**文章编号**：1007-5321(2019)06-0126-08　　**DOI**：10. 13190/j. jbupt. 2019-202

# 通过检测语义分歧识别无答案问题

刘咏彬[1]，　王小捷[1]，　袁彩霞[1]，　易　炼[2]

（1. 北京邮电大学 计算机院，北京 100876；2. 阿里巴巴(北京)软件服务有限公司，北京 100022）

**摘要**：机器阅读理解中存在无法仅从给定文档中获取问题答案的特殊情况，为此，基于语义冲突检测的机器阅读理解网络(SCDNet)提出应通过检测问题与文档内容之间的语义分歧来识别这种情况．经分析发现，文档无法为问题提供答案的根本原因主要分为两类：一是文档中不包含问题所需的语义信息；二是二者包含的语义成分之间存在分歧．据此推断，可以通过检测文档语义信息是否全面涵盖问题所需的信息来识别问题是否可由文档信息给出回答．此外，通过在损失函数中加入答案文本长度惩罚项，网络优化目标函数更接近评测指标，系统性能得到提升．网络模型使用联合训练模型建模无答案的问题识别与答案抽取 2 个子任务，并使用端到端的方式训练．实验结果证明，其对无答案问题类别预测的正确率超过了性能先进的基线模型 SAN2.0，在 SQuAD2.0 数据集上取得了 72.43 的 F1 值和 76.96 的无答案问题识别正确率．

**关　键　词**：机器阅读理解；问答系统；无答案的问题

**中图分类号**：TN929.53　　　**文献标志码**：A

# Unanswerable Questions Recognition by Semantic Discrepancy Detection

LIU Yong-bin[1]，　WANG Xiao-jie[1]，　YUAN Cai-xia[1]，　YI Lian[2]

(1. School of Telecommunication Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China;
2. Alibaba (Beijing) Software Services Company Limited, Beijing 100022, China)

**Abstract**: Machine reading comprehension (MRC) with unanswerable questions is challenging to the field of natural language processing research. Unlike previous work which ignores the mechanism of answerable and unanswerable, the semantic conflicts detection-based MRC network (SCDNet) was proposed aiming at detections of no-answer (NA) questions through semantic conflicts detection network. The basic idea is that if the given question is unanswerable, there exists semantic absence or conflicts between the question and the reference passages. Therefore, SCDNet predicts the NA probability by checking whether the passage covers the integral semantics of the question. Besides, in order to extract the exact answer from the passage, SCDNet is applied an answer length penalty in the loss function, which helps the learning objective to be more consistent with the evaluation metrics. SCDNet packs the NA question predictor and the answer extractor in a joint model and is trained in an end-to-end manner. Experiments show that SCDNet performs better than some strong baseline models, and achieve an F1 score of 72.43 and 76.96 NA accuracy on SQuAD 2.0 dataset.

**Key words**: machine reading comprehension; question answering; unanswerable question

**收稿日期**：2019-09-28
**基金项目**：中央高校基本科研业务费专项资金项目(500419302)

**作者简介**：刘咏彬(1977—)，女，讲师，E-mail：liuyb@bupt.edu.cn.

Machine reading comprehension (MRC) is a kind of question answering system based on the facts in the reference text. It has received considerable attention over the past few years. With the benefit of the first high-quality and large MRC dataset SQuAD 1.1[1], MRC models with deep learning architectures are proposed and have achieved promising results on a variety of tasks. However, most of them are trained to choose the most probable answer by comparing the candidate answers under the hypothesis that the given text always has the correct answer in its context. However, this hypothesis cannot be guaranteed in real world, some questions might be unanswerable only by its reference text. SQuAD 2.0[2] released recently offers a no-answer (NA) option to each question. Table 1 gives an example in SQuAD 2.0 which cannot be answered from the given reference text. The unanswerable questions in SQuAD 2.0 are written specially to be similar to answerable ones, all of the questions' contents are relevant to the passage and each of the unanswerable questions is provided a plausible answer which is not real. So, there are no obvious differences between answerable questions and unanswerable questions, and they must be distinguished by deep semantic matching.

**Table 1　Examples of SQuAD 2.0**

| No. | Examples |
|---|---|
| 1 | Passage: The first full-scale working railway steam locomotive was built by Richard Trevithick in the United Kingdom and, on 21 February 1804, the world's first railway journey took place as Trevithick's unnamed steam locomotive hauled a train along the tramway from the Pen-y-darren ironworks, near Tydfil to Abercynon in south Wales. |
|  | Question: Where did the world's first railway journey terminate? |
|  | Ground truth answer: Abercynon |
| 2 | Passage: Tobacco smoking (including secondhand smoke) and short-term exposure to air pollution such as carbon monoxide, nitrogen dioxide, and sulfur dioxide (but not ozone) have been associated with MI. |
|  | Question: What forms of air pollution does smoking tobacco cause? |
|  | Ground truth answer: <No Answer> |
|  | Plausible answer: carbon monoxide, nitrogen dioxide, and sulfur dioxide |

To deal with the NA problem, most current models append a special symbol to the passage to represent NA, models are trained to point to this special symbol when there comes an unanswerable question. Additionally, UNet[3] proposed a universal node for classifing the NA problems, SAN 2.0[4] used a binary NA classifier to be a joint training target, Read + verify[5] applied a binary generative pre-training (GPT) verifier to check the entailment relation between the question and the predicted answer sentence. Although they got great success in improving the NA prediction, they all ignore the mechanism of NA.

An NA question is difficult to recognize because there is no special syntactic or semantic NA-feature, it only depends on the question-passage semantic relation, and can only be distinguished from their semantic matching results. From this point of view, the NA problem is similar to the natural language inference (NLI) task[6-7] except that the NLI task focuses on the relationship between two sentences while MRC task cares more about a question and a passage. A question would be unanswerable if the given passage doesn't have enough information to support the facts the question asked, or if the semantics conveyed by the passage conflicts with the facts asked by the question, while when a question is answerable by a passage, every semantic component of the question can be found in the passage. Table 1 demonstrates two examples to explain our claim. Example 1 shows that the highlighted semantic components of the question all exist in the passage, they locate concentrated in the question but scattered in the passage. Example 2 shows that the counterpart of the question word "cause" or the question phrase "smoking tobacco cause" doesn't exist in the passage. Inspired by this semantic location pattern and the works from NLI, it is proposed that the NA problem can be formulated as a semantic matching task that detects semantic conflicts and absence between question and passage.

To recognize an answerable question, every semantic constituents of thequestion needs to find its counterpart in the passage to ensure the passage matches all the semantic components from the question, and the semantic integrity checking is more reasonable to be modeled from the question side. In SCDNet, the question's semantic matching counterpart is collected from the passage by query-to-passage attention, the attention averaged passage vectors are concatenated after the query vectors forming the passage-aware question, and then fed into a BiLSTM layer for the absence and conflicts detection.

Furthermore, the answer prediction subtask is jointly trained with the NA classifier, the answer prediction network utilizes an iterative pointer network to predict the answer's boundary. An answer boundary penalty is used in the loss function to constraint the start-end pair in a reasonable relative position. The penalty is a function of the distance of the highest confident start-end pair, it improves the model's F1 performance for about 1 percent.

Our contributions can be summarized as follows:

· SCDNet, a novel neural model is proposed which predicts the unanswerable questions and extracts answers for answerable questions, while the two procedures are packed together into a joint model and are trained jointly.

· SCDNet uses a simple Bi-directional long short term memory (BiLSTM) + maxout network to predict whether there is semantic gaps or semantic conflicts between the passage and the fact that question is concerned about, and improves the original pointer network to make it predict the answer span with a suitable length.

· Through extensive experiments on benchmark datasets,it is demonstrated that SCDNet's effectiveness over the competitive state-of-the-art approaches by 1.6 percent in NA classification accuracy.

## 1 Related work

Machine reading comprehension, a challenge to enable machines to answer questions after reading given textual evidence, has attracted considerable attention from both academic and industrial communities. SQuAD 1.1[1] was released as the first large-scale dataset created by humans through crowdsourcing, and it constrains the answer to be a fragment of the given passage, while SQuAD 2.0[2] contains a collection of questions that might be unanswerable. MRC models must not only extract an answer but also determine if the question is unanswerable and refrain from answering an unanswerable question.

A typical MRC framework is shared among the previous models. First, it encodes the questions and passages, then refines passage representation to get a more elaborated question-aware passage from a matching network, finally predicts the answer span and output the final answer. Most of the methods focused on how to improve question-aware passage representation, question-passage fusion process, and the attention mechanism.

As the answer defined in SQuAD is a continuous span of the passage, the broadly used strategy for extract answer is to predict the probability of each passage position being the start or end of an answer span. Most current models predict the probability directly from the question-aware passage word features with a fully connected layer and a softmax function. Pointer network[8] is usually used to predict the start and end positions sequentially, which makes the end prediction step depend on the previous start prediction. Reinforced mnemonic reader (RMR)[9], stochastic answer network (SAN)[10] used iterative output layer to refine their answer by multi-step reasoning.

The predicted start and end points of the candidate answer must obey a position constraint: the start always goes before the end, and the answer length is usually not too lengthy. So, the models usually use an extra span probability rectification step following the network's output to lower or zero the illegal start-end probabilities such as in[3-4]. Despite their success, the models don't include the answer length limitation in the training process, and the answer being evaluated maybe not the one from the start-end pair with the highest confidence.

Since SQuAD 2.0 was released, several models try to solve the NA problem by adding an extra choice to their original SQuAD model. BiDAF-no-answer (BNA)[11] adds a trainable bias as the no-answer representation to compete for the answer boundary with passage words. RMR + Verify[5] adds two independent loss items to its loss function and use a GPT formed verifier to predict the NA probability. SAN 2.0 treats the end-of-sentence (EOS) padding of the passage as a representative position for NA, and build a binary NA classifier additionally to train the model in a joint way. U-Net[3] finds that a universal node vector encoded between the question and passage to be a powerful information collector for NA detection. However, these works totally rely on the question-aware passage to predict NA, and ignore to reveal the mechanic that why a question is unanswrable, therefore are hard to explain.

## 2 Model

An MRC problem can be typically represented by a 3-tuples $(Q, C, A)$. The $Q = q_1, q_2, \cdots, q_m$ is the question with $m$ words. $C = c_1, c_2, \cdots, c_n$ is the passage with $n$ words. $A = a_s, a_e$ when the answer exists, $A =$ NA when the question can't be answered according to the given passage, where $a_s$ and $a_e$ are the start and the end boundaries of the answer span respectively.

SCDNet is composed of three main blocks, they are Encoding, Interaction, and Prediction, as shown in Fig. 1. The encoding part encodes questions and reference text respectively, the interaction part fuses the information by attention mechanism to extract question-aware passage features and passage-aware question features. Based on the question-aware passage features, the answer prediction network predicts the answer boundary probability. Based on the passage-aware question features, the NA prediction network predicts the NA probability. The details of each part are given in the following subsections.

### 2.1 Encoding

The Encoding layer is used to transform the input word sequence into its contextual embedding. The words are first mapped into fixed word embeddings with pre-trained GloVe[12] and CoVe[13]. The embedding extraction method from document reader question answering (DRQA) is applied, part of speech (POS), named entity recognition (NER) and lemma embedding features are appended after the pre-trained word vector. Let $Q_{\mathrm{E}} = \mathbb{R}^{m \times d}$ be the question vectors and $C_{\mathrm{E}} = \mathbb{R}^{n \times d}$ be the passage vectors. Two 1D convolutional neural networks (CNNs) is deployed after the embedding layer to extract the uni-gram and bi-gram features of the word sequences, a maxout layer is used on the uni-gram and bi-gram vectors to reduce dimension as shown in Eq (1) and (2). The bi-gram matrix is padded using its last word to make it the same length as the uni-gram sequence.

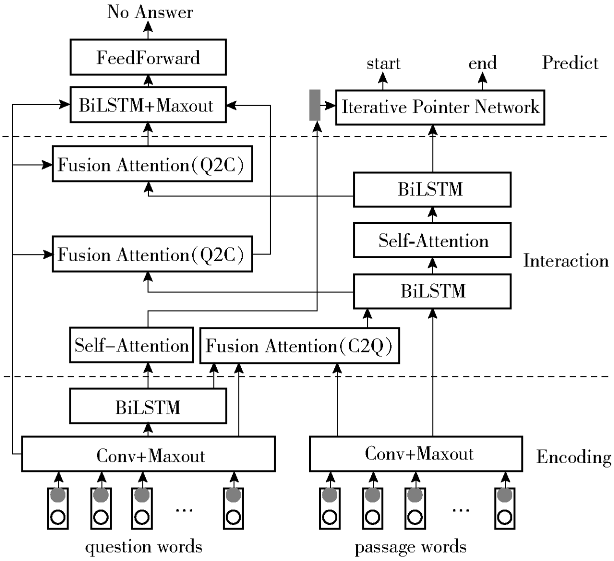


Fig. 1　Architecture of the SCDNET

$$C_{\mathrm{e\_m}} = \mathrm{maxout}(\mathrm{conv}_{\mathrm{uni}}(C_{\mathrm{E}}), \mathrm{conv}_{\mathrm{bi}}(C_{\mathrm{E}})) \tag{1}$$

$$Q_{\mathrm{e\_m}} = \mathrm{maxout}(\mathrm{conv}_{\mathrm{uni}}(Q_{\mathrm{E}}), \mathrm{conv}_{\mathrm{bi}}(Q_{\mathrm{E}})) \tag{2}$$

A BiLSTM layer is then used to encode the contextual information as shown in Eq (3).

$$Q_{\mathrm{e}} = \mathrm{BiLSTM}(Q_{\mathrm{e\_m}}) \tag{3}$$

### 2.2 Interaction

The Interaction part is composed of two layers, a fusion layer and a self-attention layer. The fusion layer is to extract question-aware passage representations, the attention mechanism is applied to fulfill the fusion process. Let $X \in \mathbb{R}^{l_x \times d_x}$, $Y \in \mathbb{R}^{l_y \times d_y}$ be the input vectors, $W_1 \in \mathbb{R}^{k \times d_x}$, $W_2 \in \mathbb{R}^{k \times d_y}$ be trainable weights matrices, $D \in \mathbb{R}^{k \times k}$ be a diagonal matrix, $[\,;\,]$ indicates the matrix/vector concatenation operator. The attention averaged question vectors are calculated based on the similarity matrix and concatenated to their corresponding passage word as shown in Eq (4) – (7), where $S_{\mathrm{C1}} \in \mathbb{R}^{m \times n}$, $S_{\mathrm{C2}} \in \mathbb{R}^{m \times n}$ are two similarity matrices calculated respectively, the similarity score function from FusionNet[14] is applied.

$$S(\boldsymbol{X}, \boldsymbol{Y}) = \mathrm{softmax}(\boldsymbol{W}_1 \boldsymbol{X} \boldsymbol{D}(\boldsymbol{W}_2 \boldsymbol{Y})^{\mathrm{T}}) \in \mathbb{R}^{l_x \times l_y} \quad (4)$$

$$S_{\mathrm{C1}} = S([Q_{\mathrm{GloVe}}; Q_{\mathrm{CoVe}}; Q_{\mathrm{e\_m}}], [C_{\mathrm{GloVe}}; C_{\mathrm{CoVe}}; C_{\mathrm{e\_m}}]) \quad (5)$$

$$S_{\mathrm{C2}} = S([Q_{\mathrm{GloVe}}; Q_{\mathrm{CoVe}}; Q_{\mathrm{e\_m}}], [C_{\mathrm{GloVe}}; C_{\mathrm{CoVe}}; C_{\mathrm{e\_m}}]) \quad (6)$$

$$Q_{\mathrm{C}} = [\boldsymbol{S}_{\mathrm{C1}}^{\mathrm{T}} Q_{\mathrm{e\_m}}; S_{\mathrm{C2}}^{\mathrm{T}} Q_{\mathrm{e}}] \quad (7)$$

At last, a BiLSTM is used to encode the contextual information as shown in Eq(8).

$$C_{\mathrm{fusion}} = \mathrm{BiLSTM}([C_{\mathrm{e\_m}}; Q_{\mathrm{C}}]) \quad (8)$$

The self-attention layer is used to capture the long-distance dependencies in the passage. The similarity matrix $S_{\mathrm{self}}$ is calculated as in Eq (10), the corresponding passage vector for every passage word is calculated as in Eq (11).

$$C_{\mathrm{H}} = [C_{\mathrm{GloVe}}; C_{\mathrm{CoVe}}; C_{\mathrm{e\_m}}; C_{\mathrm{fusion}}] \quad (9)$$

$$S_{\mathrm{self}} = S(C_{\mathrm{H}}, C_{\mathrm{H}}) \quad (10)$$

$$C_{\mathrm{self}} = \boldsymbol{S}_{\mathrm{self}}^{\mathrm{T}} C_{\mathrm{e\_m}} \quad (11)$$

At last, a BiLSTM is used to encode the contextual information as shown in Eq (12). This layer output the final passage features $H_{\mathrm{C}}$ for predicting the answer boundary.

$$H_{\mathrm{C}} = \mathrm{BiLSTM}([C_{\mathrm{fusion}}; C_{\mathrm{self}}]) \quad (12)$$

## 2.3 Prediction

### 2.3.1 No answer classification

Our model use an attention layer to collect the question's most relevant information from the two passage layers $C_{\mathrm{fusion}}$ and $H_{\mathrm{C}}$, and a BiLSTM to check whether there is a noteworthy semantic absence or a conflict as shown in Eq (13) and (14).

$$S_{\mathrm{Q1}} = S(Q_{\mathrm{e}}, C_{\mathrm{fusion}}), S_{\mathrm{Q2}} = S(Q_{\mathrm{e}}, H_{\mathrm{C}}) \in \mathbb{R}^{m \times n} \quad (13)$$

$$H_{\mathrm{Q}} = \mathrm{BiLSTM}([Q_{\mathrm{e}}; S_{\mathrm{Q1}} C_{\mathrm{fusion}}; S_{\mathrm{Q2}} H_{\mathrm{C}}]) \in \mathbb{R}^{m \times h_Q} \quad (14)$$

A maxout layer is deployed after the BiLSTM to aggregate the checking results throughout the question sequence into a $h_{\mathrm{Q}}$ dimensional vector. And based on this vector, a binary classifier is built to predict the question's NA probability, $W_{\mathrm{NA}} \in \mathbb{R}^{h_Q}$ is a trainable weight vector, $b_{\mathrm{NA}}$ is a trainable bias value as shown in Eq (15).

$$p_{\mathrm{NA}} = \mathrm{sigmoid}(W_{\mathrm{NA}} \max_m (H_{\mathrm{Q}}) + b_{\mathrm{NA}}) \quad (15)$$

The predicted answer would be set to NA if its NA probability exceeds a threshold. Experiments show that it brings 2.3 percent improvement than predicting from the EOS vector of the passage.

### 2.3.2 Answer prediction

An iterative pointer network is used as our answer network. The memory is the output feature of the passage $H_{\mathrm{C}}$, the initial hidden vector $h_{\mathrm{s},0}$ is a self-attention averaged question vector, as shown in Eq (16).

$$h_{s_0} = \mathrm{softmax}(W_{\mathrm{Q}} Q_{\mathrm{e}} + b_{\mathrm{Q}}) Q_{\mathrm{e}} \quad (16)$$

A gated recurrent unit (GRU) formed pointer network is used to refresh the hidden vector as shown in Eq (18), $h_{\mathrm{s},t}, h_{\mathrm{e},t}$ are hidden vectors of the GRU. The start and end are predicted through a bilinear function as in Eq (17) and (19).

$$p_t(s) = \mathrm{softmax}(h_{\mathrm{s},t} W_{\mathrm{s}} H_{\mathrm{C}}) \in \mathbb{R}^n \quad (17)$$

$$h_{\mathrm{e},t} = \mathrm{GRU}(h_{\mathrm{s},t}, p_t(s) H_{\mathrm{C}}),\ h_{\mathrm{s},t+1} = h_{\mathrm{e},t} \quad (18)$$

$$p_t(e) = \mathrm{softmax}(h_{\mathrm{e},t} W_{\mathrm{e}} H_{\mathrm{C}}) \in \mathbb{R}^n \quad (19)$$

$W_{\mathrm{Q}}, b_{\mathrm{Q}}, W_{\mathrm{s}}, W_{\mathrm{e}}$, are all trainable weight matrices, $p_t(s), p_t(e)$ are the probabilities of the start and end at time step $t$. The final prediction is the output at time step $T$, here $T$ is a hyperparameter.

The model is jointly trained, and the total loss is expressed as in Eq (20). $\mathrm{loss}_{\mathrm{ans}}$ is the loss item for the answer span prediction task, $\mathrm{loss}_{\mathrm{NA}}$ is for the NA prediction task, $\mathrm{loss}_{\mathrm{span}}$ is the penalty item to penalize the highest confidence answer's out-of-range length.

$$\mathrm{loss} = \mathrm{loss}_{\mathrm{ans}} + \mathrm{loss}_{\mathrm{NA}} + \mathrm{loss}_{\mathrm{span}} \quad (20)$$

The cross-entropy loss function for both $\mathrm{loss}_{\mathrm{ans}}$ and $\mathrm{loss}_{\mathrm{NA}}$ can be written as Eq (21) ~ (23) which only consider a single training example.

$$\mathrm{loss}_{\mathrm{ans}} = -\log p_T(s) - \log p_T(e) \quad (21)$$

$$\mathrm{loss}_{\mathrm{NA}} = -\log p_{\mathrm{NA}} \quad (22)$$

$$\mathrm{loss}_{\mathrm{span}} = \begin{cases} \log(\hat{e} - \hat{s} - T + 1), & \hat{e} - \hat{s} > L \\ \log(\hat{s} - \hat{e} + 1), & \hat{s} - \hat{e} > 0 \\ 0, & \text{other} \end{cases} \quad (23)$$

## 3　Experiments

### 3.1　Dataset

SQuAD 2.0 has 86 821 answerable questions and 43 498 unanswerable questions in its training data, 5 928 answerable questions and 5 945 unanswerable questions in its development data. Two metrics are used to evaluate the model performance: exact match (EM) and a marcro-averaged F1 score which measures the weighted average of the precision and recall rate at character level.

### 3.2　Implementation details

The spacy tool is utilized to tokenize all the both the question and the passage, and generate lemma, part-of-speech and named entity tags. PyTorch is used to implement our models.

The model uses word embedding with 300-dimensional GloVe and 600-dimensional CoVe word vectors, and only finetunes a 1 000 most frequent word embedding weights during training. The embeddings for the out-of-vocabulary are set to 0. All the hidden sizes of BiLSTM in the interaction layer and prediction layer are set to 300. Weight normalization is used. The dropout rate is 0.1, the mini-batch size is set to 16. Adamax optimizer is used and its learning rate is initialized to 0.001 and decrease it by 0.5 after each 10 epochs. The threshold of the NA classifier is set to 0.5. The threshold of the answer length threshold $L$ is 10. The prediction iterative step $T$ is also set to 10. The best results are reached after 13 training epochs.

In this model lots of codes and ideas are borrowed from SAN 2.0.

### 3.3　Results and analysis

SCDNet is evaluated on the development dataset of SQuAD 2.0, and it achieves a F1 score of 72.43 with GloVe and CoVe embeddings. Our model can achieve a NA accuracy of 76.96 which proves that SCDNet is good at NA prediction. This is 1.6 percent higher than our compared model SAN 2.0[4], which is also tested on the development dataset with the classifier threshold set to 0.5.

**Table 2　Experiment results**

| Configuration | All |  | NoAns |
|---|---|---|---|
|  | EM | F1 | ACC |
| BNA[2] | 59.8 | 62.6 | – |
| DocQA[2] | 65.1 | 67.6 | – |
| RMR + Verify[5] | 70.58 | 74.8 | – |
| U-Net[3] | 70.3 | 74.0 | 80.2 |
| SAN 2.0[4] | 69.27 | 72.66 | 75.3 |
| SCDNet (with CoVe) | 69.19 | 72.44 | 76.96 |

SAN 2.0 has a simple but effective MRC network structure, its network structure is easy to understand and suitable for testing the differences between the two ways of NA prediction, from the question-side or the passage-side. Based on the frame of SAN 2.0, SCDNet model is built and shows the effectiveness of predicting NA probability by detecting the semantic discrepancy between the question and the passage.

Comparing with SAN 2.0, SCDNet simplifies the 3-layer encoding BiLSTM in SAN 2.0 into 1-layer BiLSTM. SCDNet adds a bi-gram convolution layer to its encoding network, because it can help SCDNet to be more sensitive of the order of words in the texts.

U-Net and RMR + Verify in Table 2 all get very strong performances. But they are more complicated in structure than SAN 2.0 and they all use ELMo which is a more advanced pre-trained word embedding model than CoVe.

### 3.4　Ablation study

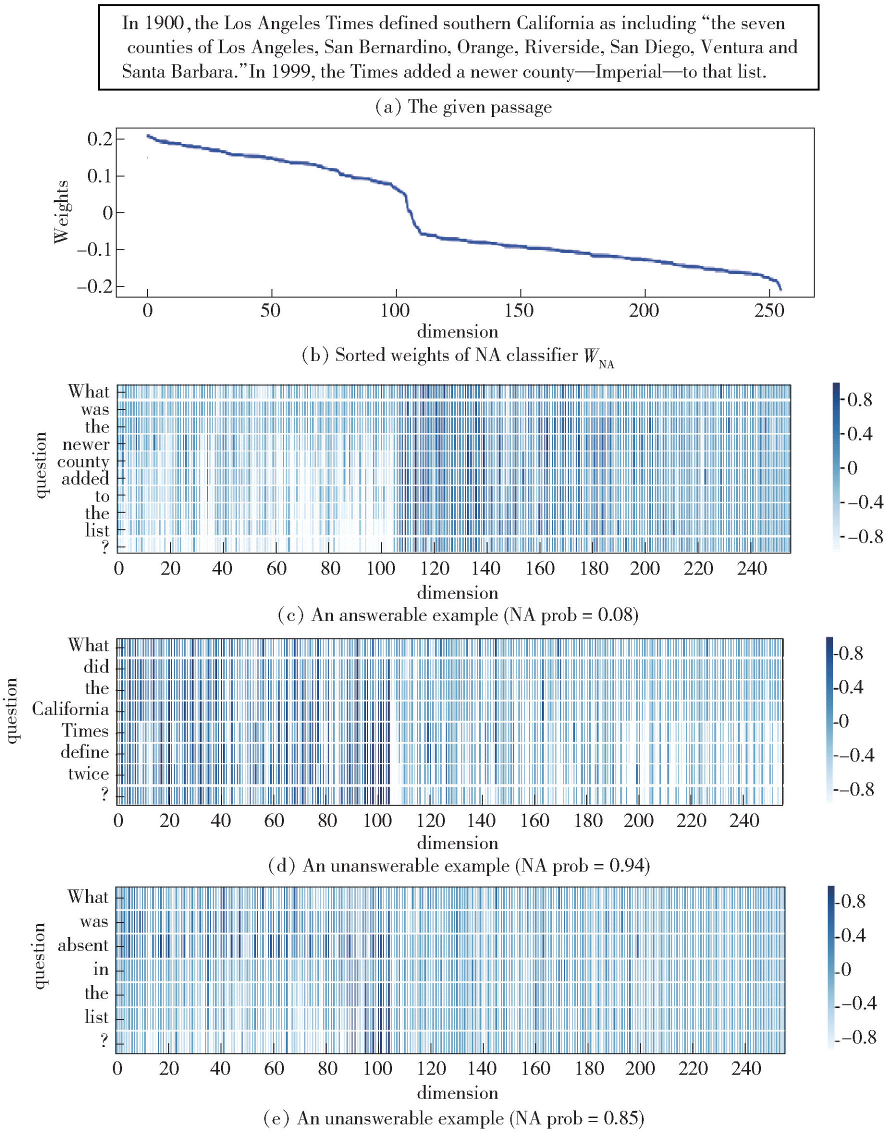
To illustrate the effectiveness of the NA classifier and answer length penalty, ablation studies are shown in Table 3. Here are five models trained on the development set: 1) SCDNet is our proposed model; 2) -length penalty model is SCDNet without length penalty; 3) -length rectification is SCDNet without the rectification on the candidate answer-span probabilities; 4) -length penalty and length rectification is an SCDNet without both the length penalty in the loss function and the probability rectification before evaluation; 5) EOS NA means to remove the question-side NA prediction part, and use the NA prediction from the appended EOS vector of the passage feature $H_C$ instead. 6) one Q2C means to remove the second last fusion attention Q2C layer.

**Table 3 Ablation experiments**

| Configuration | HasAns | | All | | NoAns |
|---|---|---|---|---|---|
| | EM | F1 | EM | F1 | ACC |
| 1) SCDNet | 64. 04 | 70. 54 | 69. 19 | 72. 44 | 76. 96 |
| 2) -length penalty | 68. 21 | 71. 46 | 68. 21 | 71. 47 | 76. 13 |
| 3) -length rectification | 60. 46 | 65. 89 | 63. 09 | 67. 03 | 76. 59 |
| 4) -length penalty and length rectification | 58. 69 | 63. 38 | 62. 69 | 66. 52 | 75. 91 |
| 5) EOS NA | 65. 20 | 69. 08 | 68. 46 | 71. 89 | 74. 69 |
| 6) one Q2C | 63. 87 | 70. 22 | 68. 33 | 71. 99 | 76. 35 |

By adding the answer-length penalty, it gives a 0. 93 percent performance boost compared to 2), the best performance is got at the threshold 10 instead of 15 although 15 is commonly used as the answer length limitation by other researchers. In addition, the answer-span probability rectification step is indispensable for our model as is shown in 3), this means that although our network improves its performance by add the answer length penalty, it still can benefit from using an extra step of probability rectification. The rectification method from SAN 2. 0 codes is applied to reduce the answers' probability by multiplying a punish factor $\frac{1}{\log(l_{\text{ans}}+1)}$ for those answers which have more than 5 word, $l_{\text{ans}}$ stonds for the answer length.

By comparing with 5), the SCDNet get a 0. 55 percent promotion in F1 and 2. 3 percent promotion in



In 1900, the Los Angeles Times defined southern California as including "the seven counties of Los Angeles, San Bernardino, Orange, Riverside, San Diego, Ventura and Santa Barbara." In 1999, the Times added a newer county—Imperial—to that list.

(a) The given passage

(b) Sorted weights of NA classifier $W_{\text{NA}}$

(c) An answerable example (NA prob = 0.08)

(d) An unanswerable example (NA prob = 0.94)

(e) An unanswerable example (NA prob = 0.85)

Fig. 2 Visualization of the sorted outputs of the NA BiLSTM layer

NA accuracy which proves the effectiveness of our question-side based conflicts detection model. Moreover, our NA prediction network can also improve the network's performance just as a jointly trained task for the EOS based NA prediction model.

The two Q2C fusion attention layers in our model are used to collect two different levels of passage information, 6) in Table 3 shows that if we only use the last passage information collection layer, the F1 score will drop about 0.45.

### 3.5 Case study

To illustrate the effectiveness of our NA prediction, the visualization of BiLSTM output $H_Q$ and maxout output $\max_m(H_Q)$ from Eq (13) are shown in Fig. 2. The BiLSTM is used as NA semantic matching network and the maxout is applied after it to detect the existence of semantic absence or conflict. The 256-dim $H_Q$ are sorted according to the weights of the NA classifier $W_{NA}$ because the elements multiplied with bigger weights have more powerful impacts on the predicting results. The NA classifier weights are sorted in a descending order as shown in Fig 2 (b). Here, 3 different question-answer cases for the same passage are given in Fig 2. Fig 2 (a) is the passage. Fig 2 (c) shows an answerable question. Fig 2 (d) shows an unanswerable case because of a semantic conflict, "California Times" is conflicted with "Los Angeles Times" in the passage. Fig 2 (e) is an unanswerable case for semantic absence of the word "absent", the word "absent" gets high values at the top 100 dimensions which can be regarded as a sign of semantic discrepancy, this discrepancy is caught by the maxout operation and be recognized by the NA classifier.

## 4 Conclusion

SCDNet, a simple yet effective network for machine reading comprehension with unanswerable questions is proposed. It incorporates a BiLSTM + maxout made semantic matching mechanism to check the semantic absence and conflicts through the question words, and get a 1.6 percent improvement in no-answer accuracy compared to our baseline model SAN 2.0. Additionally, it is very useful to add an answer-length penalty in the answer span prediction loss function, which brings about 1% improvement in F1 score.

**References:**

[1] Rajpurkar P, ZhangJian, Lopyrev K, et al. SQuAD: 100, 000 + questions for machine comprehension of text [C] // Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA, USA: Association for Computational Linguistics, 2016: 2383-2392.

[2] Rajpurkar P, Jia R, Liang P. Know what you don't know: unanswerable questions for SQuAD[C] // Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Stroudsburg, PA, USA: Association for Computational Linguistics, 2018, 2: 784-789.

[3] Sun F, Li L Y, Qiu X P, et al. U-net: machine reading comprehension with unanswerable questions[EB/OL]. 2018(2018-10-12)[2019-11-18]. https://arxiv.org/abs/1810.06638.

[4] Liu X D, Li W, Fang Y W, et al. Stochastic answer networks for SQuAD 2.0[EB/OL]. 2018(2018-09-24)[2019-11-18]. https://arxiv.org/abs/1809.09194.

[5] Hu Minghao, Wei Furu, Peng Yuxing, et al. Read + verify: machine reading comprehension with unanswerable questions[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33: 6529-6537.

[6] Wang S H, Jiang J, Learning natural language inference with LSTM[EB/OL]. 2015(2016-11-10)[2019-11-18]. https://arxiv.org/abs/1512.08849v2.

[7] Parikh A P, Täckström O, Das D, et al. A decomposable attention model for natural language inference[EB/OL]. 2016(2016-09-25)[2019-11-18]. https://arxiv.org/abs/1606.01933v2.

[8] Vinyals Q, Fortunato M, Jaitly N, Pointer networks[C] // Advances in Neural Information Processing Systems. New York: Curran Associates, 2015: 2692-2700.

[9] Hu Minghao, Peng Yuxing, Huang Zhen, et al. Reinforced mnemonic reader for machine reading comprehension[C] // Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence. California: International Joint Conferences on Artificial Intelligence Organization, 2018: 4099-4106.

（下转第 141 页）

Conference on Cognitive Informatics & Cognitive Computing（ICCI＊CC）. Beijing，China：IEEE Press，2015：399-404.

[8] Lena T，Lior R，Bracha S. Identification of label dependencies for multilabel classification[C]// Proceedings of the 2nd International Workshop on Learning from Multi-Label Data. Dublin，Ireland：IEEE，2010：53-60.

[9] Tsoumakas G，Katakis I，Vlahavas I. Random *k*-labelsets for multilabel classification[J]. IEEE Transactions on Knowledge and Data Engineering，2011，23(7)：1079-1089.

[10] Charte F，Rivera A，del Jesus M J，et al. Improving multi-label classifiers via label reduction with association rules[M]//Lecture Notes in Computer Science. Berlin，Heidelberg：Springer Berlin Heidelberg，2012：188-199.

[11] Liu Caizhi，Sheng Yanxiu，Wei Zhiqiang，et al. Research of text classification based on improved TF-IDF algorithm[C]//2018 IEEE International Conference of Intelligent Robotic and Control Engineering（IRCE）. Lanzhou，China：IEEE Press，2018：218-222.

[12] Moreno-Leon J，Robles G，Roman-Gonzalez M. Comparing computational thinking development assessment scores with software complexity metrics[C]//2016 IEEE Global Engineering Education Conference（EDUCON）. Abu Dhabi，UAE：IEEE Press，2016：1040-1045.

[13] Wang Zezhong，Cao Shuo. A power load association rules mining method based on improved FP-growth algorithm[C]//2018 China International Conference on Electricity Distribution（CICED）. Tianjin，China：IEEE Press，2018：2833-2837.

[14] Gibaja E，Ventura S. Atutorial on multilabel learning[J]. ACM Computing Surveys，2015，47(3)：1-38.

（上接第 133 页）

[10] Liu Xiaodong，Shen Yelong，Duh K，et al. Stochastic answer networks for machine reading comprehension[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics（Volume 1：Long Papers）. Stroudsburg，PA，USA：Association for Computational Linguistics，2018：1694-1704.

[11] Levy O，Seo M，Choi E，et al. Zero-shot relation extraction via reading comprehension[C]//Proceedings of the 21st Conference on Computational Natural Language Learning（CoNLL 2017）. Vancouver，USA：Association for Computational Linguistics，2017：333-342.

[12] Pennington J，Socher R，Manning C. Glove：global vectors for word representation[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing（EMNLP）. Doha：Association for Computational Linguistics，2014：1532-1543.

[13] McCann B，Bradbury J，Xiong C M，et al. Learned in translation：contextualized word vectors[C]//Advances Inneural Information Processing Systems. New York：Curran Associates，2017：6294-6305.

[14] Huang H Y，Zhu C G，Shen Y L，et al. Fusionnet：fusing via fully-aware attention with application to machine comprehension[EB/OL]. 2017（2018-02-04）[2019-11-18]. https://arxiv.org/abs/1711.0734-1v2.