A Multiple Granularity Co-Reasoning Model for Multi-choice Reading Comprehension

研究问题：多选式阅读理解
方法：多粒度；co-reasoning
idea：多粒度文本匹配模块完成三者的交互；
多语句共理性模块完成跨语句的句子推理
建模的重点是 交互特征

Hang Miao, Ruifang Liu and Sheng Gao
School of Information and Communication Engineering
Beijing University of Post and Telecommunications, Beijing, China
{miaohang123, lrf, gaosheng}@bupt.edu.cn

***Abstract*—We propose a multi-granularity co-reasoning model for multi-choice reading comprehension task, which aims to select the correct option based on the interaction between passage, question and candidate options. Firstly, we introduce a multiple granularity text matching module to interact passage with question and each option. We take advantage of information extracted from diverse semantic spaces to conduct more extensive matching between text sequences. With this help, we could better match the passage against the question and each option to gather relevant information. Furthermore, we employ a multi-sentence co-reasoning module for sentence inference across multiple sentences. Specifically, we utilize 1D Convolutional Neural Network (1D-CNN) with different kernel sizes and self-attentive Recurrent Neural Network (RNN) to model the relationships of relevant sentences. This module could better synthesize and aggregate sentence-level evidence to make decisions. Experimental results demonstrate that our proposed model achieves state-of-the-art performance for single models on the RACE dataset.**

阅读理解任务定义:像人一样理解自然语言

I. Introduction

Machine Reading Comprehension (MRC) is a challenging task which aims to make computers capable of comprehending natural language as humans. In particular, it requires an AI agent to answer questions based on a given passage. With the release of several large-scale datasets, as well as the rapid development of deep learning, great progress has been made on the field of machine reading comprehension. Now researchers could train deep learning based machine reading systems in an end-to-end manner and some of them are near human-level performance on specific datasets [1].

数据集（任务）分类：填空式、抽取式和多选式。

Basically, there are three main kinds of machine reading comprehension datasets: cloze-style dataset (e.g., CNN/Daily Mail [2]), span-based dataset (e.g., SQuAD [3]) and examination dataset (e.g., RACE [1] and MCTest [4]), which is usually in the form of multi-choice answer selection. In this paper, we mainly focus on multi-choice reading comprehension task, which aims to select correct option based on the interaction between the given passage, question and option candidates. Compared with other MRC tasks, multi-choice answer selection requires a higher level reasoning and comprehending capacity, because the candidate options are not directly extracted from the passages but human generated. For instance, in the span-based dataset SQuAD, above 70% of the questions can be addressed through word-based matching and paraphrasing, with only about 20% questions requiring deep reasoning. While in the multi-choice dataset RACE released in 2017, half of the questions involve single sentence reasoning and multiple sentence reasoning. In order to do this task well, we need to focus more on sentence-level reasoning and inference, synthesizing information and evidence across multiple sentences in the passage.

多选式难度最高，因为这类任务中的答案并不能直接在原文中提取，而是需要更多的推理

TABLE I
AN EXAMPLE WHICH REQUIRES MULTI-SENTENCE REASONING TO SELECT THE CORRECT ANSWER.

Passage: However, many parents are afraid of these young bloggers. Parents see the kids talking about how they got drunk last weekend and how they don't like studying. They are using language that is surprising to their parents. Besides hearing from their friends, teen bloggers also get message from strangers. Most of the time, it's older men asking to meet teenage girls. **"These strange men are dangerous for my kids**. They sometimes teach my kids bad words, " said Cara Cabral, a mother of two. Many teens and young adults know it's not safe to use blogs on the Internet. They know they are putting information about themselves in a place they can be seen by anyone. But teens are unlikely to give up these new communication tools that have becomes a way of life for many of them.
Question: Parents think it's dangerous for their kids to use blogs because .
Candidate Options: A. their kids use a surprising language B. their kids talk about how they don't like studying C. teen bloggers got messages from strangers D. their kids talk about their girlfriends or boyfriends.
Ground Truth Answer: C

该文章的具体方向：句子级推理；跨语句推理

先前工作的思路：词级文本匹配

Most previous researches treat multi-choice MRC task as a word-level text sequence matching problem [1] [5] [6]. Firstly, a sequence encoder is employed to represent passage, question and options. Then a word-level attention matrix is computed to match the word similarity of {passage, question} pairs and obtain question-aware contextual representations. Finally, the correct option is selected by calculating the similarity between the refined passage representation and each option representation. However, these approaches may lack the ability of sentence reasoning and inference. The matching process is based on word level, which does well on word matching and paraphrasing questions, but for cases when answer clues are far apart in the passage, this method may fail to capture the intra information of a sentence [7]. On the other hand, these work did not take into account the logical relationship of different sentences. For questions involving multi-sentence reasoning, the answer clues may not be located in a single sentence but distributed across multiple sentences, so sentence-

先前工作的具体做法：对三者编码，通过注意力机制得到段落与问句的词相似矩阵（问句感知的上下文表示），将其与各个候选答案进行匹配

这种方法的缺点：仅根据词级的相似进行匹配，缺少句子级的推理。当答案的多个支撑语句在原文中相距较远时，这种方法就不行了

978-1-7281-2009-6/$31.00 ©2019 IEEE

level reasoning and inference are needed to capture the relation and interaction of multiple sentences.

To tackle the aforementioned problems, we propose a multi-granularity co-reasoning model for text sequence matching, reasoning and inference. Instead of simple word-by-word matching mechanism based on just one representation of the text sequences, we conduct a multi-granularity text matching mechanism to match the passage against question and each option on a multiple semantic neural space. To be more specific, given the passage and question representation, we firstly employ a series of 1D-CNN with diverse kernel sizes to capture the local information of each sequence on multiple granularity. After that, each sequence could be represented by multiple representation vectors. Then a set of matching matrices of the two sequences is obtained by computing the similarity between each representation pairs on different semantic neural spaces. This is inspired by human reading strategy. When we are matching the semantic of two text sequences, we do not just compare them word by word. Instead, we will jointly consider the meaning of words and phrases of the sentences to better comprehend the compared texts.

As for multi-sentence reasoning, [8] employed a hierarchical Long Short-Term Memory (LSTM) to capture sentence-level information. [7] applied an option gating module to gather evidence from the passage for each candidate option. However, these approaches may lack the capability of reasoning across multiple sentences since they did not consider the local relationship of a certain group of sentences. When we humans are doing reading comprehension task, if we find a sentence highly related to the given question and a certain option, we tend to look forward and backward a few sentences and pay attention to the logic relationship of them. Table I illustrates an example of this strategy. According to the question, we could find the most relevant sentence of the question in the passage: "These strange men are dangerous for my kids." However, in order to choose the correct option C, we have to look forward and backward the matched sentence and jointly consider the relationships of these sentences. Motivated by this strategy, we consider to use CNN with different kernel sizes to capture the relationships of multiple sentences and gather sentence-level information to make predictions. Unlike RNN which tries to synthesize the whole structure information of a passage, CNN is capable of modeling local relationships of sentences in certain regions and help do sentence reasoning better. We compare adopting 1D-CNN to aggregate sentence-level information with self-attentive RNN [9] based method. Experimental results show that CNN based sentence reasoning module performs better than RNN based method.

To conclude, the overall contributions of our work are summarized as follows:

1) We propose a multi-granularity matching mechanism for text sequence matching and alignment, compared with word-by-word attention, our method is capable of capturing more extensive interactions of two sequences on multiple semantic representation spaces.
2) We employ a multi-sentence co-reasoning module to capture sentence-level local relationships of multiple sentences based on refined passage representation. Our method could help better tackle questions involving multiple sentence reasoning and inference.

The rest of this work is organized as follows. Section II introduces the related works of multi-choice MRC task and text sentence matching mechansim. Section III describes the overall structure of our model. Section IV summarizes the training details and main results of our proposed model.

II. R

A. Multi-choice machine reading comprehension

Multi-choice machine reading comprehension (multi-choice MRC) is a common task to evaluate humans reading and comprehending ability. Compared with cloze-style task and span-based answer extracting task, multi-choice MRC is more challenging since the ground-truth answer is not restricted to a continuous text span in the passage. Furthermore, it requires more for sentence reasoning, summarizing and inference, which is a long-standing problem in natural language processing and understanding. MCTest [4] is among the earliest released multi-choice MRC dataset with high-quality questions and candidate answers, which are human designed by crowdsourcing. However, this dataset is relatively small to train deep neural models.

RACE dataset [1] is a recently released large-scale dataset, collected from English reading comprehension examinations in Chinese middle and high schools. Compared with MCTest, RACE is more complicated and requires more reasoning and inference across multiple sentences. [1] refined two deep neural network models: Stanford Attentive Reader (SAR) [5] and Gated-Attention Readers (GA) [6] as baselines for this dataset. Both models build question-aware passage representation by attention mechanisms and the option is selected by computing the similarity between the option and refined passage representation through a bilinear attention layer. Considering that the two models did not take into account the interaction between passage and options, [8] proposed to build the interactions between passage, question and options with each other at hierarchical levels and utilized option corrections to refine option representation. [10] made use of a dynamic multiple matching strategies to fuse passage, questions and options into attention vectors and designed a multi-step reasoning for answer selection. [8] employed a co-matching strategy to jointly match the passage against the question and candidate options and incorporated a hierarchical aggregation component to capture sentence structure of the passage.

B. Sequence matching and alignment mechanisms

Sequence matching and alignment mechanisms serve as a core step for MRC task, which build the interaction of the given text sequence pairs. In fact, these mechanisms have been used on a wide range of NLP tasks such as natural language inference (NLI), paraphrase identification and answer sentence selection. Most of the sequence matching models

该模型与先前模型的不同：先前研究仅基于文本的一种表示实现词级匹配机制 该方法构造多粒度文本匹配机制

具体流程：得到三者的编码表示后，先通过多个不同大小卷积核的卷积网络进行编码，随后建立多个不同粒度的匹配矩阵。。这篇文中的粒度是指不同卷积核所提取到的信息，即卷积核多大，粒度就为多大

多语句推理先前方法的做法：LSTM捕获句子级特征；门控模块来控制信息的融合

上述方法的缺点：未能考虑局部信息的影响

受人类阅读策略的启发，应用不同卷积核的CNN来捕获局部信息，使得模型可以有效捕获相关语句间的关系信息

贡献1：提出了多粒度匹配机制，可以在多中语义空间表示中捕获更加丰富的交互特征

贡献2：提出了多语句共理模块，可以捕获多语句间的句子级局部信息

多选阅读理解的定义，及与另外两种的区别

现有方法的具体做法

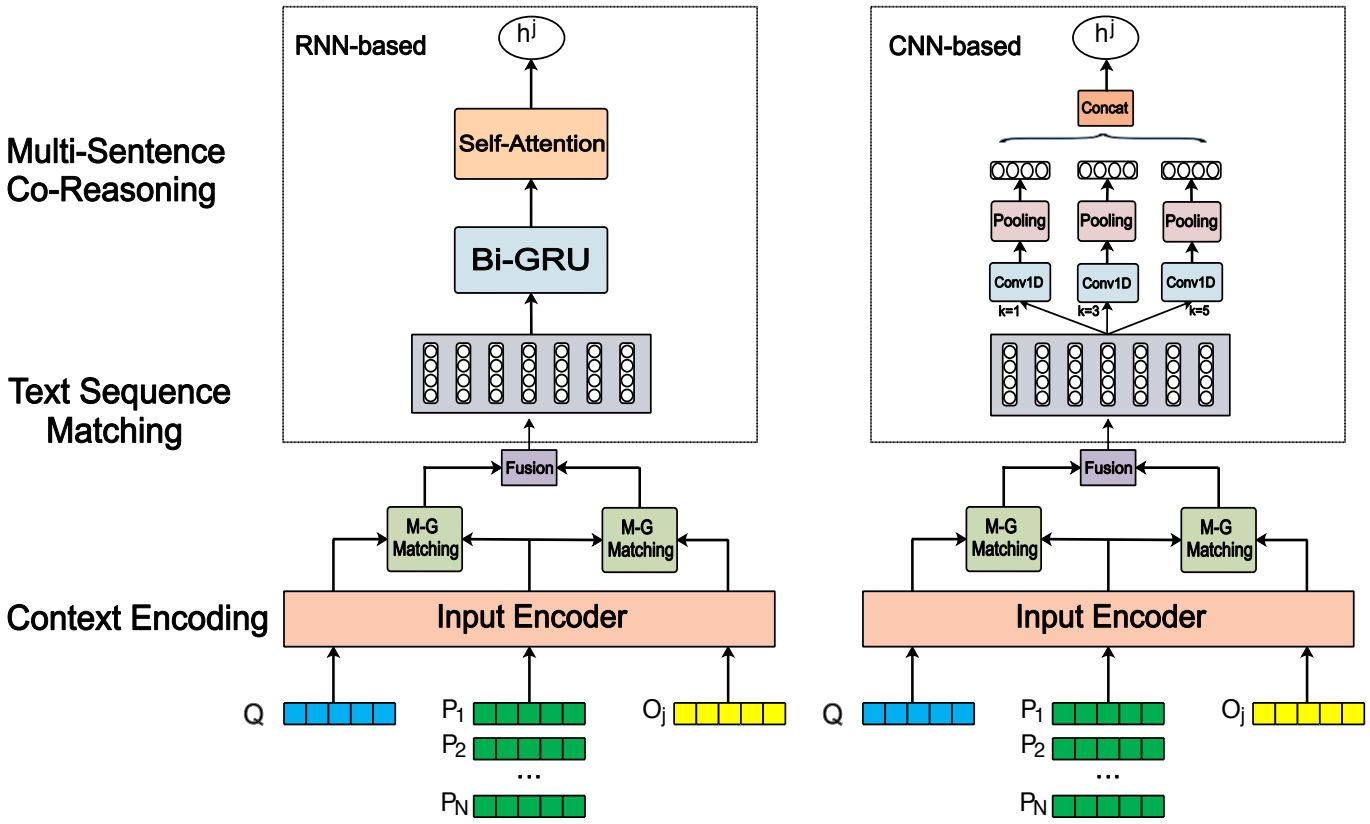

Fig. 1. An overview of our proposed model for multi-choice MRC task. Our model is composed of four main components: context encoding module, text sequence matching module, multi-sentence co-reasoning module and answer selection module (not shown in this figure). We adopt self-attentive BiGRU (left) and 1D-CNN (right) in the multi-sentence co-reasoning module to aggregate sentence-level information respectively.

are based on word-by-word attention mechanism. Typically, two sequence is encoded by either CNN or RNN. Then word-by-word attention is applied to match each state in one sequence representation to the representation of the other. Eventually we could obtain refined sentence pair encoding via soft-alignment. [11] proposed a compare-aggregate architecture to compare the similarity of two text sequences and aggregate matching features for making decisions. [12] used one BiLSTM to encode sequences and another BiLSTM to aggregate the matching information. All these approaches used one representation of the two compared sequences to complete the matching process. [13] introduced a co-stack residual affinity network to match sequence pairs, which achieves state-of-the-art performance on four kinds of text matching task. This method calculated the matching scores based on representations encoded by multiple stacked recurrent encoder and could obtain more extensive matching information. Our multiple granularity matching mechanism is inspired by this work. But we do not stack RNN to extract features from different hierarchies, instead, we use CNN with different kernel sizes to extract local features, which is able to measure the similarity of the sequence pairs on multiple semantic neural space.

现有的匹配模型的具体做法

III. MULTI-GRANULARITY CO-REASONING MODEL

In this section, we describe the overall structure of our proposed model for multi-choice MRC task. Our model is composed of four main modules: context encoding module, multi-granularity text sequence matching module, multi-sentence co-reasoning module and answer selection module, as Fig. 1 shows. In the multi-sentence co-reasoning module, we adopt self-attentive BiGRU and 1D-CNN to aggregate sentence-level information respectively. For simplification, we suppose the input of our model is a triple of passage, question and candidate options, which is denoted by $\{P, Q, O\}$. The target is to select the correct option from the candidates based on the interaction of passage, question and options.

这几篇文章的第一个共性：已有好几篇文章都是先把段落划分为句子，这是什么约定？还是单纯的只是因为这样做效果更好？

A. Context encoding layer

Given the input passage, question and candidate options, we firstly split the passage into several sentences and denote the input as a triple $\{Q, P_i, O_j\}$, where P_i denotes the i-th sentence of the passage, $i \in [1, 2, \ldots, N]$, N is the number of the sentences in the passage; O_j denotes the j-th option of the option candidates, $j \in [1, 2, 3, 4]$,

Then we convert each word in the sequences to a d-dimensional vector by pre-trained word embedding models. After that, we could obtain the word embeddings of the input

预训练嵌入+BiRU编码

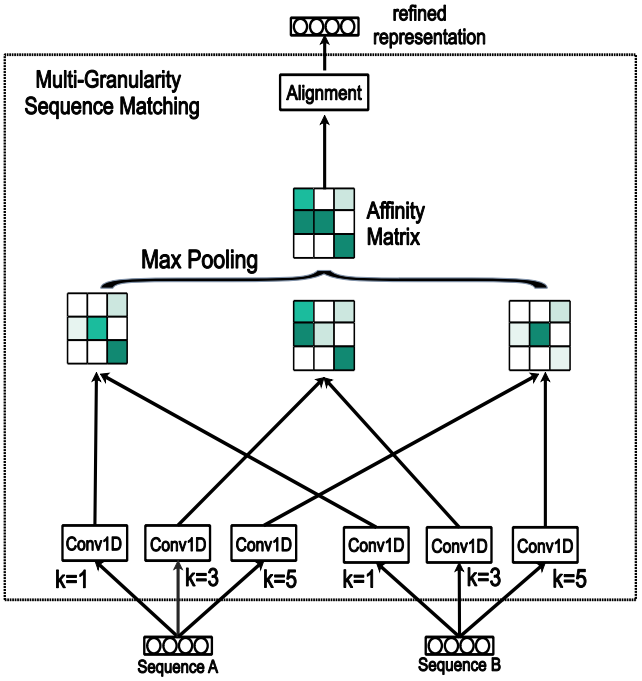

Fig. 2. Multi-granularity text sequence matching between two text sequences

triple, denoted as $\{Q^{emb}, P_i^{emb}, O_j^{emb}\}$. Next, we feed these word embeddings into a Bidirectional Gated Recurrent Unit (BiGRU) to process the word sequences and obtain contextual encodings from the surrounding words in each sequence:

$$u^{P_i} = \text{BiGRU}(P_i) \tag{1}$$

$$u^{Q} = \text{BiGRU}(Q) \tag{2}$$

$$u^{O_j} = \text{BiGRU}(O_j) \tag{3}$$

where $u^{P_i} \in \mathbb{R}^{h \times p_{len}}$, $u^{Q} \in \mathbb{R}^{h \times q_{len}}$ and $u^{O_j} \in \mathbb{R}^{h \times o_{len}}$ are the context encoding of the input triple produced by the BiGRU; p_{len}, q_{len} and o_{len} denote the max length of the i-th sentence of the passage, the question and the j-th option, respectively. h is the hidden size of BiGRU encoder.

B. Multi-granularity text sequence matching module

This module is responsible for capturing the deep interaction between two encoded sequences on multiple semantic space. We take the matching process between P_i and Q as an example to illustrate this module, as Fig. 2 shows. For brevity, we drop the subscript i of P_i. Given the contextual representation u^P and u^Q, we perform a set of one-layer 1-D CNN with diverse pre-determined kernel sizes to extract the local features of each contextual representation. Then we could obtain a series of new feature representation on diverse semantic neural space.

Typically, we use zero-padding and set the input and output channel of each 1-D CNN equally, in order to make the dimension of new representation equal to the dimension of corresponding contextual representation. For each convolution kernel, we have

$$c_k{}^P = \text{Conv1D}_k(u^P) \tag{4}$$

$$c_k{}^Q = \text{Conv1D}_k(u^Q) \tag{5}$$

where $c_k^P \in \mathbb{R}^{h \times p_{len}}$, $c_k{}^Q \in \mathbb{R}^{h \times q_{len}}$ are the local representation extracted by convolution kernel of size k. Then we compute the similarity matrices by soft attention mechanism:

$$S^k = \text{Softmax}((W_p c_k{}^Q)^T c_k{}^P) \tag{6}$$

where $S^k \in \mathbb{R}^{q_{len} \times p_{len}}$ represents the soft-alignment score matrix of the kernel size k, its element $S_{ij}{}^k$ represents the relevance score between the i-th hidden state of $c_k{}^Q$ and j-th hidden state of $c_k{}^P$; W_p is a trainable parameter. Inspired by [13], to determine on which semantic neural space the two sequences have the highest matching score, we apply a max pooling layer to select the most informative matching score across each soft-alignment matrix and get the final affinity matrix:

$$S = \text{MaxPooling}([S^1; S^2; \ldots; S^k]) \tag{7}$$

where $S \in \mathbb{R}^{q_{len} \times p_{len}}$. After that, we use this matrix to refine the passage contextual representation,

$$\overline{u^Q} = u^Q S \tag{8}$$

where $\overline{u^Q} \in R^{h \times p_{len}}$. Following a popular and effective matching trick used by [8] [13] [14] [15], we integrate the original contextual representation u^P and the refined representation $\overline{u^Q}$ by:

$$v^Q = \text{Relu}(W_c[u^P; \overline{u^Q}; u^P - \overline{u^Q}; u^P \otimes \overline{u^Q}] + b_c) \tag{9}$$

where $v^Q \in R^{h \times p_{len}}$, $\otimes$ denotes element-wise product, $W_c \in \mathbb{R}^{h \times 4h}$ and $b_c \in \mathbb{R}^h$ are trainable parameters. Inspired by the co-matching model proposed by [8], we also match the passage against the options in order to make use of the interaction between passage and each candidate option. Through this process, we could obtain the final refined representation of each sentence in the passage, based on its interaction with the question and the candidate options, denoted as v^Q and v^O, respectively.

C. Multi-sentence co-reasoning module

This module is responsible for fusing information from question-aware and option-aware passage representation, gathering information and evidence across multiple sentences in the passage. Given an input a triple $\{Q, P, O_j\}$, we could obtain two refined representation by text sequence matching module. Then the whole passage can be represented as $\{v_i^Q\}_{i=1}^N$ and $\{v_i^{O_j}\}_{i=1}^N$, based on its interaction between question and the j-th candidate option, respectively. First of all, we fuse the information from different matching results by the same sequence matching trick used in Section III-B:

$$c_i = [v_i{}^Q; v_i^{O_j}; v_i{}^Q - v_i{}^{O_j}; v_i{}^Q \otimes v_i{}^{O_j}] \tag{10}$$

where $c_i \in R^{h \times p_{len}}$. To get richer aggregated information of the sentence representation, we apply a max pooling and a

这几篇顶会文章的第二个共性：
编码结构很简单（比如这里只用了一层BiGRU）
也许这个任务的重点并不是编码器的设计，而是三者之间的交互

从这里开始，下边的操作不再是三者对称的。
这里用多个卷积网络先对篇章和问题表示进行编码

这里理解错误，依然是对称处理，只不过这里只是拿其中两者进行举例

在多个注意力矩阵中，选出信息量最大的那个

将问句感知的文章表示和答案感知的文章表示进行融合
融合方式与上文一样

paper N-19172.pdf

TABLE II
STATISTICS OF THE RACE DATASET.

Dataset	RACE-M	RACE-H	RACE
Passage numbers(train/dev/test)	6409/368/362	18728/1021/1045	25137/1389/1407
Question numbers(train/dev/test)	25421/1436/1436	62445/3451/3498	87866/4887/4936
Average Sentence Length in the Passage	17.2	19.2	18.7
Average Passage Length	231.1	353.1	321.9
Average Question Length	9.0	10.4	10.0
Average Option Length	3.9	5.8	5.3
Single-sentence reasoning ration	31.3%	34.1%	33.4%
Multi-sentence reasoning ration	22.6%	26.9%	25.8%

融合后应用两种池化，并联接结果

average pooling layer on top of c_i and concatenate the pooling results by:

$$m_i = [\text{Maxpooling}(c_i); \text{AvgPooling}(c_i)] \tag{11}$$

where $m_i \in R^{2h}$, $i = 1, 2, \ldots, N$. We apply two kinds of multi-sentence co-reasoning methods for capturing and aggregating high-level information of sentence vectors.

a) Self-attentive RNN based sentence reasoning: We use BiGRU to capture the long-range dependencies of the passage:

拆开写的注意力机制

$$s^j = \text{BiGRU}([m_1, m_2, \ldots, m_N])) \tag{12}$$

where $s^j \in R^{2h \times N}$. Instead of using pooing operation to aggregate sentence-level information, we perform a self-attention mechanism to extract semantic aspects of the passage:

$$a^j = \text{Softmax}(W_{s_2} tanh(W_{S_1} s^j))) \tag{13}$$

where $a^j \in R^{1 \times N}$ represents the attention vectors, $W_{s_1} \in R^{2h \times 2h}$ and $W_{s_2} \in R^{1 \times 2h}$ are weight matrices. Finally we sum up the hidden state of s^j according to the attention vectors:

$$h^j = a^j {s^j}^T \tag{14}$$

where $h^j \in R^{2h}$ is the final output vector based on the interaction of the passage, question and the j-th option.

b) CNN based sentence reasoning: In order to capture the local relations of relevant sentences in the passage, we employ a set of 1-D CNNs with diverse kernel sizes and a max pooling layer to extract local features of the sentence-level passage representation:

$$h_k^j = \text{Maxpooling}(\text{Conv1D}_k([m_1, m_2, \ldots, m_N])) \tag{15}$$

where $h_k^j \in R^{2h}$ denotes the sentence-level aggregated passage representation of kernel size k, $k \in [1, 3, 5]$. CNN with different kernel sizes is able to model the relationships of sentences in different ranges. After that, the final passage representation h^j is obtained by concatenation of h_k^j:

$$h^j = [h_1^j; h_3^j; h_5^j] \tag{16}$$

where $h_k \in R^{6h}$.

多卷积联接

D. Option selection module

全连接；多分类

For each candidate option O_j, we feed the final output vector h^j into a fully connected layer to get the matching score of this option:

$$r^j = W_P h^j \tag{17}$$

where W_P is a trainable parameter. Then a softmax layer is added to get the probability of option O_j being correct:

$$Pr(O_j|P, Q, O) = \frac{exp(r^j)}{\sum_{i=1}^{4} exp(r^i)} \tag{18}$$

The whole model is trained by minimize the negative likelihood loss function through back propagation.

IV. EXPERIMENT

In this section, we firstly describe the dataset used to train and test the performance of our model. Then we make a brief introduction of existing baselines and state-of-the-art models for multi-choice MRC task. Finally we conduct experiments on the aforementioned dataset to evaluate the performance of our model. Experimental results demonstrate that our proposed method outperforms all the compared baselines. Additionally, we carry out a ablation study to analyze the contributions of each module of our proposed model.

A. Dataset

介绍数据集

We mainly focus on the RACE dataset to train and evaluate our model. RACE is a popular multi-choice MRC dataset collected from English reading examinations in China. Each passage is associated with several questions and each question only has one correct answer. This dataset consists of two parts: high school part and middle school part, denoted as RACE-M and RACE-H. The main difference between RACE-M and RACE-H is that RACE-H contains more questions involving sentence reasoning and inference, which makes it more complex than RACE-M. The statistics of this dataset is shown in Table II.

B. Experimental setups

这个数字也与AAAI2019的一致

We firstly split the passage into sentences. The maximum number of sentences in each passage and maximum number of words in each sentence are both set to 50. While the maximum sequence lengths of question and each candidate option are fixed to 25. Then we use NLTK to do the word stemming and

这个与AAAI2019的不一样

文章划分为句子，每篇最多不超过50个句子，每个句子最多不超过50个词；
问题和选项最长不超过25个词；
预训练Glove300维嵌入；
词表为50000；
GRU的dropout为0.25，全连接的dropout为0.5；
GRU的神经元数量为128；
卷积核尺寸为[1,3,5]
最大池化为按行池化；
Adam为优化器；
学习率为0.001；
batch为32

TABLE III
PERFORMANCE COMPARISON OF ALL PUBLISHED SINGLE MODELS.

Single Models	RACE-M	RACE-H	RACE
Random	24.6	25.0	24.9
Sliding Window [4]	37.3	30.4	32.2
Stanford AR [5]	44.2	43.0	43.3
GA Reader [6]	43.7	44.2	44.1
ElimiNet [16]	44.5	44.5	44.5
Hierarchical Attention Flow [17]	45.3	44.2	44.1
Dynamic Fusion Network [10]	51.5	45.7	47.4
Hierarchical Co-Matching [8]	55.8	48.2	50.4
BiAttention + Simple MRU [18]	57.7	47.5	50.4
Multiple Granularity Co-Reasoning (RNN based)	57.7	48.4	51.1
Multiple Granularity Co-Reasoning (CNN based)	**58.4**	**48.7**	**51.5**
Turkers [1]	85.1	69.4	73.
Ceiling [1]	95.4	94.2	94.5

lemmatization. The word embedding matrices are initialized by 300-dimensional Glove embeddings [19] pre-trained on Wikipedia. Empirically, the size of our vocabulary is fixed to $50k$. The word embedding vectors of unknown words are set to zero vectors. In order to avoid over-fitting, we add a recurrent dropout of 0.25 to each GRU cell and a dropout of 0.5 to each linear layer. Additionally, the hidden size of BiGRU is set to 128. In the text sequence matching module, we choose a range values of [1, 3, 5] for the convolution kernel size. While in the CNN based sentence reasoning module, we also set the kernel size as [1, 3, 5]. And the maxpooling function used in this module is the row-wise max pooling operation. Our model is trained by adopting Adam optimizer with an initial learning rate of 10^{-3} and the batch size is set to 32. For fair comparison, we do not use external contextualized embeddings such as ELMo [20], GPT [21] and so forth.

C. Compared Method

大同小异

We compare the performance of our model with the following baselines and state-of-the-art models.

Stanford AR [5]. The question-aware passage representation is obtained by soft attention mechanism and a bilinear-attention is adopted to calculated the matching score between the summarized passage representation and each option representation.

GA Readers [6]. A gated mechanism is introduced to iteratively extract and refine the passage representation based on its interaction with the question.

ElimiNet [16]. An elimination module is introduced to remove irrelevant options and then the passage representation is refined by iteratively interacted with uneliminated options.

Hierarchical Attention Flow [17]. Interactions between passage, question and options are calculated at hierarchical levels and option correction is utilized to refine option representation.

Dynamic Fusion Network [10]. A dynamic multiple matching strategies is employed to fuse passage, questions and options into attention vectors and a multi-step reasoning is used to select the correct option.

Hierarchical Co-Matching [8]. The passage is jointly matched against the question and options to obtain co-matching states of the passage, and a hierarchical LSTM layer is applied to aggregate sentence-level information.

BiAttention + Simple MRU [18]. An sentence encoder with multi-ranged gates is added to a bidirectional attention to integrate passage, question and candidate options.

TABLE IV
RESULTS OF ABLATION STUDY.

MGCR Single Model	RACE-M	RACE-H	RACE
Full model (CNN based)	**58.4**	**48.7**	**51.5**
- Multi-granularity matching	57.0	48.0	50.6
- Multi-sentence co-reasoning	54.9	47.3	49.4

D. Main Results

Accuracy is used to evaluate the model performance on RACE dataset. We run our model 5 times with randomized initialization of the parameters and report the average performance. Experimental results of our model and all the compared models are summarized in Table III. We report the results of compared approaches from respective original papers. Here we just include the performance of single models. As is shown in III, our model outperforms all the compared single models on RACE-M, RACE-H and RACE, which demonstrates the effectiveness of our method. On RACE-H dataset which requires a higher ability of sentence reasoning and inference, our model outperforms Hierarchical Co-Matching and BiAttention with simple MRU encoder by 0.5% and 1.2%, respectively. We also find that using CNN based sentence reasoning method could achieve slightly better performance than self-attentive RNN based method. This indicates that using CNN to capture local relations across relevant sentences is more helpful to gather sentence-level evidence for making decisions on this dataset.

E. Ablation Study

We conduct a ablation study to determine the contributions of each component of our model, as illustrated in Table IV. We mainly focus on investigating the influence of two modules: multi-granularity matching module and multi-sentence co-reasoning module. Firstly, we remove the multi-granularity

matching module and calculate the refined passage representation by word-level attention mechanism. We note the accuracy decreases by 0.9 percentage on RACE. Secondly we remove the multi-sentence co-reasoning module and treat the passage as a plain sequence instead of splitting it into sentences. After text sequence matching module, we just concatenate the two refined passage representation and use a pooling operation to get the final output vector. A decrease of 2.1 percentage on the RACE is reported, which shows that our multi-sentence co-reasoning module does strength the overall performance .

他这里不划分句子也能达到49，说明我的模型不是这个地方的问题。。也就是说，肯定是交互过程太简单导致的准确率上不去。

V. Conclusion

In this paper, we introduce a multiple granularity co-reasoning model to tackle the sentence matching and reasoning problems of multi-choice MRC task. We propose a multi-granularity sequence matching module for text sentence matching. This module is capable of matching two text sequences on different semantic neural spaces. In this way, we could better match the passage against the question and options to gather relevant information. Furthermore, we design a sentence-level co-reasoning module for sentence inference across multiple sentences. We employ 1D CNN with diverse kernel sizes and self-attentive RNN to model the relationships of multiple refined sentence vectors. Experimental results show our model outperforms all the compared baselines and achieves the state-of-the-art performance for single models on the RACE dataset. In the future, we will focus on multiple sentence reasoning and try to improve the sentence inference capability of deep neural networks.

Acknowledgment

This work was supported by National Natural Science Foundation of China (61702047,61300080).

References

[1] G. Lai, Q. Xie, H. Liu, Y. Yang, and E. H. Hovy, "Race: Large-scale reading comprehension dataset from examinations," *empirical methods in natural language processing*, pp. 785–794, 2017.

[2] K. M. Hermann, T. Kocisk, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, and P. Blunsom, "Teaching machines to read and comprehend," *neural information processing systems*, pp. 1693–1701, 2015.

[3] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "Squad: 100,000+ questions for machine comprehension of text," *empirical methods in natural language processing*, pp. 2383–2392, 2016.

[4] M. Richardson, C. J. C. Burges, and E. Renshaw, "Mctest: A challenge dataset for the open-domain machine comprehension of text," pp. 193–203, 2013.

[5] D. Chen, J. Bolton, and C. D. Manning, "A thorough examination of the cnn/daily mail reading comprehension task," *meeting of the association for computational linguistics*, vol. 1, pp. 2358–2367, 2016.

[6] B. Dhingra, H. Liu, Z. Yang, W. W. Cohen, and R. Salakhutdinov, "Gated-attention readers for text comprehension," *meeting of the association for computational linguistics*, pp. 1832–1846, 2017.

[7] X. Lin, R. Liu, and Y. Li, "An option gate module for sentence inference on machine reading comprehension," in *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. ACM, 2018, pp. 1743–1746.

[8] S. Wang, M. Yu, J. Jiang, and S. Chang, "A co-matching model for multi-choice reading comprehension," *meeting of the association for computational linguistics*, vol. 2, pp. 746–751, 2018.

[9] Z. Lin, M. Feng, C. N. D. Santos, M. Yu, B. Xiang, B. Zhou, and Y. Bengio, "A structured self-attentive sentence embedding," *international conference on learning representations*, 2017.

[10] Y. Xu, J. Liu, J. Gao, Y. Shen, and X. Liu, "Dynamic fusion networks for machine reading comprehension," *arXiv preprint arXiv:1711.04964*, 2017.

[11] S. Wang and J. Jiang, "A compare-aggregate model for matching text sequences," *international conference on learning representations*, p. 1, 2017.

[12] Q. Chen, X. Zhu, Z. Ling, S. Wei, H. Jiang, and D. Inkpen, "Enhanced lstm for natural language inference," *meeting of the association for computational linguistics*, vol. 1, pp. 1657–1668, 2017.

[13] Y. Tay, A. T. Luu, and S. C. Hui, "Co-stack residual affinity networks with multi-level attention refinement for matching text sequences," *empirical methods in natural language processing*, pp. 4492–4502, 2018.

[14] W. Wang, M. Yan, and C. Wu, "Multi-granularity hierarchical attention fusion networks for reading comprehension and question answering," *meeting of the association for computational linguistics*, vol. 1, pp. 1705–1714, 2018.

[15] S. Wang, M. Yu, J. Jiang, W. Zhang, X. Guo, S. Chang, Z. Wang, T. Klinger, G. Tesauro, and M. Campbell, "Evidence aggregation for answer re-ranking in open-domain question answering," *international conference on learning representations*, p. 1, 2018.

[16] S. Parikh, A. Sai, P. Nema, and M. M. Khapra, "Eliminet: A model for eliminating options for reading comprehension with multiple choice questions," pp. 4272–4278, 2018.

[17] H. Zhu, F. Wei, B. Qin, and T. Liu, "Hierarchical attention flow for multiple-choice reading comprehension," pp. 6077–6085, 2018.

[18] Y. Tay, L. A. Tuan, and S. C. Hui, "Multi-range reasoning for machine comprehension." *arXiv: Computation and Language*, 2018.

[19] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," pp. 1532–1543, 2014.

[20] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. G. Clark, K. Lee, and L. S. Zettlemoyer, "Deep contextualized word representations," *north american chapter of the association for computational linguistics*, vol. 1, pp. 2227–2237, 2018.

[21] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," *URL https://s3-us-west-2. amazonaws. com/openai-assets/research-covers/languageunsupervised/language understanding paper. pdf*, 2018.