计算机科学
COMPUTER SCIENCE

基于多层感知机和语义矩阵的答案选择模型

罗亮, 程春玲, 刘倩, 归耀城

引用本文

罗亮, 程春玲, 刘倩, 归耀城.基于多层感知机和语义矩阵的答案选择模型[J]. 计算机科学, 2023, 50(5): 270-276.

LUO Liang, CHENG Chunling, LIU Qian, GUI Yaocheng. Answer Selection Model Based on MLP and Semantic Matrix [J]. Computer Science, 2023, 50(5): 270-276.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.220400275

基于多层感知机和语义矩阵的答案选择模型

罗　亮[1]　程春玲[1]　刘　倩[1]　归耀城[2]
1 南京邮电大学计算机学院、软件学院、网络空间安全学院　南京 210023
2 南京邮电大学现代邮政学院　南京 210023
(1220045114@njupt.edu.cn)

摘　要　答案选择是问答系统领域的关键子任务,其性能表现支撑着问答系统的发展。基于参数冻结的BERT模型生成的动态词向量存在句级语义特征匮乏、问答对词级交互关系缺失等问题。多层感知机具有多种优势,不仅能够实现深度特征挖掘,且计算成本较低。在动态文本向量的基础上,文中提出了一种基于多层感知机和语义矩阵的答案选择模型,多层感知机主要实现文本向量句级语义维度重建,而通过不同的计算方法生成语义矩阵能够挖掘不同的文本特征信息。多层感知机与基于线性模型生成的语义理解矩阵相结合,实现一个语义理解模块,旨在分别挖掘问题句和答案句的句级语义特征;多层感知机与基于双向注意力计算方法生成的语义交互矩阵相结合,实现一个语义交互模块,旨在构建问答对之间的词级交互关系。实验结果表明,所提模型在WikiQA数据集上*MAP*和*MRR*分别为0.789和0.806,相比基线模型,该模型在性能上有一致的提升,在SelQA数据集上*MAP*和*MRR*分别为0.903和0.911,也具有较好的性能表现。

关键词:答案选择;BERT模型;动态词向量;多层感知机;语义矩阵
中图法分类号　TP391.1

Answer Selection Model Based on MLP and Semantic Matrix

LUO Liang[1],CHENG Chunling[1],LIU Qian[1] and GUI Yaocheng[2]
1 School of Computer Science,Nanjing University of Posts and Telecommunications,Nanjing 210023,China
2 School of Modern Posts,Nanjing University of Posts and Telecommunications,Nanjing 210023,China

Abstract　Answer selection is a key sub-task in the field of question answering systems,and its performance supports the development of question answering systems. The dynamic word vector generated by the BERT model based on parameter freezing also has problems such as lack of sentence-level semantic features and the lack of word-level interaction between question and answer. Multilayer perceptrons have a variety of advantages,they not only can achieve deep feature mining,but also have low computational costs. On the basis of dynamic text vectors,this paper proposes an answer selection model based on multi-layer perceptrons and semantic matrix,which mainly realizes the semantic dimension reconstruction of text vector sentences,and generates semantic matrix through different calculation methods to mine different text feature information. The multi-layer perceptron is combined with the semantic understanding matrix generated by the linear model to implement a semantic understanding module,which aims to excavate the sentence-level semantic characteristics of the question sentence and the answer sentence respectively; the multi-layer perceptron is combined with the semantic interaction matrix generated based on the two-way attention calculation method to achieve a semantic interaction module,which aims to build the word-level interaction relationship between the question and answer pairs. Experimental results show that the proposed model has a *MAP* and *MRR* of 0.789 and 0.806 on the WikiQA dataset,respectively,which has a consistent performance improvement over the baseline model,on the SelQA dataset,*MAP* and *MRR* is 0.903 and 0.911,respectively,which also has a good performance.

Keywords　Answer selection,BERT model,Dynamic word vector,Multilayer perceptron,Semantic matrix

1　引言

问答系统的技术进步衍生出了许多人类生活中已普遍存在的智能产品,例如语音助手、智能客服等,帮助人与机器使用自然语言交流,为用户提供了便利、高效、智能的生活方式。问答系统能够依据语境和语义,理解用户更深层次的真实目的,从而提供一个相对精准且符合人类思维的答案。答案选择任务作为问答系统的关键子任务,为问答系统的性能

到稿日期:2022-04-27　返修日期:2022-09-10
基金项目:江苏省双创博士项目(JSSCBS20210507);南京邮电大学引进人才科研启动基金(NY220176)
This work was supported by the Foundation of Jiangsu Provincial Double-Innovation Doctor Program(JSSCBS20210507) and NUPTSF (NY220176).
通信作者:程春玲(chengcl@njupt.edu.cn)

保障提供了重要支撑。答案选择任务是经典的文本匹配任务之一，它的目标是给定一个问题句和以句子为单位构成的答案句候选池，然后以候选答案句是否包含问题句的答案为依据，从候选池中匹配出一个或多个正确的答案句并进行优先级排序。

在传统的答案选择模型[1-2]中，词编码层主要由 GloVe[3]，Word2Vec[4]等静态词向量技术实现，将自然语言转换为词向量形式，将离散的语言符号映射到高维向量空间。这一类词编码过程是静态的，以单词为单位，无论单词所在的文本句上下文如何，单词的词向量始终是固定的，这将带来词向量表征不详的问题，即不同语境下一词多义和同义词等问题，导致答案选择任务准确率不高。为此，传统答案选择模型多采用 LSTM 和 Transformer 等神经网络作用在文本向量上，来辅助提取文本句中每个单词的上下文语境信息，以丰富语义表达。

近年来，基于上下文信息编码的动态词向量技术，例如 BERT，RoBERTa 等，受到了广泛关注[5-6]。这些动态词向量技术在词编码阶段充分考虑了文本句的语境信息，为下游任务提供了更好的中间文本特征。在答案选择任务中，引入 BERT 模型构建新的词编码层，为解决词义匮乏和一词多义等问题提供了新的思路。为充分利用 BETR 模型生成的动态文本向量特征，提升答案选择任务的准确度，现有工作还面临着两个挑战：1）基于动态词向量技术生成的文本向量仅是词级维度的拼接，缺乏句级维度的语义逻辑关系；2）难以准确捕捉问题句中的单词与答案句中高度相关的对应词之间的词级语义交互信息。

为解决上述难题，本文提出了一种基于多层感知机和语义矩阵的答案选择模型。本文主要的贡献包括：

（1）应用多层感知机来解决复杂的语义特征提取问题，其性能表现和计算复杂度均较优。

（2）针对答案选择任务，将语义矩阵嵌入多层感知机，构建了两个网络模块，实现了句级语义特征的挖掘和词级交互关系的构建。

2　相关工作

传统的答案选择模型大多基于静态词向量技术生成的文本向量，采用神经网络实现对问题句和候选答案句之间的语义关系建模，然后通过最终的问题句和答案句文本向量衡量问答对之间的匹配程度。根据答案选择模型整体架构可以将这些方法分为基于孪生网络和基于比较聚合架构两种。基于孪生网络的答案选择模型采用共享参数机制，应用神经网络和注意力机制对静态文本向量实现语义建模，无论是问题句还是答案句都将经过统一参数的网络结构，该共享参数机制不仅解决了问题句数量与候选答案句数量严重不平衡的问题，丰富了问题句文本特征，还使得模型量级更小，加快了模型收敛。Attentive LSTM[7]通过双向 LSTM 共同捕捉问题句和答案句文本的双向长距离依赖关系。基于位置注意力机制的 RNN 答案选择模型[8]，以增强 RNN 所生成的文本特征为目标，通过位置注意力机制捕获问题句原始文本的词序关系，然后附加于 RNN 所生成的答案句文本表示向量。AMMSNN[9]采用多尺度词级特征思想，基于不同尺度的卷积核的卷积网络组，将各个语言粒度词组信息考虑在内，合理区分问题句和答案句中不同词长的词组特征。基于比较聚合的答案选择模型[10]通过多种方法，例如神经网络、欧几里得和余弦计算公式、相减或点乘等方法，比较问题句和答案句中的词单元，将比较结果用 LSTM 或 CNN 网络重新聚合成新的问答对文本向量，然后通过余弦相似度计算公式计算问答对之间的匹配程度。该架构更细致地探究了问题句和答案句之间的词级交互关系，体现了更细粒度的文本词级特征，但复杂的比较聚合逻辑导致模型参数量较大，计算成本较高。随后 Dynamic-Clip[11]，Comp-Clip＋LM＋LC[12]和 DAMPM[13]基于基础的比较聚合答案选择模型架构添加了动态注意力[11]、潜在聚簇[12]和多角度匹配[13]等辅助方法，能够一定程度地提高答案选择任务的准确度。无论是基于孪生网络还是基于比较聚合的答案选择模型，它们都有一个共同的特点，就是在应用各种神经网络和注意力机制等深度学习方法解决静态词向量所带来的词义匮乏和特征隐藏等问题时，集中于挖掘文本句子内词级上下文信息，而忽略了句子级别语义逻辑关系对答案选择任务的重要性。

动态词向量技术从大规模语料库中学习各种语境下的词语表达形式，通过输入整个句子为每个词建立对应的词向量，相同的词会依据语境上下文信息生成不同的词向量，有效地解决了不同语境下一词多义和同义词等难题。如 ELMo[14]利用双向 LSTM 网络结构同时获取句子的正向和逆向语义表达并加以整合，捕捉文本句中词与词的长距离依赖关系和词序关系，表达每个词所属的语境特征。受到 Transformer[15]网络结构的启发，大量基于 Transformer 的预训练模型涌现，包括 GPT[16]，Bert[17]，RoBERTa[18]等，这些技术提供了更为丰富的文本向量特征。以 BERT 等预训练模型为答案选择模型词编码层的技术支持，为词级上下文信息的嵌入提供了新的选择。$\text{BERT}_{\text{base}}$＋Transformer Encoder[19]采用冻结参数的 BERT 模型生成对应的文本向量，然后通过神经网络结构实现答案选择。该方法对于答案选择任务的准确度有一定提升，但缺乏对动态文本向量和答案选择任务之间关系特点的分析，几乎是将传统的答案选择模型的词编码层之外的网络结构直接应用于 BERT 模型所提供的文本向量之上，其网络模型并不契合动态文本向量，未能充分利用动态文本向量特征，不能很好地应对所面临的挑战，其准确度还有一定的上升空间。

除此之外，很多答案选择任务的研究还借助了多方面的技术支持，例如基于协作对抗网络的答案选择模型[20]、借助外部知识丰富词义表达的答案选择模型[21]等。而本文主要探究传统模式下，动态文本向量特征在答案选择任务中新的应用方式，从所忽略的句子层面深度语义理解和文本匹配领域至关重要的词语级别文本对深度交互两方面出发，以提升答案选择模型准确度和降低模型复杂度为目标，设计和提出了一个新颖的基于 BERT 模型的答案选择模型架构。

3　基于多层感知机和语义矩阵的答案选择模型

3.1　整体模型架构

BERT 模型虽然能够根据句子语境提供语义丰富的文本向量特征，但依旧缺乏针对答案选择任务的特征处理过程。

为此，本文将在动态词编码层后添加专门处理答案选择任务的网络结构。本文所提答案选择模型的整体架构如图 1 所示。

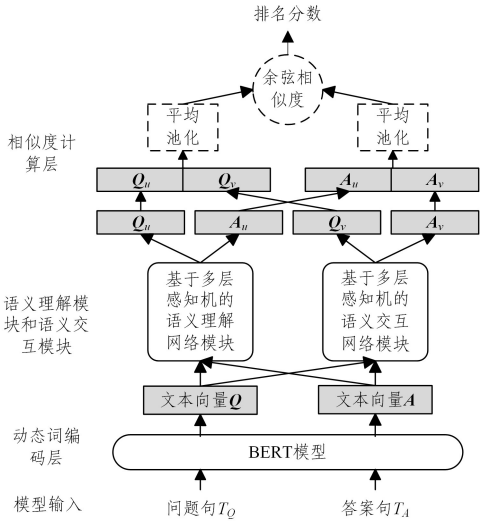


图 1 基于多层感知机和语义矩阵的答案选择模型整体架构

Fig. 1 Overall architecture of answer selection model based on multilayer perceptrons and semantic matrices

该模型共包含动态编码层、语义理解模块、语义交互模块和相似度计算层 4 个部分。答案选择任务中，问题句和答案句之间的对应关系是一对多，实际训练过程中问题句的数量远少于答案句。为避免问题句一方不拟合，本文模型采用孪生网络的共享参数机制，使得语义理解模块和语义交互模块在问题句和答案句之间共享参数。

(1)动态词编码层。本文模型实现的动态词编码层所使用的是预训练阶段后、微调之前的 BERT 模型，即禁止 BERT 模型在答案选择任务训练过程中更新参数，该模型称为冻结参数的 BERT 模型，其目的是将冻结参数的 BERT 模型充当一个适用于多种不同任务的词编码工具。

(2)语义理解模块。该模块的主体是多层感知机结构，额外添加了语义矩阵的生成、原始文本向量与语义矩阵之间的映射两个操作，作用于文本向量的句级维度。文献[22-23]指出，基于多层感知机的网络架构足以媲美常用来解决计算机视觉领域问题的卷积神经网络和基于注意力机制的网络，这是一种具有竞争力的替代方案，但概念和技术上更为简单。因此本文同样尝试使用基于多层感知机的网络架构来解决答案选择任务所面临的挑战，该架构将替代传统答案选择模型中的 RNN，LSTM 和 Transformer 等参数量大且计算复杂度高的网络结构。

(3)语义交互模块。本文模型还将延续传统答案模型中交互层行使的职责，从词级维度捕获问题句和答案句之间的交互关系，提出了一个语义交互模块，其结构与语义理解模块相似，也是由多层感知机与语义矩阵构成，但语义矩阵的生成方式不同，语义矩阵与原始文本向量的映射方法也不同，该语义矩阵的相关操作将实现问答对之间的交互。由于语义理解模块在挖掘句级维度特征的过程中会对原始词向量特征产生影响，为避免原始词级语义丢失问题，语义交互模块采用和语义理解模块并行的方案，即与语义理解模块共同置于动态词编码层之后。

(4)相似度计算层。该层通过拼接的特征融合方式生成最终的文本向量，计算问题句和答案句文本向量之间的余弦相似度，衡量两者的匹配程度。

该模型以问题句 $T_Q=\{w_1^q,w_2^q,\cdots,w_m^q\}$ 和单个答案句 $T_A=\{w_1^a,w_2^a,\cdots,w_n^a\}$ 为模型的输入，其中文本句 $T_Q(T_A)$ 包含 $m(n)$ 个单词，通常来说 $m\neq n$。BERT 模型将问题句和答案句中的每个单词 w 映射到高维向量空间，然后通过拼接的方式，将所有词向量拼接形成最终的文本向量 $\boldsymbol{Q}(\boldsymbol{A})$，文本向量 $\boldsymbol{Q}(\boldsymbol{A})$ 的维度为 $l_Q(l_A)\times d$。为了实现问题句和答案句在语义理解模块和语义交互模块上参数共享，动态词编码层还需通过填充或截断的方式，将问题句和答案句文本向量长度从 $l_Q(l_A)$ 修改成固定长度 l。随后，文本向量 $\boldsymbol{Q}$ 和 $\boldsymbol{A}$ 进入语义理解模块，输出对应的文本向量 $\boldsymbol{Q}_u$ 和 $\boldsymbol{A}_u$，同时 $\boldsymbol{Q}$ 和 $\boldsymbol{A}$ 进入语义交互模块，输出对应的文本向量 $\boldsymbol{Q}_v$ 和 $\boldsymbol{A}_v$。最后，$\boldsymbol{Q}_u$ 和 $\boldsymbol{Q}_v$ 按序拼接，从句子维度采取平均池化操作生成问题句对应的最终文本向量 $\boldsymbol{Q}_{uv}$，同理生成答案句对应的 $\boldsymbol{A}_{uv}$。相似度计算层基于 $\boldsymbol{Q}_{uv}$ 和 $\boldsymbol{A}_{uv}$ 两个向量计算余弦相似度，衡量两者的匹配关系，相似度越高，优先级就越高。

3.2 语义理解模块

语义理解模块主要解决文本向量句子维度上向量特征离散化的问题，通过多层感知机构建新的句级语义空间，依赖线性模型挖掘句子级别语义特征。该模块由两个输入输出维度对称的全连接层构成，并在全连接层之间添加了语义矩阵生成和映射操作。语义理解模块能够挖掘句子维度的潜在特征，有利于在相似度计算层计算问题句和答案句的相似度时，提高句子间语义相似度的准确率。为了实现降低网络架构复杂度的目标，本模块选用线性模型来生成语义矩阵，学习句级维度语义特征。语义理解模块的网络结构如图 2 所示。

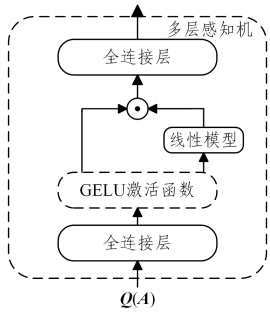


图 2 语义理解模块

Fig. 2 Semantic understanding module

本模块首先将全连接层作用于文本向量 $\boldsymbol{Q}(\boldsymbol{A})$ 词级维度之上，将原本维度为 d 的每个词向量重新映射到维度为 d_f1 的维度空间中，生成文本向量 $\boldsymbol{Q}_1(\boldsymbol{A}_1)$，然后将其从句子维度拆分，形成 d_f1 个 l 维的句向量，采用线性模型挖掘每个句向量的潜在语义特征，生成对应的语义理解矩阵，语义理解矩阵的生成方法如下：

$$\boldsymbol{M}_1=\boldsymbol{W}\boldsymbol{Q}_1+b \tag{1}$$

其中，$\boldsymbol{W}$ 是 $l\times l$ 维的权重矩阵，b 是偏重，$\boldsymbol{W}$ 和 b 将在训练过程中更新学习得到。随后通过点乘操作，将文本向量 $\boldsymbol{Q}_1(\boldsymbol{A}_1)$ 映射到新的语义空间，最后通过一层全连接层，以便更好地协助文本对之间的语义匹配。

文本向量 $\boldsymbol{W}$ 在整个模块中的计算过程可定义为：

$$\boldsymbol{Q}_1=\sigma_1(\boldsymbol{Q}\boldsymbol{W}_1+b_1) \tag{2}$$

$$\boldsymbol{Q}_2=\boldsymbol{Q}_1\odot\sigma_2(\boldsymbol{M}_1) \tag{3}$$

$$\boldsymbol{Q}_u=\boldsymbol{Q}_2\boldsymbol{W}_2+b_2 \tag{4}$$

其中，σ_1是激活函数 GELU，σ_2是激活函数 Tanh，式(2)指代全连接层，其输入维度是 d，输出维度是 d_f1，式(4)也指代全连接层，其输入维度是 d_f1，输出维度是 d，$\boldsymbol{M}_1$的维度为 $l\times d\_f1$，$\odot$表示点乘操作，文本向量 $\boldsymbol{A}$ 的计算过程与 $\boldsymbol{Q}$ 相同，且经过相同的网络结构以实现参数共享。整个模块基于基本的矩阵运算，模型参数量仅与文本向量的句级维度呈线性相关，模型的计算复杂度也是线性的。

3.3　语义交互模块

语义交互模块主要挖掘问题句和答案句之间的交互关系，主要是通过多层感知机构建新的句级语义空间和依赖注意力机制挖掘词与词之间的交互关系。本模块同样也采用双层感知机的基础网络架构，与语义理解模块不同的是，本模块的语义矩阵将指示问题句和答案句各个词向量之间的交互关系，其计算方法如下：

$$\boldsymbol{M}_2=\boldsymbol{Q}_3\boldsymbol{W}\boldsymbol{A}_3 \tag{5}$$

其中，$\boldsymbol{Q}_3$和$\boldsymbol{A}_3$是全连接层生成的重构文本向量，$\boldsymbol{W}$ 是维度为 $d\_f2\times d\_f2$ 的权重矩阵。这是一种双向注意力机制计算方式[24]，仅依赖简单的矩阵计算实现两个句子的交互关系。与之不同的是，本文为实现与双层感知机的结合，将其池化操作删去。模块网络的结构如图 3 所示。

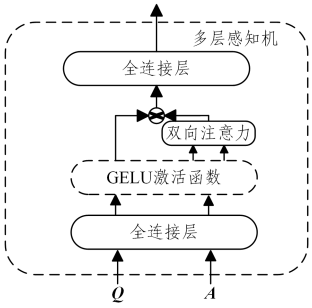


图 3　语义交互模块

Fig. 3　Semantic interaction module

语义交互模块中通过如下定义建模文本向量 $\boldsymbol{Q}$ 和 $\boldsymbol{A}$ 之间的交互关系，并映射生成问题句 $\boldsymbol{Q}$ 对应的文本向量。

$$\boldsymbol{Q}_3=\sigma_1(\boldsymbol{Q}\boldsymbol{W}_3+b_3) \tag{6}$$

$$\boldsymbol{Q}_4=\boldsymbol{Q}_3\times\sigma_3(\boldsymbol{M}_2) \tag{7}$$

$$\boldsymbol{Q}_v=\boldsymbol{Q}_4\boldsymbol{W}_4+b_4 \tag{8}$$

其中，σ_3是激活函数 softmax，问题句文本向量 $\boldsymbol{Q}$ 对应语义交互矩阵$\boldsymbol{M}_2$的维度为 $l\times l$，是基于文本向量 $\boldsymbol{Q}_3$和 $\boldsymbol{A}_3$生成的，而输出答案句文本向量对应的语义矩阵仅需将$\boldsymbol{Q}_3$和$\boldsymbol{A}_3$的位置调换实现计算即可。同样地，其参数在问题句和答案句之间共享。整个模块同样只涉及基本的矩阵计算和矩阵转置操作，模型参数量与维度 d 呈二次非线性关系，模块计算复杂度为线性。

4　实验与分析

本文将提出的模型与选定的基准数据集上的先进模型进行比较，以衡量本文模型的整体有效性。紧接着，统计了本文提出的两个基于多层感知机的网络模块的参数量。再者，在 WikiQA 数据集上开展了消融实验，以评估本文网络模块的独立作用。最后，通过一个案例详细说明本文模型的优势。

4.1　实验数据集和对比模型

本文在典型的答案选择任务数据集 WikiQA 和 SelQA 上开展实验。WikiQA 数据集是一个开放域问答数据集，其答案来源于维基百科，WikiQA 数据集中存在不包含任何正确答案的问题，本文与其他有关答案选择的研究一致[12-13,20-21,25]，将不包含正确答案的问题句剔除，最终 WikiQA 数据集分别包含 873 个、126 个和 243 个问题，每个问题对应多个答案句，共包含 8 672 个、1 130 个和 2 351 个问答对用于训练、验证和测试，其中答案句长度普遍比问题句要长，训练集问题句长度最长为 21，答案句长度大部分在 100 以内，超过 100 的答案句有 5 个。SelQA 数据集是专门为问答系统而创建的答案选择数据集，分别有 5 529 个、785 个和 1 590 个问题，共包含 66 438 对、9 377 对和 19 435 对问答对用于训练、验证和测试，是 WikiQA 数据集问答对的 7 倍以上，其中训练集问题句长度最长 44，答案句长度超过 100 的共有 205 个。

本文在 WikiQA 数据集上采用的对比模型可分为两大类，一类是基于冻结参数的 BERT 模型实现词编码的答案选择模型。

(1)$\text{BERT}_{\text{base}}$+Transformer Encoder[19]：BERT 模型结合 Transformer 编码层实现的答案选择模型。

(2)$\text{BERT}_{\text{base}}$+BiLSTM+AP：其来源于经典的答案选择模型 Attentive LSTM[7]，该模型其他网络结构包括 BiLSTM 和 Attentive Pooling Networks[24]（简称 AP），本文将其词编码工具替换成 BERT 模型，其他网络结构不变。

(3)$\text{BERT}_{\text{base}}$ + Transformer Encoder + AP：由于 $\text{BERT}_{\text{base}}$+Transformer Encoder 缺乏问答对之间的语义交互关系挖掘，本文在其基础上添加了答案选择任务领域提出的用于挖掘交互关系的 AP 网络结构。

另一类是其他答案选择模型。

(1)CAN[20]：基于问答对之间均被特征建模的协作对抗网络。

(2)MVFNN[25]：一个基于 BiLSTM 实现的多视图融合神经网络。

(3)Comp-Clip+LM+LC[12]：基于 ELMo 编码和潜在聚簇优化的比较聚合模型。

(4)WEHM[21]：基于外部知识丰富词义表达和注意力机制的答案选择模型。

(5)DAMPM[13]：基于动态注意力的比较聚合模型，引入多种匹配策略完成句子向量之间的信息交互。

考虑到本文模型是基于 BERT 模型实现的，在 SelQA 数据集上重点对比了所有基于 BERT 模型的答案选择模型和未采用 BERT 模型中性能表现最好的 WEHM 模型。

4.2　实验设置

本文模型是由 pytorch 深度学习框架实现的。动态词编码层所使用的 BERT 模型是由 Hugging Face 发布的 bert-base-uncased 版本，词向量维度 d 设为 768。我们统计了 WikiQA 和 SelQA 数据集上训练集问题句和答案句的长度，

99%的句子长度在100以内，因此设置句子向量的固定长度 l 为100。语义理解模块和语义交互模块中的全连接层的参数 d_f1 和 d_f2 均被设置为512。本文采用自适应矩阵估计(Adaptive Moment Estimation，Adam)优化器来实现网络参数更新，学习率设置为0.00005。与文献[19]中的基于参数冻结的BERT模型的答案选择方法一样，本文选择均值平方差损失函数来训练本文模型。

在实验中，与其他近期的答案选择任务研究工作一样[12,19-21,25]，本文选用了平均准确率均值(Mean Average Preision，MAP)和平均倒数排名(Mean Reciprocal Rank，MRR)作为本文模型性能的评价指标，用于衡量候选答案句的优先级排序正确性。

4.3 实验结果与分析

本文首先从模型整体的角度验证本文模型对于解决答案选择任务的有效性。本文模型在WikiQA和SelQA数据集上的实验结果与其他基线模型的对比结果如表1、表2所列。

表1 不同模型在WikiQA数据集上的实验结果

Table 1 Experimental results of different models on WikiQA dataset

对比模型	*MAP*	*MRR*
CAN(2018)	0.730	0.743
MVFNN(2018)	0.746	0.758
Comp-Clip+LM+LC(2019)	0.764	0.784
WEHM without WordNet knowledge(2020)	0.732	0.746
WEHM(2020)	0.770	0.788
DAMPM(2021)	0.761	0.772
基于BERT的模型		
$BERT_{base}$+Transformer Encoder(2020)	0.727	0.741
$BERT_{base}$+Transformer Encoder	0.754	0.770
$BERT_{base}$+BiLSTM+AP	0.730	0.747
$BERT_{base}$+Transformer Encoder+AP	0.767	0.779
Our model	**0.789**	**0.806**

从表1可以看出，本文提出的模型在WikiQA数据集上取得了显著的性能提升。相比借助外部知识的WEHM模型，本文模型在MAP指标上提升了1.9%，在MRR指标上提升了1.8%，表明本文模型能在一定程度上抗衡基于外部知识库实现的答案选择模型。与此同时，本文模型相比未添加外部知识的WEHM模型和表1中第一栏的其他对比模型展现了较大的优势。

由表1还可以看出，在WikiQA数据集上，与以冻结参数的BERT模型作为词编码工具的答案选择模型相比，本文模型具有一定优势。其中对比模型 $BERT_{base}$+Transformer Encoder保留了全部的WikiQA训练集中20360个问答对，而在验证集和测试集删除了不包含正确答案的问题句，因此，其训练集规模是本文和其他对比模型所用训练集8762个问答对的2倍以上。为了保证对比的公平性，本文通过复现 $BERT_{base}$+Transformer Encoder模型，使用与本文相同的训练集来训练，其结果如表1第2栏第2行所列。复现的 $BERT_{base}$+Transformer Encoder模型的性能表现优于文献[19]给出的数据，主要的原因可能是训练集数量过多导致过拟合，从而准确度降低。本文提出的语义理解模块和语义交互模块的网络结构相比自注意力网络结构(Transformer Encoder)更适合解决答案选择任务。这主要是因为Transformer Encoder的网络结构中的自注意力机制仍然是以词级维度为出发点，去探索句子中词和词之间的权重关系，而这并不符合BERT模型生成的文本向量特征。本文的语义理解模块主要探索动态文本向量的句级维度特征，首先通过多层感知机作用于单个词向量，多层感知机能够通过训练形成新的统一的句级语义空间，其次通过线性模型挖掘每一个句级维度上的潜在语义。除此之外，$BERT_{base}$+Transformer Encoder的模型结构忽略了文本对的词间交互关系，因此本文在该模型的基础上添加了AP网络结构，$BERT_{base}$+Transformer Encoder+AP在MAP和MRR性能指标上都带来了略微的提升，但其结果与本文模型相比，仍然存在一定差距，这体现了本文的语义交互模块能够更好地应用词与词之间的相互作用。

表2 不同模型在SelQA数据集上的实验结果

Table 2 Experimental results of different models on SelQA dataset

对比模型	*MAP*	*MRR*
WEHM without WordNet knowledge(2020)	0.849	—
WEHM(2020)	0.917	0.922
基于BERT的模型		
$BERT_{base}$+Transformer Encoder	0.870	0.875
$BERT_{base}$+BiLSTM+AP	0.868	0.876
$BERT_{base}$+Transformer Encoder+AP	0.873	0.877
Our model	0.903	0.911

从表2可以看出，SelQA数据集上的实验结果进一步验证了本文模型在答案选择任务上的有效性。本文模型相比未添加外部知识的WEHM模型和所有基于BERT的模型都展现了较大的优势，但是与添加了外部知识的WEHM模型还存在一定差距，在MAP和MRR指标上分别相差1.4%和1.1%。根据表1和表2，本文模型相比添加了外部知识的WEHM模型在WikiQA数据集上的表现更优，而在SelQA数据集上表现不佳，通过分析主要有两方面原因。首先，SelQA数据集涉及的问答内容主题众多，包括艺术、城市、历史事件等，且语料库体量庞大，其语料库内容的丰富性和主题领域的多样性远超WikiQA数据集。其次，添加了外部知识的WEHM模型能够借助WordNet中丰富多样的语义信息对问答对进行建模，涉及的领域越丰富，其优势越明显。本文模型与未添加外部知识的WEHM模型相比，在WikiQA和SelQA两个数据集的MAP性能指标上分别有6.9%和5.4%的提升。总的来说，对于语料库内容主题和领域范围较大的答案选择任务，选取适合的外部知识辅助能够提高模型性能表现的上限，而本文提出的模型更适用于面向通用领域的问答系统。在基于BERT的模型中，本文模型具有一致的优势，进一步表明了语义理解模块和语义交互模块在答案选择任务上的有效性。

4.4 参数量统计

由于主流答案选择模型所应用的方法各异，为真实衡量本文模型在参数量上的表现，选择统计与本文方法一致的答案选择模型，即基于BERT的答案选择模型，参数量的衡量不包含BERT模型，其计算方法通过python程序实现，结果如表3所列。表3中，本文模型的语义理解模块的参数量为794432，而语义交互模块为1050880，构成了本文用于更新的参数量1845312。计算其他3个对比模型的参数量时，都基于本文模型动态词编码层所生成的文本向量维度。对于Transformer Enocder，其层数设置为1，与文献[19]一致。

BiLSTM 的隐藏单元设置为 256，与文献[7]一致。从表中可以看出，本文模型在取得性能提升的同时，其参数量与计算成本也具有巨大优势。

表 3　网络模块参数量对比

Table 3　Comparison of network module parameters

模型	参数量
Transformer Encoder(2020)	3 548 160
BiLSTM	2 101 248
AP	591 360
Our model	1 845 312

4.5　消融实验

为了更好地探究本文模型的真实性和有效性，本节将在 WikiQA 数据集上开展消融实验，分析本文模型各个网络模块的独立作用。

首先，对于语义理解模块，本文选用的对比模型包括 BiLSTM[7] 和 Transformer Encoder[19]，BiLSTM 来源于经典的答案选择模型 Attentive LSTM 中用于实现语义理解的网络模块，其中 BiLSTM 的隐藏单元设置为 256，同时，本文实验过程中将其固定词向量技术 word2vec 替换成 BERT 模型，而 Transformer Encoder 则是目前自然语言处理技术中公认的能够很好地实现语义理解的网络结构，其参数设置与文献[19]一致。本文模型的实现仅包括语义理解模块，其结果如表 4 所列。

表 4　语义理解模块与其他对比模型的消融实验

Table 4　Ablation experiments of semantic understanding module and other contrasting models

模型	*MAP*	*MRR*
$BERT_{base}$＋BiLSTM	0.642	0.661
$BERT_{base}$＋Transformer Encoder	0.754	0.770
$BERT_{base}$＋语义理解模块	0.772	0.788

从表 4 可以看出，本文的语义理解模块效果更好，这是因为 BiLSTM 和 Transformer Encoder 主要是为了挖掘一个句子中词与词之间的影响，从而实现语义建模，而 BERT 模型提供的文本向量很好地考虑了该特征。语义理解模块从句级维度实现特征挖掘的设计更适用于动态文本向量。BiLSTM 模型的结果不占优势，主要原因是缺乏句子维度语义空间的重构，而语义理解模型不仅注重于句子级别语义空间建模和句级维度语义挖掘，而且展现了更大的优势。

其次，对于语义交互模块，以挖掘问题句和答案句词与词之间的交互关系为目标，本文选用 AP 作为对比模型，该网络结构是经典的以挖掘问答对的交互信息为目的而提出的网络结构，以衡量性能优势，其参数设置与本文 BERT 模型参数相匹配，本实验的语义交互模块及对比模型全都直接作用于 BERT 模型生成的文本向量之上。其实验结果如表 5 所列。

表 5　语义交互模块与其他对比模型的消融实验

Table 5　Ablation experiments of semantic interaction modules and other contrasting models

模型	*MAP*	*MRR*
$BERT_{base}$＋AP	0.530	0.534
$BERT_{base}$＋语义交互模块	0.773	0.790

从表 5 可知，相比 AP，语义交互模块具有非常大的优势，其交互矩阵的计算方法与 AP 一致，而语义交互模块的计算是基于句级维度语义重建后的文本向量，实现重建的关键在于多层感知机。虽然多层感知机的添加导致语义理解模块与 AP 相比参数量略大，但本文提出的语义交互模块很好地贴合了 BERT 模型所生成的文本向量的特征，能够更好地挖掘问答对之间的交互关系。

4.6　案例研究

答案选择任务本质上是通过自然语言理解实现候选答案句优先级的排序，本节选择了一个案例开展研究，通过给出一个问题和各个候选答案句的详细排名，而非全局的 MAP 和 MRR 性能指标，来说明各个模型的真实情况。对比模型包括 BiLSTM＋AP，Transformer Encoder 和 Transformer Encoder＋AP，这些对比模型都采用了与本文一致的动态词编码层，其结果如表 6 所列。

表 6　案例在各个对比模型中的优先级排名

Table 6　Case priority ranking in each comparison model

问题句：what can be powered by wind? 候选答案句	对比模型				标签
	BiLSTM＋AP	Transformer Encoder	Transformer Encoder＋AP	Our model	
1 burbo bank offshore wind farm at the entrance to the river mersey in north west England	3	3	2	4	0
2 the shepherds flat wind farm is a 845megawatt mw wind farm in the us state of oregon	4	1	3	3	0
3 wind power is the conversion of wind energy into a useful form of energy such as using wind turbines to make electrical power windmills for mechanical power wind pumps for water pumping or drainage or sails to propel ships	5	2	1	1	1
4 wind power as an alternative to fossil fuels is plentiful renewable widely distributed clean produces no greenhouse gas emissions during operation and uses little land	2	4	4	2	0
5 wind power is veryconsistent from year to year but has significant variation over shorter time scales	1	5	5	5	0

表 6 中最后一列的标签是数据给出的真实标签，标签 1 表示该答案句可以回答问题句所提的问题，而标签 0 则代表不能。表 6 中对比模型中的结果表示各个模型预测出候选答案句的优先级，优先级由 1 至 5 逐渐降低。其中本文模型和 Transformer Enocder＋AP 模型相比 BiLSTM＋AP 和 Transformer 而言，都将标签为 1 的正确候选答案句 3 排在了第一位，这说明本文模型能够选出正确的答案句，各个模型排名情况也符合所有模型整体的 MAP 和 MRR 性能指标趋势。

结束语 本文基于BERT模型，实现了一个动态词编码层，用于替换传统答案选择模型中的编码层，并分析其所生成的动态文本向量特点，结合答案选择任务的需求，提出了一个新颖的基于多层感知机的答案选择模型全局架构，该模型包括语义理解模块和语义交互模块。本文在公开的标准答案选择WikiQA和SelQA数据集上进行了评测，实验结果表明，本文模型相比其他基线模型，不仅性能表现上具有一定的优势，而且参数量更少，计算成本更低。

除此之外，本文在实验过程中发现了在自然语言处理领域，外部知识对于基于网络的模型的释义和推理能力的重要性。下一步的工作将在模型中引入外部知识的帮助，重点研究知识和数据的匹配，在答案选择任务中有效使用外部知识来进一步提升准确率。

参 考 文 献

[1] TAN M, DOS SANTOS C, XIANG B, et al. Improved representation learning for question answer matching[C]// Proceedings of the 54th Annual Meeting of the Associationfor Computational Linguistics. 2016: 464-473.

[2] MIN S, ZHONG V, SOCHER R, et al. Efficient and robust question answering from minimal context over documents[J]. arXiv: 1805.08092, 2018.

[3] PENNINGTON J, SOCHER R, MANNING C D. Glove: Global vectors for word representation[C]// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014: 1532-1543.

[4] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[J]. arXiv: 1301.3781, 2013.

[5] LIU R H, YE X, YUE Z Y. Review of pre-trained models for natural language processing tasks[J]. Journal of Computer Applications, 2121, 41(5): 1236-1246.

[6] QIU X, SUN T, XU Y, et al. Pre-trained models for natural language processing: A survey[J]. Science China Technological Sciences, 2020, 63(10): 1872-1897.

[7] TAN M, DOS SANTOS C, XIANG B, et al. Improved representation learning for question answer matching[C]// Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. 2016: 464-473.

[8] CHEN Q, HU Q, HUANG J X, et al. Enhancing recurrent neural networks with positional attention for question answering [C]// Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2017: 993-996.

[9] HUANG J. A Multi-Size Neural Network with Attention Mechanism for Answer Selection[J]. arXiv: 2105.03278, 2021.

[10] WANG S, JIANG J. A compare-aggregate model for matching text sequences[J]. arXiv: 1611.01747, 2016.

[11] BIAN W, LI S, YANG Z, et al. A compare-aggregate model with dynamic-clip attention for answer selection[C]// Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. 2017: 1987-1990.

[12] YOON S, DERNONCOURT F, KIM D S, et al. A compare-aggregate model with latent clustering for answer selection[C]// Proceedings of the 28th ACM International Conference on Information and Knowledge Management. 2019: 2093-2096.

[13] LI Z C, TURDI T, ASKAR H. Answer selection model based on dynamic attention and multi-perspective matching[J]. Journal of Computer Applications, 2021, 41(11): 3156-3163.

[14] PETERS M E, NEUMANN M, IYYER M, et al. Deep contextualized word representations[J]. arXiv: 1802.05365, 2018.

[15] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]// Advances in Neural Information Processing Systems. 2017: 5998-6008.

[16] RADFORD A, NARASIMHAN K, SALIMANS T, et al. Improving language understanding by generative pre-training[J/OL]. [2022-06-19]. http://cs.ubc.ca/~amuham01/LING530/papers/radford2018improving.pdf.

[17] DEVLIN J, CHANG M W, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding [J]. arXiv: 1810.04805, 2018.

[18] LIU Y, OTT M, GOYAL N, et al. Roberta: A robustly optimized bert pretraining approach[J]. arXiv: 1907.11692, 2019.

[19] LASKAR M T R, HUANG X, HOQUE E. Contextualized embeddings based transformer encoder for sentence similarity modeling in answer selection task[C]// Proceedings of The 12th Language Resources and Evaluation Conference. 2020: 5505-5514.

[20] CHEN Q, HU Q, HUANG J X, et al. Can: Enhancing sentence similarity modeling with collaborative and adversarial network [C]// The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. 2018: 815-824.

[21] LI W, WU Y. Exploiting WordNet Synset and Hypernym Representations for Answer Selection[C]// Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing. 2020: 106-115.

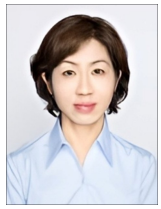
[22] TOLSTIKHIN I O, HOULSBY N, KOLESNIKOV A, et al. Mlp-mixer: An all-mlp architecture for vision[J]. Advances in Neural Information Processing Systems, 2021, 34: 24261-24272.

[23] LIU H, DAI Z, SO D, et al. Pay attention to MLPs[J]. Advances in Neural Information Processing Systems, 2021, 34: 9204-9215.

[24] SANTOS C, TAN M, XIANG B, et al. Attentive pooling networks[J]. arXiv: 1602.03609, 2016.

[25] SHA L, ZHANG X, QIAN F, et al. A multi-view fusion neural network for answer selection[C]// Thirty-second AAAI Conference on Artificial Intelligence. 2018.

LUO Liang, born in 1998, postgraduate. His main research interests include deep learning and natural language processing.



CHENG Chunling, born in 1972, professor, is a member of China Computer Federation. Her main research interests include data mining and data management

（责任编辑：喻藜）