretations

In this section, we explore whether the illogical attributions are constrained after retraining. We compare the illogical score of the retrained model to the original model. Figure 8 shows an example of change in illogical scores after retraining. We can see from the figure that most samples' illogical scores are constrained close to zero after retraining. In this example, the average value of illogical scores declines from **1.22** to **0.11** on the training set and declines from **1.43** to **0.37** on the test set. Table 3 shows changes in the average illogical score after retraining. The average illogical scores decline consistently in six experiments, which demonstrates the effectiveness of the retraining strategy.

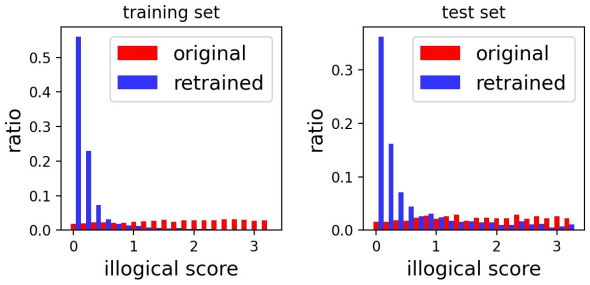


Figure 8: The change in illogical scores. The results are from ALBERT$_{xxlarge}$-DREAM.

Set	training set	test set
BERT	retrained/original	retrained/original
RACE	0.19	0.36
MULTIRC	0.16	0.78
DREAM	0.08	0.35
ALBERT	retrained/original	retrained/original
RACE	0.06	0.09
MULTIRC	0.19	0.63
DREAM	0.09	0.26

Table 3: Changes of the average illogical scores after retraining. We show the ratio of the retrained value and the original value.

According to the illogical score of the original model, we divide the test set into ten subsets with the same number of samples . Model performance on these subsets is shown in Figure 10. On subsets with high illogical scores, the model performance gets a remarkable gain after retraining. However, the model performance declines after retraining on some low-score subsets. We hypothesis that the punishment of illogical interpretations will affect the model's confidence in using the right reasoning form in some samples. For example, although we let the model to learn the difference between examples in Figure 2, punishment in example1 may affect the model's confidence in using the same form in example2.

5 Discussion

5.1 Using Post-hoc Interpretations to Improve NLP Models

Existing work focusing on using post-hoc interpretations to improve NLP models has forced the model to generate the 'correct' interpretation. Although conceptually simple, 'correct' interpretations served as the ground truth are difficult to get. For example, Liu and Avci (2019) uses human-selected terms as the target attributions, which is noisy and hard to be generalized to other datasets. Chen and Ji (2020) does sampling during training and resorts to mean-field approximation (Blei et al., 2017) to get target attributions, which leads to difficult training and unstable results. Moreover, their improvements are limited, and they all choose to experiment on simple text classification tasks, in which some words can be regarded as the decisive factors for prediction. However, for MRC tasks that often require complex reasoning, getting interpretations served as the ground truth to guide the model is more difficult and costly.

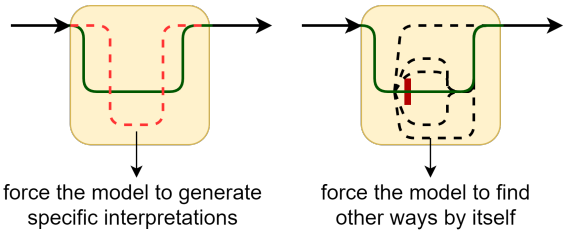


Figure 9: Two ways of using post-hoc interpretations to improve NLP models

Different from existing methods, we focus on finding illogical parts of interpretations of trained models instead of the ground-truth interpretations of the task. As shown in Figure 9, we punish the illogical parts and force the model to find other ways to get the prediction by itself. Because forms found by humans are usually easy to learn for deep models, it is hard to create interpretations helpful for

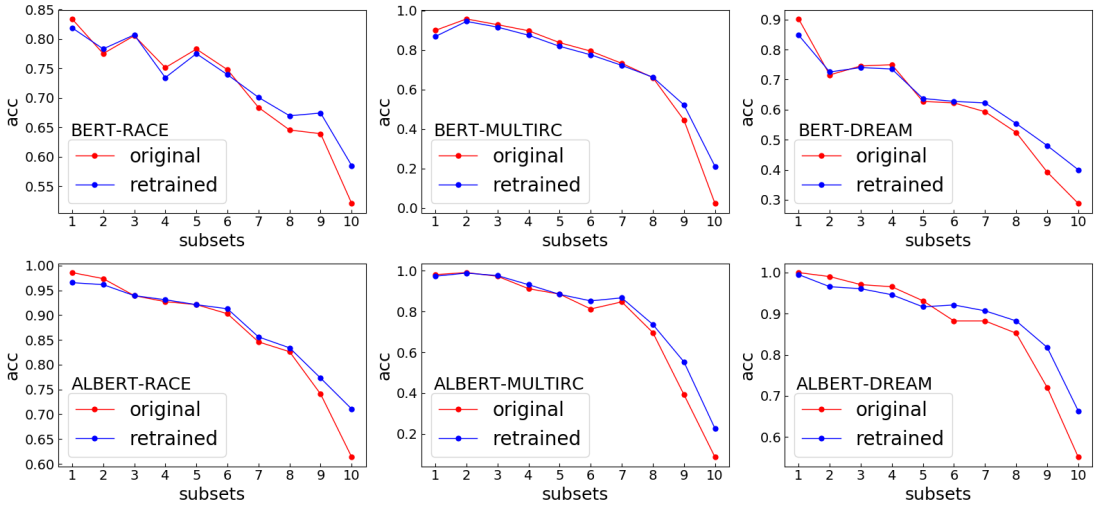


Figure 10: Model performance on subsets where the larger index corresponding to the higher illogical score. Since the F1 score is affected by the ratio of positive and negative samples, we use accuracy as the metrics for MULTIRC.

deep models. We believe analyzing models' interpretations and finding problems is a more suitable way to improve model performance.

5.2 Guilding Strong Models by Penalizing Errors

We get similar average performance improvement for the strong baseline ALBERT$_{xxlarge}$ and the trivial baseline BERT$_{base}$. This is contrary to many methods, which are usually effective on trivial baselines but difficult to get improvement on strong ones. For example, similar work (Niu et al., 2020) focusing on improving multiple-choice MRC models designs a sentence selector for learning evidence sentences. Their method gets significant improvement on BERT$_{base}$ but fails to apply to a stronger baseline RoBERTa$_{large}$ (Liu et al., 2019). Telling a strong model which sentences are evidence sentences is more difficult because the model's strong learning ability might makes this extra guidance redundant. This suggests a hypothesis that it is more effective to penalize errors than to promote correct answers for a strong model when high-quality labeled correct answers are not available.

5.3 Case Study

We can observe the wrong reasoning process of deep MRC models through analysis of the illogical interpretations, some of which are interesting and unexpected. For example in Table 4, the model was not distracted by the distractor `'March 5th'` which exactly appears in the document. However, $para2$ made a very high attribution (strong support) to the wrong choice `'March 15th'`. We hypothesis that the model understands `'not really'` is a negation of `'March 5th'`. However, the model notices `'three'` in $para2$ and believes that the choice is 3 times 5 equals 15. We observed the linguistic characteristics of high illogical score examples on the test set, and found they are different between BERT$_{base}$ and ALBERT$_{xxlarge}$. For example, examples with negation and transition tend to have high illogical scores on BERT$_{base}$, but have low illogical scores on ALBERT$_{xxlarge}$. We hypothesis that is because ALBERT$_{xxlarge}$ perform better than BERT$_{base}$ in not being distracted by these grammatical phenomena. We suggest analyzing interpretations and finding problems can help humans get a more comprehensive understanding of deep models.

DOCUMENT:
para1: M: Were you here on March 5th?
para2: W: Mm. not really. In fact I arrived three days later.
QUESTION:
When pid the woman arrive?
ANWSERS:
March 15th. **(illogical attr: (para2, 14.8))**
March 5th. (illogical attr: None)
March 8th. $\surd$ (illogical attr: None)

Table 4: An example of illogical attributions in ALBERT-DREAM. ($\surd$: the correct choice. (parai, score): paragraphs and the illogical score, (None) means there are no illogical attribution for this choice.) The example is from the test set of DREAM.

6 Conclusion and Future Work

In this work, we improve multiple-choice MRC resort to attribution interpretations. Experimental results show that our method can remarkably improves model performance on three representative datasets. We believe using post-hoc interpretations to improve NLP models is a promising research field. The future work contains two aspects:

1. We plan to experiment with our method on other tasks, such as natural language inference and sentiment analysis, and explore methods applicable to tasks without choice options or specific classes, such as generative MRC and span extractive MRC.

2. In addition to attributions, we plan to use other forms of post-hoc interpretations, such as feature interaction, to improve NLP models.

Acknowledgements

This work is supported by the National Key Research and Development Program of China (No. 2020AAA0106400), the National Natural Science Foundation of China (No.61922085, No.61976211, No.61906196). This work is also supported by Beijing Academy of Artificial Intelligence (BAAI2019QN0301), the independent research project of National Laboratory of Pattern Recognition and in part by the Youth Innovation Promotion Association CAS.

References

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140.

Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. 2009. Pearson correlation coefficient. In *Noise reduction in speech processing*, pages 1–4. Springer.

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. 2017. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877.

Danqi Chen. 2018. *Neural reading comprehension and beyond*. Stanford University.

Hanjie Chen and Yangfeng Ji. 2020. Learning variational word masks to improve the interpretability of neural text classifiers. *arXiv preprint arXiv:2010.00667*.

Hanjie Chen, Guangtao Zheng, and Yangfeng Ji. 2020. Generating hierarchical explanations on text classification via feature interaction detection. *arXiv preprint arXiv:2004.02015*.

Nicola De Cao, Michael Schlichtkrull, Wilker Aziz, and Ivan Titov. 2020. How do decisions emerge across layers in neural models? interpretation with differentiable masking. *arXiv preprint arXiv:2004.14992*.

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*.

Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Shi Feng, Eric Wallace, Alvin Grissom II, Mohit Iyyer, Pedro Rodriguez, and Jordan Boyd-Graber. 2018. Pathologies of neural models make interpretations difficult. *arXiv preprint arXiv:1804.07781*.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. *Advances in neural information processing systems*, 28:1693–1701.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness? *arXiv preprint arXiv:2004.03685*.

Sarthak Jain and Byron C Wallace. 2019. Attention is not explanation. *arXiv preprint arXiv:1902.10186*.

Di Jin, Shuyang Gao, Jiun-Yu Kao, Tagyoung Chung, and Dilek Hakkani-tur. 2020. Mmm: Multi-stage multi-task learning for multi-choice reading comprehension. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8010–8017.

Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. 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 252–262.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. Race: Large-scale reading comprehension dataset from examinations. *arXiv preprint arXiv:1704.04683*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Jiwei Li, Will Monroe, and Dan Jurafsky. 2016. Understanding neural networks through representation erasure. *arXiv preprint arXiv:1612.08220*.

Frederick Liu and Besim Avci. 2019. Incorporating priors with feature attribution on text classification. *arXiv preprint arXiv:1906.08286*.

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*.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in neural information processing systems*, pages 4765–4774.

Christoph Molnar. 2020. *Interpretable Machine Learning*. Lulu. com.

Weili Nie, Yang Zhang, and Ankit Patel. 2018. A theoretical explanation for perplexing behaviors of backpropagation-based visualizations. *arXiv preprint arXiv:1805.07039*.

Yilin Niu, Fangkai Jiao, Mantong Zhou, Ting Yao, Jingfang Xu, and Minlie Huang. 2020. A self-training method for machine reading comprehension with soft evidence extraction. *arXiv preprint arXiv:2005.05189*.

Gregory Plumb, Denali Molitor, and Ameet S Talwalkar. 2018. Model agnostic supervised local explanations. In *Advances in Neural Information Processing Systems*, pages 2515–2524.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Sofia Serrano and Noah A Smith. 2019. Is attention interpretable? *arXiv preprint arXiv:1906.03731*.

Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. 2019. Megatron-lm: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053*.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning important features through propagating activation differences. *arXiv preprint arXiv:1704.02685*.

Leon Sixt, Maximilian Granz, and Tim Landgraf. 2019. When explanations lie: Why many modified bp attributions fail. *arXiv*, pages arXiv–1912.

Charles Spearman. 1961. The proof and measurement of association between two things.

Kai Sun, Dian Yu, Jianshu Chen, Dong Yu, Yejin Choi, and Claire Cardie. 2019. Dream: A challenge data set and models for dialogue-based reading comprehension. *Transactions of the Association for Computational Linguistics*, 7:217–231.

Kai Sun, Dian Yu, Dong Yu, and Claire Cardie. 2018. Improving machine reading comprehension with general reading strategies. *arXiv preprint arXiv:1810.13441*.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. *arXiv preprint arXiv:1703.01365*.

Shikhar Vashishth, Shyam Upadhyay, Gaurav Singh Tomar, and Manaal Faruqui. 2019. Attention interpretability across nlp tasks. *arXiv preprint arXiv:1909.11218*.

Hui Wan. 2020. Multi-task learning with multi-head attention for multi-choice reading comprehension. *arXiv preprint arXiv:2003.04992*.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *arXiv preprint arXiv:1905.00537*.

Vikas Yadav, Steven Bethard, and Mihai Surdeanu. 2020. Unsupervised alignment-based iterative evidence retrieval for multi-hop question answering. *arXiv preprint arXiv:2005.01218*.

Pengfei Zhu, Hai Zhao, and Xiaoguang Li. 2020. Dual multi-head co-attention for multi-choice reading comprehension. *arXiv preprint arXiv:2001.09415*.

A Experiments Details

Getting Attributions

There is no need to select the model used to generate attributions carefully. In our experiment, we use the last saved results of the maximum training step. We use the original paragraph division of the dataset for MULTIRC and DREAM. For RACE with a chaotic paragraph division, we limit the maximum length and minimum length of paragraphs based on the original division. Specifically, if a paragraph's length is less than 10, then combine it with the previous one. If the length exceeds 30, the beginning of the next sentence after 30 is the beginning of a new paragraph. For MULTIRC with an uncertain number of correct choices, which can be formalized as a binary classification task, we see the task as a single-choice task with two choices: the choice is right, and the choice is wrong. Because of the opposition relation between these two

Model	BERT$_{base}$			ALBERT$_{xxlarge}$			DUMA
Dataset	RACE	MULTIRC	DREAM	RACE	MULTIRC	DREAM	DREAM
max seq-length	512	512	512	512	512	512	512
learning rate	2e-5	2e-5	2e-5	1e-5	1e-5	1e-5	1e-5
batch size	32	32	24	8	24	8	8
epoch$_{max}$	8	14	16	4	4	4	2
warmup steps	1000	1000	400	1000	400	300	150
α	0.5	0.1	0.5	0.1	1	0.5	0.1

Table 5: Hyperparameters of retraining.

choices, we only record the paragraph indexes of the option representing the choice is wrong for retraining.

Retraining

For hyperparameters, all three tasks use 512 as the maximum sequence length. We adopted the same learning rate and batch size as the baseline models for retraining. We use the default model settings in Transformer[2]. We search the coefficient $alpha$ among 0.1, 0.5, and 1. The details are shown in Table 5. We follow the experimental settings from the leaderboards and corresponding papers. If there is no relevant information, we retrain the model three times and pick the model with the best accuracy on the dev set. We use FP16 training from Apex[3] for accelerating the training process, and all the experiments are run on two Titan RTX GPUs and two Tesla V100 GPUs. We calculate the task-specific loss and regulation loss separately during retraining because of the limitation of video memory.

B Case Study

We present cases with the top 5 high illogical scores on the DREAM test set:

BERT
√: the correct choice
(pi: score): paragraphs and the illogical score

DOCUMENT:
p1: M: pid you go shopping yesterday?
p2: W: Yes. I bought a bag for my mother and some tea for my father.
QUESTION:
What pid the woman buy for her father?
ANWSERS:
A bag. **(p2: 3.6)**
Some tea. √ (None)
A bag and some tea. **(p2: 11.8)**

DOCUMENT:
p1: M: Is that Ann?
p2: W: Yes.
p3: M: This is Mike. How are things with you?
p4: W: Oh, very well, but I'm very busy.
p5: M: Busy? But you've finished all your exams?
p6: W: Yes, but I have to help my little sister with her foreign language.
p7: M: How about coming out with me this evening? There's a new film on.
p8: W: I'm afraid I can't. A friend of mine is coming from the south and I have to go to the station to meet him.
p9: M: What a pity! How about the weekend then?
p10: W: No, I've arranged to go to an art exhibition with my parents.
p11: M: What about next week sometime?
p12: W: Maybe.
QUESTION:
What is the woman going to do tonight?
ANWSERS:
Help her sister with English. **(p6: 11.3)**
Meet her friend at the station. √ (None)
Go to an exhibition with her parents. **(p10: 11.7)**

DOCUMENT:
p1: W: You are interested in sports, aren't you?
p2: M: Yes. I go swimming once a week and play tennis twice a month.
QUESTION:
How often does the man go swimming?
ANWSERS:
Once a week. √ (None)
Twice a week. **(p2: 11.6)**
Once a month. **(p2: 11.7)**

DOCUMENT:
p1: W: Can you tell me where the café is?
p2: M: Yes. It's to the left of ...no, to the right ... of the tree.
p2: W: I thought it was behind the Magic Castle.
QUESTION:
Where is the café, accorping to the man?
ANWSERS:
To the right of the tree. **(p2: 11.6)**
To the left of the tree. **(p2: 11.6)**
Behind the Magic Castle. √ (None)

[3]https://github.com/NVIpiA/apex

DOCUMENT:
p1: W: Good afternoon! Dr. Perkins' office.
p2: M: Good afternoon. I'd like to speak to the doctor. Is he in?
p3: W: Who is that calling, please?
p4: M: My name is Li Hong. I'm from China.
p5: W: I'm sorry. Dr. Perkins is now at an important meeting and can't choice your call.
p6: M: I'm an exchange scholar. Dr. Perkins asked me to give a lecture. There are some details I want to piscuss with him.
p7: W: I see, but you must speak to himself about that. Oh, well, if you leave your number, I'll tell him to ring you as soon as he is available.
p8: M: Thanks. My number is 7838298.
QUESTION:
Why does the man want to talk to Dr. Perkins on the phone?
ANWSERS:
To piscuss something with the doctor. √ (None)
To ask Dr. Perkins to give a lecture. **(p6: 11.3)**
To see him about his illness. **(p2: 0.58)**

ALBERT
√: the correct choice
(pi: score) : paragraphs and the illogical score

DOCUMENT:
p1: M: Were you here on March 5th?
p2: W: Mm. not really. In fact I arrived three days later.
QUESTION:
When pid the woman arrive?
ANWSERS:
March 15th. **(p2: 14.8)**
March 5th. (None)
March 8th. √ (None)

DOCUMENT:
p1: W: We'd like some information, please. We want to go to England.
p2: M: OK. What do you want to know?
p3: W: Well, first of all, we want to know the air fare to London.
p4: M: When do you want to go?
p5: W: We don't really know, maybe July.
p6: M: I see. Well, in May and June the fare is 480 dollars, but it's less in March and April. It's only 460 dollars.
p7: W: And what about July?
p8: M: It's more in July.
p9: W: More? How much is it then?
p10: M: It's 525 dollars.
p11: W: Oh... I'll think it over before I make the final decision.
QUESTION:
In which month or months is the fare to London the most expensive?
ANWSERS:
In March and April. **(p6: 5.5)**
In May and June. **(p6: 14.7)**
In July. √ **(p1: 0.2)**

DOCUMENT:
p1: W: I wish I hadn't hurt Linda's feeling like that yesterday. You know I never meant to.
p2: M: The great thing about Linda is that she doesn't hold any grudges. By tomorrow she'll have forgotten all about it.
QUESTION:
What does the man say about Linda?
ANWSERS:
She is forgetful. **(p2: 14.1)**
She is considerate. **(p2: 4.2)**
She is forgiving. √ (None)

DOCUMENT:
p1: W: Are you traveling alone?
p2: M: No, I will take my family abroad this time. My wife and our three children are all going along with me.
p3: W: What a wonderful experience that will be! I wish I could travel abroad some day.
QUESTION:
How many people will go with the man?
ANWSERS:
Three. (None)
Five. **(p2: 14.0)**
Four. √ (None)

DOCUMENT:
p1: W: Hello. Can I speak to Linda, please?
p2: M: Sorry, there's no Linda here.
QUESTION:
What does he mean?
ANWSERS:
The girl can't speak to Linda. **(p2: 13.5)**
Linda isn't here now. **(p2: 13.7)**
The girl has pialed the wrong number. √ (None)