

多任务学习增强的多视角语义特征融合网络

陈泽林^{1,2}, 陈羽中^{1,2*}

¹ 福州大学 计算机与大数据学院 福州350116

² 福建省网络计算与智能信息处理重点实验室 福州 350116

yzchen@fzu.edu.cn

摘要: 本论文针对CCKS2021:医疗科普知识答非所问识别任务提出了多任务学习增强的多视角语义特征融合网络MVMT(Multi-view Semantic Feature Fusion Network enhanced by Multi-task Learning)。答非所问识别任务是自然语言推理任务与对话系统的结合,需要模型根据对话的问题推理出不符合问题的回答。本文提出的MVMT架构从加强预训练语言模型的推理能力和对话理解能力两个方向出发,构建了多视角的语义特征。模型首先采用多任务学习策略根据推理任务的特点加强预训练语言模型的推理能力,并提出多层次匹配模块加强学习问题与回复的关系。此外,我们提出了多视角表示聚合机制来提取融合各个类别的语义特征,充分发挥了各个模块的优势。我们在答非所问任务的两个评测阶段均对MVMT架构进行了实验评估,在第一轮的评测中取得了任务的第四名,并且MVMT架构在最终评测阶段的测试中F1分数达到了69.78%,取得了任务第三名的成绩。

关键词: 自然语言推理; 对话系统; 多任务学习; 问答匹配;

* 通讯作者(Corresponding author)

Multi-view Semantic Feature Fusion Network enhanced by Multi-task Learning

Zelin Chen^{1,2}, Yuzhong Chen^{1,2*}

¹ College of Computer and Data Science, Fuzhou University, Fuzhou 350116

²

Fujian Provincial Key Laboratory of Networking Computing and Intelligent Information Proces-

sing, Fuzhou 350116

yzchen@fzu.edu.cn

Abstract: This paper presents the Multi-view Semantic Feature Fusion Network enhanced by Multi-task Learning (MVMT) for the CCKS2021: irrelevant answer recognition in medical scientific knowledge task. The irrelevant answer recognition task is a combination of the natural language inference and the dialogue system, which requires the model to infer the irrelevant answer from the question in dialogue. The proposed MVMT architecture constructs the multi-view semantic features to strengthen the inferential capability and the dialogue understanding ability of the pre-trained language model. We first adopt the Multi-task Learning Strategy to enhance the inferential capability of the pre-trained language model, and further propose the Hierarchical Matching Module to stimulate the model's learning of the relationship between question and response. Moreover, to exploit all modules' advantages to the full, we propose the Multi-view Representations Aggregation mechanism to extract and fuse the semantic features of each category. We have conducted model evaluation on both evaluation stages of the irrelevant answer recognition task. The experimental results show that our proposed solution ranks fourth place in the first evaluation stage, and further achieves an F1-score of 69.78% in the final evaluation stage, which ranks third place in the task.

Keywords: Natural Language Inference; Dialogue System; Multi-task Learning; Question/Answer Matching;

1 引言

CCKS2021: 医疗科普知识答非所问识别任务¹的目标是识别一段医疗科普文本是否为特定问题的正确回答, 该任务是自然语言推理任务(NLI)与对话系统(Dialogue System)的结合。自然语言推理任务(NLI)是自然语言处理领域中的一项重要任务, 任务需要模型根据一个句子推理出它与目标句的语义关联性, 语义关联性可以分为蕴含、矛盾、中立三种。而对话系统是现在自然语言处理研究领域的热点方向, 任务需要模型理解对话问题中的语义信息, 结合上下文为用户给出合适的回复。一个优秀的对话系统能够大幅降低电商客服、教育辅导等领域的人力成本, 更加高效准确的响应用户提出的问题。答非所问识别赛题需要模型在学习医疗问答数据的同时, 掌握医疗问答文本的特点, 深度理解问题与回复之间的语义关联性, 判断回复能否由特定问题推理得出, 这大大增加了任务的难度。无论是自然语言推理任务领域, 还是对话系统中的优秀算法, 都不能够很好的独立完成答非所问识别任务。

表1 医疗科普知识答非所问识别任务的数据样例

种类	问题	描述	回复
1	脸总是起痘是闭口粉刺, 还总是泛红怎么治疗呢?	脸总是起痘泛红是闭口粉刺, 该用什么药	皮肤光亮泛红, 很可能是由化妆品过敏引起的过敏, 如果脸部表皮薄, 可能有皮肤冲洗过敏, 所以应停止使用化妆品, 然后应用炉甘石洗剂, 也可以采取抗过敏药物治疗内部, 并避免辛辣刺激食物, 别熬夜, 多喝水, 没特效药, 注意生, 可用扑尔敏试试。
2	气管狭窄是怎么回事?	我老公最近胸闷特别厉害, 而且咳嗽不断去医院检查, 是右上叶支气管狭窄, 担心是癌症。右上叶支气管狭窄是癌症吗?	一般良性的狭窄, 多数可以考虑用支气管镜下球囊扩张的治疗方法。如果是短暂的狭窄, 比如气管异物引起的肉芽导致的狭窄, 把气管异物通过支气管镜取出来之后, 就可以看肉芽很快消失。如果是恶性狭窄, 恶性狭窄主要是指肿瘤, 这种情况的治疗就更加复杂, 除了肿瘤是否能够切除, 是否能够外科治疗之外, 还要考虑支气管镜的介入治疗。

¹ https://www.biendata.xyz/competition/ccks_2021_tencentmedical_2/

此外，答非所问识别赛题还为其数据集中的答非所问数据进行了划分，一共划分为两种，如表1所示，第一种答非所问的样本中，问题的主题是“脸上长痘的治疗”，而回答却偏向于化妆品的过敏问题，回复与问题的关联度很低，其关键词重合较少。第二种答非所问的样本中，患者的问题是询问病因，而回答偏向于该病的治疗方法，尽管同样是文不对题，但是关键词的重合度很高，问题与回复也具有较强的关联性。这种高关联性的答非所问样本会让一些依靠关键词匹配的机器学习算法失效，也会大大限制目前最先进的对话、推理算法的准确度。

综上，为了解决上述的问题，构建更加完备的答非所问识别模型，我们从加强模型推理能力、捕捉问题与回复的匹配关系等角度出发，提出了多任务学习增强的多视角语义特征融合网络MVMT (Multi-view Semantic Feature Fusion Network enhanced by Multi-task Learning)。针对上述第二种（高关联度样本）所体现出来的困难，我们在基于预训练语言模型微调的基础之上，提出了多任务学习策略MLS(Multi-tasks Learning Strategy)来让预训练语言模型分别从问题(question)和描述(description)对回复(answer)进行推理，从两个角度充分加强预训练模型对于医疗问答数据的推理能力；针对现有推理和对话算法难以独立完成答非所问任务的问题，我们提出了多层次匹配模块HMM(Hierarchical Matching Module)，将对话匹配领域中常用的学习文本关联度的匹配方法与预训练模型结合起来，加强了模型对于对话和推理两个任务的理解能力；此外，我们提出了多视角表示聚合MRA(Multi-view Representations Aggregation)机制，能够从上述模块中提取出多种特征并进行聚合，充分利用各个模块的优势，特征可分为四种，分别是全局视角(global-view)，问题视角(question-view)，描述视角(description-view)，匹配视角(matching-view)。总体来说，我们的贡献有四点：

第一，我们提出了能够加强预训练语言模型推理能力的多任务学习策略(MLS)，能够让预训练语言模型以自监督学习的方式，在主任务中学习推理信息的同时，分别地从问题、描述推理回复，促进预训练语言模型学习到更多可用于推理的输出。

第二，我们提出了能够同时加强对话学习和推理能力的多层次匹配模块(HMM)，将传统的对话匹配方法与预训练语言模型相结合，先层次化地学习问题上下文信息，理解问题的真正语义，再通过多种匹配方法学习问题与回复的关系，促进模型对于问答数据的理解。

第三，我们提出了能够同时挖掘预训练语言模型、多任务学习策略、多层次匹配模块三个部分潜力的多视角表示聚合(MRA)机制。MLS方法使用不同位置的特征表示分别学习问题角度和描述角度的推理信息，这使得我们可以根据不同的位置从预训练语言模型中分别提取出问题视角(question-view)和描述视角

(description-view)的信息。进而我们可以提取HMM模块的输出，获得匹配视角(matching-view)的信息。最后，将上述三种视角的特征表示与带有全局视角(global-view)表示的预训练语言模型标签[CLS]的表示进行结合，促进模型从多方位理解答非所问识别任务。

第四，我们将MVMT网络在CCKS2021:医疗科普知识答非所问识别任务的最终测试集上进行了测试，其中F1分数达到了69.78%，取得了赛事第三名的成绩。

2 相关工作

2.1 自然语言推理任务

自然语言推理任务是自然语言处理领域中的一个重要分支，早期的算法中，主要以统计学习的方法为主，但是需要人工构建特征的机器学习方法难以取得良好的效果，在高质量的大规模推理数据集Stanford NLI (SNLI)和Multi-Genre NLI (MultiNLI)推出之后，研究人员将注意力转向深度学习方法，衍生了大量基于深度学习的方法，均取得了不错的成果。基于深度学习的算法主要可以分为两个类型，分别是句子编码型和句子交互型。

句子编码型

句子编码型的方法将推理任务中的前提句与假设句编码为语义向量，然后对两个语义向量之间的语义距离进行计算，最后通过语义距离来判断推理任务的类别(蕴含、矛盾、中立)[1-4]。Talman等人[1]使用层次化的Bi-LSTM[5]来编码前提句与假设句，并将编码后的两个结果使用拼接、相减、点积等手段进行处理，最后使用全连接层对处理后的向量进行分类；Shen等人[3]将强化学习的方法与软硬注意力结合起来对前提句和假设句进行编码，更准确的学习了语句的上下文信息。这些工作从构建句子语义信息的角度入手，能够理解前提句、假设句的语义，但是忽视了对两个句子之间关系的构建，在模型最后计算语义距离的方法难以让模型学习到前提句与假设句之间真正的关系。

句子交互型

针对上面的问题，研究人员开始尝试在编码阶段就对前提句与假设句之间的关系进行学习，提出了句子交互型的自然语言推理任务算法[6-10]。Chen等人[7]提出局部推断(local inference modeling)的思想，在使用Bi-LSTM编码之后，使用提出的可分解的注意力机制(Decomposable Attention)对前提句和假设句计算词粒度上的注意力分数，更好的从词级别上学习了前提句与假设句的关系；

Tan等人[10]在Chen等人工作的基础上进一步加强了前提句与假设句的交互深度,提出需要使用多种注意力机制更加细致的从词粒度计算语义相关度,并使用了效率更高的双向GRU[11]来编码句子。这些工作加强了词语级别的交互深度,解决了句子编码型的缺陷,但是使用RNN算法分别编码前提句与假设句的方式难以学习全局的文本信息,且这些方法往往使用Word2Vec[12]、Glove[13]这样的静态词向量,难以解决一词多义的问题。

近年来,随着预训练语言模型技术的发展,自然语言推理任务研究进入了新的阶段,BERT[14],Roberta[15],ELECTRA[16],MacBERT[17]等预训练模型在提出之时均针对自然语言推理任务进行微调来验证它们的效果,实验结果显示直接将预训练语言模型针对自然语言推理任务进行微调得到的实验结果就能够大幅度超越传统句子交互型算法。预训练语言模型对全局信息进行深度编码的方法解决了句子交互型算法中存在的问题,为研究人员提供了新的思路。

2.2 对话系统

对话系统是自然语言处理领域中的热点方向,对工业界与学术界都有着重要的研究价值,目前的对话系统主要有生成式对话和检索式对话两种。生成式对话能够根据一段对话独立的生成回复,生成的结果丰富多样,而检索式对话则需要从语料库中进行检索,得到的结果较为单一。但与此同时,现有检索式对话得到的回复相较于生成式对话算法而言更加准确,更受工业界的青睐,广泛应用于阿里小蜜,Siri,小爱同学等对话系统中。检索式对话在检索的过程中需要学习上下文与回复之间的关系,从对话的角度理解一个回复是否合适,与答非所问识别任务更加相符,所以我们在这里进行重点介绍。

检索式对话算法可以分为单轮对话与多轮对话,早期的工作中主要针对单轮对话,以一些基于统计学习的短文本的匹配算法为主。2013年后,研究人员开始尝试使用深度学习的方法来编码对话文本的上下文,使用多层感知机MLP(Multi-Layer Perceptron),卷积神经网络CNN(Convolutional Neural Network)等方法从一元语法(unigram)、二元语法(bigram)等粒度信息中进行学习。但是,单轮对话的文本长度较短,不充分的上下文导致可用于推理的信息太少,因此研究人员将研究重心转向了多轮对话。多轮对话算法中,早期的工作通过将上下文和回复进行匹配来提取重要的信息[18-20]。Wu等人[19]提出的语义匹配模型使用GRU对每一个句子分别进行编码,并基于词粒度的表示和句粒度的表示构建了两两匹配策略;Yuan等人[20]提出的多跳选择模型采用DAM[21]中提出的注意力机制进行编码,并设计了一个多跳选择器来过滤与问题不相关的上下文,在模型学习的过程中加强上下文尾部句子对于回复的影响。总体上说,上述方法能够学习多轮对话中复杂的上下文信息,但是对于全局信息的学习还是有所欠缺,在对话轮数

多的场景下,单独编码多个句子后再进行结合的方法会为全局信息的学习带来噪声。

在近期,同样出现了许多与BERT, Roberta, ELECTRA等预训练语言模型相结合的工作[22-25]。Whang等人[22]提出的BERT-VFT模型,在BERT上针对检索式多轮对话任务进行微调,并且将领域适应策略应用于多轮对话任务中,让预训练语言模型在微调前就先理解对话数据的语言风格。BERT-VFT在 $R_{10}@1$ 上达到了85.8%,以5.8%的优势大幅度超过了基于匹配方法的MSN,体现了预训练语言模型强大的语言理解能力。Xu等人[25]在BERT-VFT的基础之上,设计了四个子任务,分别是对话段落预测(Next Session Prediction),话语修复(Utterance Restoration),不连贯对话检测(Incoherence Detection),对话一致性鉴别(Consistency Discrimination),这些子任务通过恢复扰乱过的对话数据来从中学习到对话数据中应该具有的特点,取得了显著的提升。

3 模型

3.1 概述

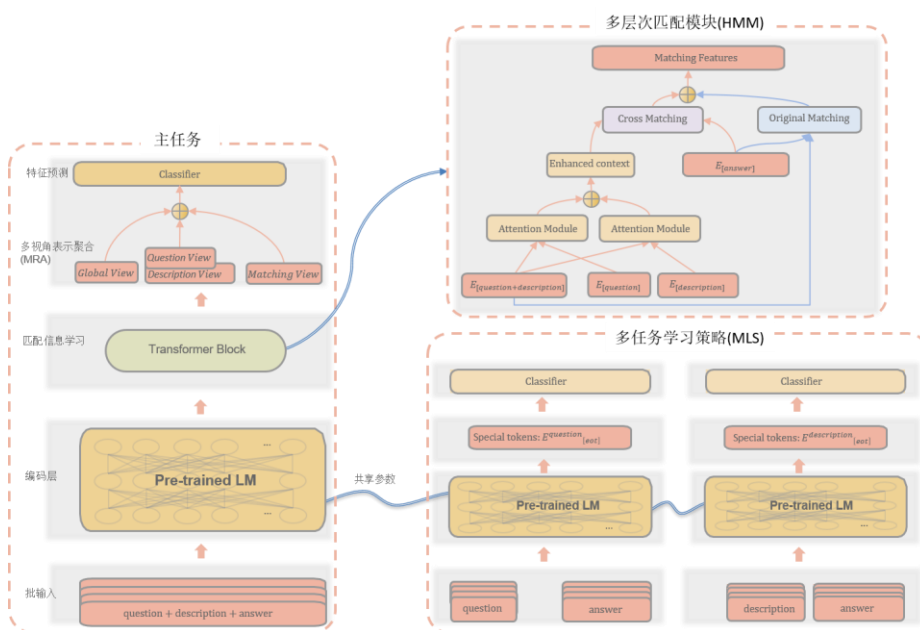


图1 多任务学习增强的多视角语义融合网络(MVMT)的整体架构

图1展示了本文所提出模型的整体架构，MVM模型主要由基于预训练语言模型的编码模块、多层次匹配模块HMM(Hierarchical Matching Module)、多视角表示聚合MRA(Multi-view Representations Aggregation)所组成，同时辅以多任务学习策略MLS(Multi-task Learning Strategy)来加强模型的推理性能。

MVM模型使用具有强大文本理解能力的预训练语言模型来学习医疗科普问答数据，借助预训练模型中特殊的上下文标记[CLS]，我们能够从中提取出代表全局语义特征的全局视角(global-view)表示；在预训练模型学习的同时由我们提出的MLS方法从问题、描述两个角度出发进一步加强预训练模型在领域数据集推理能力，学习问题视角(question-view)、描述视角(description-view)的推理信息；在预训练语言模型提取到语义特征之后，我们设计的HMM模块使用交互注意力机制对问题与回复的关系进行加强学习，输出一个能够代表用户问题与回复关系的匹配视角(matching-view)的特征向量。最后我们提出了多视角表示聚合机制来提取并结合多个视角的推理信息，充分发挥各个模块的优势。

3.2 问题定义

对于一个问答形式的医疗科普数据集 D ，内部的每一条数据形式都为 (c, a, y) 。具体的， $c = \{q, d\}$ 代表着一组问题描述对，其中 $q = \{w_1, w_2, \dots, w_n\}$ 代表包含 n 个单词的医疗问题， $d = \{w_1, w_2, \dots, w_m\}$ 代表着对应医疗问题 q 的详细描述，其中包含了 m 个单词， $a = \{w_1, w_2, \dots, w_o\}$ 代表着包含 o 个单词的回复。 $y \in \{0, 1\}$ 代表着数据的标签，其中 $y = 1$ 表示对于当前问题和问题描述来说回复是不合理的，即答非所问， $y = 0$ 则表示合理。我们的目标是在数据集 D 中学习到一个能够准确衡量医疗回答对于对应问题的不合理程度的模型 $g(c, a)$ 。

3.3 基于预训练语言模型的学习网络

模型的基本架构是基于预训练语言模型的学习网络，在预训练语言模型上根据自然语言推断任务进行微调。近年来，自然语言处理领域出现了许多优秀的预训练模型框架，有BERT, Roberta, ELECTRA, MacBERT等，我们选择ELECTRA和MacBERT作为预训练语言模型来编码医疗问答数据，将微调的过程分解为四个部分，分别是输入、编码层、特征聚合、特征预测、各个部分的流程如下。

输入

我们将医疗问答数据中的问题、描述、回答三个内容进行拼接输入进预训练模型中，与预训练模型相似的，我们在输入的头部添加[CLS]标记来学习全局的文

本信息，并在回复的前后插入[SEP]标记来区分问题描述对与回复。同时，我们在问题和描述之间添加特殊的标记[EOT]来断开问题和描述，输入可公式化为：

$$x = \{[CLS], q, [EOT], d, [EOT], [SEP], a, [SEP]\} \quad (1)$$

其中 q 代表着医疗问答数据中的问题， d 代表着描述， a 代表着回复。

编码层

编码层部分使用预训练模型进行编码，在实践中，我们将输入得到的部分直接输入进预训练模型中，利用预训练模型本身已经具有的语义理解能力，将医疗问答数据转化为语义向量 E ，以ELECTRA模型举例：

$$E = ELECTRA(x) \quad (2)$$

特征聚合

在此部分中，我们改变了传统基于预训练模型的方法中只使用[CLS]作为特征的方法，提出了新的特征聚合模块多视角特征聚合MRA(Multi-view Representations Aggregation)。在我们独立设计的多任务学习策略(MLS)和多层次匹配模块(HMM)中，模型学习到了问题视角(question-view)，描述视角(description-view)，匹配视角(matching-view)三种视角的表示，进而我们将这三种视角的表示与能代表全局视角(global-view)的[CLS]标记表示进行拼接，得到聚合的表示 $E_{ensemble}$ ，能够更加准确建模问题和回复的关系特征，这一部分会在后面的章节中详细介绍。

特征预测

特征预测部分对得到的特征进行分类学习，我们使用一个分类层进行分类，并使用交叉熵作为损失函数，其中 W 、 b 是可训练的参数， $\sigma(\cdot)$ 代表sigmoid激活函数：

$$g(c, a) = \sigma(W^T E_{ensemble} + b) \quad (3)$$

$$Loss_{main} = - \sum_{(c,a,y) \in D} y \log(g(c, a)) + (1 - y) \log(1 - g(c, a)) \quad (4)$$

在实际微调的过程中，主任务与我们提出的MLS方法同步优化预训练模型，因此损失函数为：

$$Loss = Loss_{main} + Loss_{question} + Loss_{description} \quad (5)$$

其中 $Loss_{question}$ 与 $Loss_{description}$ 为MLS方法中两个子任务的损失函数，在下一节中会进行详细介绍。

3.4 多任务学习策略

在预训练模型的基础之上，我们提出了两个自监督的学习策略来深度挖掘预训练语言模型的理解能力。由于在医疗科普数据集中，问题和描述都代表着用户的问题，其中问题中的内容更短，更加具有概括性，而描述中的内容则更长，包含更多信息。针对上述的特点，我们将问题和描述拆分开来进行考虑，将原本的推断任务拆分为两个推断子任务，分别是问题推断回复和从描述推断回复。两个推断子任务共享着主任务的预训练模型参数，同时，为了不破坏本身预训练模型中的[CLS]、[SEP]等标记已学习的语义信息，我们使用添加的[EOT]标记来进行预测，利用问题尾部的[EOT]标记和描述尾部的[EOT]标记做不同的推断任务，学习不同视角的推理信息。

从问题推断回复

为了让预训练模型有针对性的关注到问题与回复之间的关系，我们将描述部分使用[MASK]标记进行遮蔽，只对问题和回复做推断任务，输入部分的公式为：

$$x_q = \{[CLS], q, [EOT], [MASK], \dots, [MASK], [SEP], a, [SEP]\} \quad (6)$$

该子任务使用共享参数的预训练模型进行编码，以ELECTRA举例：

$$E_q = ELECTRA(x_q) \quad (7)$$

我们选择问题尾部[EOT]标记的表示作为预测的特征，并使用交叉熵作为损失函数，其中 W_q 、 b_q 是可训练的参数， $\sigma(\cdot)$ 代表sigmoid激活函数：

$$g(q, a) = \sigma(W_q^T E_{question}_{[eot]} + b_q) \quad (8)$$

$$Loss_{question} = - \sum_{(q,a,y) \in D} y \log(g(q, a)) + (1 - y) \log(1 - g(q, a)) \quad (9)$$

从描述推断回复

与前一节相似的，为了让预训练模型关注到描述与回复之间的关系，我们将问题部分使用[MASK]标记进行遮蔽，只对描述和回复做推断任务，输入部分的公式为：

$$x_d = \{[CLS], [MASK], \dots, [MASK], d, [EOT], [SEP], a, [SEP]\} \quad (10)$$

使用共享参数的预训练模型进行编码，以ELECTRA举例：

$$E_d = ELECTRA(x_d) \quad (11)$$

选择描述尾部的[EOT]的表示作为预测的特征，使用交叉熵作为损失函数，其中 W_d 、 b_d 是可训练的参数， $\sigma(\cdot)$ 代表sigmoid激活函数：

$$g(d, a) = \sigma(W_d^T E_{description}_{[eot]} + b_d) \quad (12)$$

$$Loss_{description} = - \sum_{(d,a,y) \in D} y \log(g(d,a)) + (1-y) \log(1-g(d,a)) \quad (13)$$

3.5 多层次匹配模块

为了进一步拓展加强模型对于问题与回复的感知能力，我们提出了层次化的匹配模块来学习问题和回复的关系。匹配的方法主要存在于2020年以前的自然语言处理算法中，在2020年之后被效果更好的基于预训练模型的微调方法所替代。我们为了有针对性的加强模型对于关系信息的学习能力，采用传统匹配方法与预训练模型相结合的方式，针对预训练模型的输出，通过层次化的匹配模块进一步编码学习。

上下文信息学习

给定一个问题 $D_i = \{w_{q,1}, w_{q,2}, \dots, w_{q,n}, w_{d,1}, w_{d,2}, \dots, w_{d,m}, w_{a,1}, w_{a,2}, \dots, w_{a,o}\}$ ，其中 n 为问题的长度， $w_{q,n}$ 代表问题中的第 n 个单词， m 为描述的长度， $w_{d,m}$ 代表描述中的第 m 个单词， o 为回复的长度， $w_{a,o}$ 代表回复中第 o 个单词。经过编码层编码之后，得到能够表示问题描述对的问题上下文的语义向量 $C_i \in \mathbb{R}^{(m+n) \times d}$ ，回复的语义向量 $A_i \in \mathbb{R}^{o \times d}$ ， d 代表编码层得到的向量维度。其中问题上下文的语义向量又可以根据长度拆分为问题的语义向量 $Q_i \in \mathbb{R}^{n \times d}$ ，描述的语义向量 $D_i \in \mathbb{R}^{m \times d}$ 。对于得到的语义向量，我们使用交互注意力机制对问题的上下文信息进行学习，加强上下文信息的表示，可以用下列公式表示：

$$C_i^{c,q} = \text{AttentiveModule}(C_i, Q_i, Q_i) \quad (14)$$

$$C_i^{c,d} = \text{AttentiveModule}(C_i, D_i, D_i) \quad (15)$$

由此，我们可以得到加强后的上下文信息表示为：

$$C_i^c = C_i^{c,q} \oplus C_i^{c,d}, C_i^c \in \mathbb{R}^{2 \times (m+n) \times d} \quad (16)$$

其中对于 **AttentiveModule**，我们采用由DAM模型中提出的一种注意力机制变体，注意力权重的运算公式如下：

$$V_{att} = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \quad (17)$$

使用注意力权重对模块的输入 Q 进行加权，对于得到的 sum 向量再使用一个带有 **RELU** 激活函数的线性层进行进一步的提取，其中 W_1, b_1, W_2, b_2 都是可训练的参数。

$$FFN(\text{sum}) = \max(0, \text{sum}W_1 + b_1)W_2 + b_2 \quad (18)$$

匹配信息学习

在得到了加强后的上下文信息之后，我们进一步将上下文信息与回复进行匹配，采用MSN模型中的话语与回复匹配(Utterance-Response Matching)部分的方法，通过句向量的点乘和余弦相似度计算匹配矩阵，并进一步将其分解为两种匹配方式，分别是原始匹配(Original Matching)和交互匹配(Cross Matching)。

对于原始匹配，我们将由预训练模型输出的语义向量直接进行匹配计算，则匹配矩阵的计算为：

$$M_1 = [C_i B_1 A_i^T; \cos(C_i, A_i)] \quad (19)$$

其中 \cos 代表余弦相似度计算， $B_1 \in \mathbb{R}^{d \times d}$ 是一个可学习的线性变换矩阵。

对于交互匹配，我们使用交互注意力机制建模上下文信息与回复的关系：

$$C_i^{cross} = \text{AttentionModule}(C_i^c, A_i, A_i) \quad (20)$$

$$A_i^{cross} = \text{AttentionModule}(A_i, C_i^c, C_i^c) \quad (21)$$

再计算匹配矩阵：

$$M_2 = [C_i^{cross} B_2 A_i^{crossT}; \cos(C_i^{cross}, A_i^{cross})] \quad (22)$$

其中 $B_2 \in \mathbb{R}^{d \times d}$ 是一个可学习的线性变换矩阵。

匹配特征聚合

为了增强交互注意力的影响，我们将 M_1 与两个 M_2 堆叠在一起，得到最终的匹配矩阵：

$$M = [M_1; M_2; M_2] \quad (23)$$

与MSN一致的，我们进一步使用二维卷积神经网络和最大池化层对匹配矩阵进行进一步提取，得到特征向量 h_{match} 。

3.6 多视角表示聚合

在此部分中，我们改变了传统基于预训练模型的方法中只使用[CLS]表示作为特征的做法，通过多个模块的联合作用，我们为特征预测模块提供了更多高质量的特征信息，对于这些特征信息，我们可以归类为四种：全局视角(global-view)、问题视角(question-view)、描述视角(description-view)、匹配视角(matching-view)。

全局视角(global-view)

使用带有全局语义信息的[CLS]标记的表示作为特征信息，[CLS]标记在预训练时从大量的文本中学习到了高质量的语义信息，能够准确的衡量全局文本的合理程度。

问题视角(question-view)&描述视角(description-view)

问题视角：使用问题尾部的[EOT]标记的特征表示 $E_{[eot]}^{question}$ 作为特征信息；
 描述视角：使用描述尾部的[EOT]标记的特征表示 $E_{[eot]}^{description}$ 作为特征信息。
 模型在从问题推断回复的子任务和从描述推断回复的子任务中，让[EOT]标记的特征表示学习掌握了由问题和描述单独推断回复的能力，并且不同位置的[EOT]表示能够根据其所属的文本（问题或回复）而代表不一样的语义信息。

匹配视角(matching-view)

使用多层次匹配模块(HMM)提供的特征向量 h_{match} 作为特征信息，在层次化匹配模块的作用下， h_{match} 能够从匹配的角度学习到问题上下文与回复的关系。

特征聚合

为了能够利用到每一种视角的特征信息，我们将四种视角的特征进行拼接，得到需要进行预测的特征向量：

$$E_{ensemble} = E_{[CLS]} \oplus E_{[eot]}^{question} \oplus E_{[eot]}^{description} \oplus h_{match} \quad (24)$$

3.7 领域适应策略

我们将预训练模型先在训练集中进行领域学习，让通用预训练模型根据预测[MASK]标记单词的MLM(masked language modeling)任务和下一句话预测NSP(Next Sentence Prediction)任务在微调前先掌握医疗数据集的特点。由于竞赛数据集比较小，为了防止领域学习阶段与微调阶段的数据与任务过于相近导致过拟合，我们删除了描述部分的内容，截断了较长的回复，只使用问题和短回复进行领域学习。

3.8 额外策略

多模型融合

为了在单模上进一步提升模型的效果，我们尝试了多种架构的模型融合策略，对于验证集阶段和测试集阶段，我们使用了不同的架构组合。为了增大不同模型之间的差异性，我们在确保模型效果的不会损失太大的情况下对部分模型的架构进行精简。我们采用多数投票(Majority Vote)的融合策略，让多个模型对同一个测试样例投票进行表决，超过半数就进行确认，且效果好的单模型的可以持有更多的票数，再结合上我们提出的高精确率模型+高召回率模型的融合技巧，能够在单模的基础上有进一步的提升。

关于高精确率模型+高召回率模型技巧，我们在实验中发现，不同架构的模型在精确率和召回率上有时会有明显的不同，有的模型的精确率会比较突出，但是召回率较低，而有的则相反，所以我们将这两种模型结合在一起，让高精确率模型先预测出一部分答非所问样本，再使用高召回率模型进行补充，此技巧在验证集阶段和测试集阶段在效果上都有一定的提升。

伪样本数据增强

为了充分利用官方提供的无标签数据，在竞赛的测试集阶段，我们针对验证集中的数据进行预测，将预测结果加入到训练集中作为补充数据进行训练，能够有明显的提升。在实验中，我们尝试了两种方案，第一种是将全部预测结果加入到训练集中作为伪标签，第二种是抽取其中置信度高的条目作为伪标签，两种方案都有着一定的效果。

对抗学习

为了增加模型的鲁棒性，我们采用了对抗学习的策略，在预训练语言模型的嵌入层上添加FGM扰动，以此提升模型的泛化能力。尽管该方案在单模型上没有提升效果，但是我们使用扰动攻击后的模型与攻击前的模型进行融合，在保证模型效果的同时增大了两个模型之间的差异性，取得了一定的效果。

4 实验

4.1 数据集

表2 答非所问识别数据集统计信息

数据集(训练/验证/测试)	答非所问识别
数量	40000/5000/10000
正例：负例	1:2/-/-
平均问题上下文长度(问题+描述)	65.34/64.31/61.67
平均回复长度	129.45/130.43/129.60

我们在竞赛官方提供的数据集中进行了实验，数据集的内容是医疗科普知识问答，包含了带有标注数据的40000条训练集数据，第一阶段评测的无标签5000条验证集，最后阶段评测的无标签10000条测试集。数据集中的每一条数据包含

问题(question)、描述(description)和回复(answer)，具体数据集的分布见表2。

4.2 实验设置与评估指标

我们基于Pytorch²框架实现了模型，并在单张NVIDIA Tesla V00 GPU上进行了实验。对于模型中用于微调的预训练模型，我们使用了由哈工大讯飞联合实验室提供的chinese-electra-180g-large-discriminator³和chinese-macbert-large⁴。预训练语言模型的输入长度统一设定为512，其中我们将问题上下文长度(问题+描述)设定为192，回复长度设定为320，对短于设定长度的条目我们使用[PAD]标记进行填充。在微调的过程中，我们采用梯度累积策略，累积到40个批次后再更新梯度。梯度下降时使用Adam算法作为优化器，学习率设定为 8×10^{-6} ，模型中所有的全连接层的dropout比例统一设定为0.8。根据我们的实验结果，对于领域适应策略，我们取训练一轮后的模型进行微调；对于所有模型的微调，我们统一取微调两轮后的模型进行预测。

竞赛中官方提供的模型评测指标为F1值，F1值可由精确率P、召回率R共同计算得到。假设模型预测为答非所问的条目集合为A，标注为答非所问的条目集合为B，则F1的计算可由下面的三个公式得到：

$$P = \frac{|A \cap B|}{|A|} \quad R = \frac{|A \cap B|}{|B|} \quad F1 = \frac{2 \times P \times R}{P + R} \quad (25)$$

4.3 实验结果

由于我们设计的各个模块在验证集和测试集中的性能有所不同，因此我们在验证集和测试集中采用了不一样的框架组合，在下面的实验结果中，我们使用MLS代表完整使用了多任务学习策略(MLS)，使用MLS(description)代表只使用了MLS中的由描述推断回复子任务；我们使用HMM代表使用了多层次匹配模块(HMM)，使用MRA代表使用了多视角表示聚合(MRA)，使用Post代表使用了领域学习策略。在验证集阶段，我们提交的最佳单模型是：ELECTRA-Post-MLS-HMM-MRA。

对于验证集阶段的模型融合，我们取得最优效果的模型组合是：ELECTRA-MLS(description)-HMM-MRA、ELECTRA-Post-MLS(description)-HMM-MRA。该融合结果在第一阶段取得了第四名的成绩，使用的融合策略是高精确率模型+高召回率模型的技巧，将ELECTRA-Post-MLS(description)-HMM-MRA作为高精确率模型，将ELECTRA-MLS(description)-HMM-MRA作为高召回率模型。在表3的实验结果中，

² <https://github.com/pytorch/pytorch>

³ <https://huggingface.co/hfl/chinese-electra-180g-large-discriminator>

⁴ <https://huggingface.co/hfl/chinese-macbert-large>

我们使用ELECTRA-MVMT-Ensemble代表验证集阶段的模型融合方案，比最佳单模型方案在F1上高了接近0.4%，验证了高精确率+高召回率模型融合技巧的有效性。

表3 MVMT模型在验证集阶段的实验效果

模型名	F1	Precision	Recall
ELECTRA-Post-MLS-HMM-MRA	0.840389	0.847930	0.834270
ELECTRA-MVMT-Ensemble(rank 4)	0.844059	0.851677	0.837870

测试集阶段，为了进一步提升效果，我们将MacBERT作为我们的预训练语言模型，并且使用了伪样本数据增强技巧进一步提升效果，在下面的实验结果中，我们使用Pseudo代表使用了全部的验证集预测结果作为伪标签，共有5000条，使用Pseudo(filter)代表使用了过滤了低置信度条目后的预测结果，共有4545条。由于在测试集阶段，根据我们的实验发现领域适应策略和多任务学习策略中的MLS(question)的效果不明显，因此我们提交的最佳单模型是：MacBERT-MLS(description)-HMM-MRA-Pseudo。

对于模型融合，我们取得最优效果的模型融合策略是将多数投票策略与高精确率模型+高召回率模型技巧相结合。在多数投票策略部分，我们使用以下四个模型进行投票：MacBERT-FGM-HMM-MRA-Pseudo、MacBERT-MLS(description)-HMM-MRA-Pseudo、MacBERT-MLS(description)-HMM-MRA-Pseudo、ELECTRA-Post-MLS-HMM-MRA-Pseudo。

由于是偶数个模型，所以我们给予四个模型中效果最好的MacBERT-MLS(description)-HMM-MRA-Pseudo两张票以防止平票，投票得到的融合结果我们在表中命名为MacBERT-MVMT-Ensemble(Majority Vote)。同时我们将MacBERT-MLS(description)-HMM-MRA-Pseudo(filter)作为高精度模型，与多数投票策略的结果进行融合。我们使用MacBERT-MVMT-Ensemble代表测试集阶段的最佳模型融合方案，在最终阶段的评测中排名第三。此外，值得一提的是我们的单模型效果依然能够在最终评测中排名第三，体现了我们方案的有效性。

表4 MVMT在测试集阶段的实验效果

模型名	F1	Precision	Recall
MacBERT-MLS(description)-HMM-MRA-Pseudo	0.690136	0.675553	0.705363
MacBERT-MVMT-Ensemble(Majority Vote)	0.694786	0.680104	0.710115
MacBERT-MVMT-Ensemble(rank 3)	0.697775	0.683030	0.713170

同时从表4中的实验结果可以看出，我们应用的两种模型融合策略均可以有效提高精度，且当两个模型融合策略结合使用时效果最佳。

4.4 消融分析

表5 MVMT架构的消融学习结果

模型名	F1	Precision	Recall
MacBERT-MLS-HMM-MRA	91.13%	91.77%	90.49%
w/o MLS(question)	90.71%	88.66%	92.85%
w/o MLS(description)	90.54%	90.95%	90.13%
w/o HMM	90.61%	90.61%	90.61%
w/o MRA	90.83%	89.91%	91.76%

为了充分验证我们提出的各个模块的有效性，我们在此节中对我们提出的多任务学习策略(MLS)、多层次匹配模块(HMM)、多视角表示聚合(MRA)三个主要的优化方法展开消融实验。由于竞赛中的提交次数很有限，并且在竞赛结束之后已经无法再进行实验，所以我们从训练集中随机抽取出了5000条的数据用于验证主要模块的效果，使用抽取后剩下的35000条数据进行训练，以MacBERT作为预训练语言模型。

表6 各优化模块的组合实验

编号	模型名	F1	Precision	Recall
1	MacBERT	90.37%	87.92%	92.97%
2	MacBERT-MLS(question)	90.38%	90.02%	90.73%
3	MacBERT-MLS(description)	90.27%	88.76%	91.82%
4	MacBERT-MLS	90.61%	90.61%	90.61%
5	MacBERT-HMM	90.14%	87.69%	92.73%
6	MacBERT-MLS(question)-HMM	90.13%	89.30%	90.98%
7	MacBERT-MLS(description)-HMM	90.58%	89.77%	91.40%
8	MacBERT-MLS-HMM	90.83%	89.91%	91.76%
9	MacBERT-MLS(question)-HMM-MRA	90.54%	90.95%	90.13%
10	MacBERT-MLS(description)-HMM-MRA	90.71%	88.66%	92.85%
11	MacBERT-MLS-HMM-MRA	91.13%	91.77%	90.49%

表5中我们列举了MVM架构的消融实验结果，从实验结果中可以看出，每个主要的模块都能够发挥出明显的效果。其中w/o MLS(question)是去除掉MLS中的从问题推理回答子任务的结果，实验结果相比完整的算法在F1分数上下降了0.42%；w/o MLS(description)是去除掉MLS中的从描述推理回答子任务的结果，实验结果相比完整的算法在F1分数上下降了0.59%；上述结果显示两个子任务都能够对模型的推理能力有积极的促进效果，且从描述推断回复的子任务对预训练语言模型的影响更大。w/o HMM是去除掉HMM的实验结果，实验结果相比完整的算法在F1上下降了0.52%，证明了使用匹配模块来学习问题与回复关系的方法是有效的。w/o MRA是去除掉MRA的实验结果，实验结果相比完整的算法在F1上下降了0.3%，这证明了MRA机制的有效性。多个优化模块简单的组合在一块尽管也可以带来提升，但是无法充分发挥各个优化模块的效果，我们设计的MRA机制能够从多个视角最大化每个模块对整体模型的影响。

此外在表6中我们详细列举了各个优化模块的组合结果，从其实验结果中可以看到，编号1是我们直接使用MacBERT进行微调得到的基线模型，其中F1的分数达到了90.37%；编号2到5是我们分别单独使用各个优化策略的结果，与基线模型的对比可以看出单独使用各个优化策略带来的提升比较有限，甚至其中的一些优化策略还会带来效果的下降，证明了只从部分角度加强模型是不足的，提升模型的效果需要从多方面同时促进模型学习；编号6到8的实验中我们将多任务学习策略(MLS)与多层次匹配模块(HMM)结合起来，与前面单独使用各个优化策略的结果相比，其中编号6的模型实验效果出现了下降，编号7模型相比对应的编号3、5模型分别有0.31%、0.44%的提升，编号8模型相比对应的编号4、5分别有0.22%、0.69%的提升，上述结果表明在大多数情况下组合多个优化模块能够带来提升，但是如果不能充分利用各个模块学习到的信息，就不能够充分发挥各个模块的效果，有时甚至会出现“一加一小于二”的情况；在编号9-11的实验中我们将能够充分发挥各个模型作用的多视角表示聚合机制(MRA)与编号6-8的方案进行结合，对比结果显示结合后的方案在F1分数上都有了较大的提升，充分证明了MRA机制的有效性。其中编号11的方案中同时运用了多任务学习策略、多层次匹配模块、多视角聚合机制三个方案，在F1分数上超过基线模型0.76%，达到了最佳的实验结果。综上，从表6的实验结果中可以看出，我们提出的各个优化模块是相辅相成、相互促进的，只有同时充分利用各个模块，从多个角度优化模型才能取得最优的结果。

5 总结

在本文中，我们提出了多任务学习增强的多视角语义特征融合网络 MVMT (Multi-view Semantic Feature Fusion Network enhanced by Multi-task Learning)，从加强推理和加强对话理解两个角度出发，生成多种视角的语义特征来学习答非所问识别任务，有效解决了答非所问任务需要同时学习推理任务和对话信息的难点。我们提出的 MVMT 模型在最终的测试集阶段评测中排名第三，证明了我们方法的有效性。

参考文献

1. Talman, A., Yli-Jyrä, A., & Tiedemann, J. (2018). Natural language inference with hierarchical bilstm max pooling architecture. arXiv preprint arXiv:1808.08762.
2. Nie, Y., & Bansal, M. (2017). Shortcut-stacked sentence encoders for multi-domain inference. arXiv preprint arXiv:1708.02312.
3. Shen, T., Zhou, T., Long, G., Jiang, J., Wang, S., & Zhang, C. (2018). Reinforced self-attention network: a hybrid of hard and soft attention for sequence modeling. arXiv preprint arXiv:1801.10296.
4. Yoon, D., Lee, D., & Lee, S. (2018). Dynamic self-attention: Computing attention over words dynamically for sentence embedding. arXiv preprint arXiv:1808.07383.
5. Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. IEEE transactions on Signal Processing, 45(11), 2673-2681.
6. Parikh, A. P., Täckström, O., Das, D., & Uszkoreit, J. (2016). A decomposable attention model for natural language inference. arXiv preprint arXiv:1606.01933.
7. Chen, Q., Zhu, X., Ling, Z., Wei, S., Jiang, H., & Inkpen, D. (2016). Enhanced lstm for natural language inference. arXiv preprint arXiv:1609.06038.
8. Chen, Q., Zhu, X., Ling, Z. H., Inkpen, D., & Wei, S. (2017). Neural natural language inference models enhanced with external knowledge. arXiv preprint arXiv:1711.04289.
9. Gong, Y., Luo, H., & Zhang, J. (2017). Natural language inference over interaction space. arXiv preprint arXiv:1709.04348.
10. Tan, C., Wei, F., Wang, W., Lv, W., & Zhou, M. (2018, January). Multiway Attention Networks for Modeling Sentence Pairs. In IJCAI (pp. 4411-4417).
11. Dey, R., & Salem, F. M. (2017, August). Gate-variants of gated recurrent unit (GRU) neural networks. In 2017 IEEE 60th international midwest symposium on circuits and systems (MWSCAS) (pp. 1597-1600). IEEE.

12. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.
13. Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543).
14. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
15. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.
16. Clark, K., Luong, M. T., Le, Q. V., & Manning, C. D. (2020). Electra: Pre-training text encoders as discriminators rather than generators. arXiv preprint arXiv:2003.10555.
17. Cui, Y., Che, W., Liu, T., Qin, B., Wang, S., & Hu, G. (2020). Revisiting pre-trained models for chinese natural language processing. arXiv preprint arXiv:2004.13922.
18. Zhou, X., Dong, D., Wu, H., Zhao, S., Yu, D., Tian, H., ... & Yan, R. (2016, November). Multi-view response selection for human-computer conversation. In Proceedings of the 2016 conference on empirical methods in natural language processing (pp. 372-381).
19. Wu, Y., Wu, W., Xing, C., Zhou, M., & Li, Z. (2016). Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots. arXiv preprint arXiv:1612.01627.
20. Yuan, C., Zhou, W., Li, M., Lv, S., Zhu, F., Han, J., & Hu, S. (2019, November). Multi-hop selector network for multi-turn response selection in retrieval-based chatbots. In Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP) (pp. 111-120).
21. Zhou, X., Li, L., Dong, D., Liu, Y., Chen, Y., Zhao, W. X., ... & Wu, H. (2018, July). Multi-turn response selection for chatbots with deep attention matching network. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 1118-1127).
22. Whang, T., Lee, D., Lee, C., Yang, K., Oh, D., & Lim, H. (2020, October). An Effective Domain Adaptive Post-Training Method for BERT in Response Selection. In INTERSPEECH (pp. 1585-1589).
23. Gu, J. C., Li, T., Liu, Q., Ling, Z. H., Su, Z., Wei, S., & Zhu, X. (2020, October). Speaker-aware bert for multi-turn response selection in retrieval-based chatbots. In Proceedings of the 29th ACM International Conference on Information & Knowledge Management (pp. 2041-2044).

24. Whang, T., Lee, D., Oh, D., Lee, C., Han, K., Lee, D. H., & Lee, S. (2021, May). Do response selection models really know what's next? utterance manipulation strategies for multi-turn response selection. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 35, No. 16, pp. 14041-14049).
25. Xu, R., Tao, C., Jiang, D., Zhao, X., Zhao, D., & Yan, R. (2020). Learning an effective context-response matching model with self-supervised tasks for retrieval-based dialogues. *arXiv preprint arXiv:2009.06265*.