融合语义信息与问题关键信息的多阶段注意力答案选取模型

张仰森[1)] 王 胜[1)] 魏文杰[1)] 彭媛媛[2)] 郑 佳[2)]

[1)](北京信息科技大学智能信息处理研究所 北京 100101)

[2)](中国科学院软件研究所 北京 100190)

摘 要 自动问答系统可以帮助人们快速从海量文本中提取出有效信息,而答案选取作为其中的关键一步,在很大程度上影响着自动问答系统的性能.针对现有答案选择模型中答案关键信息捕获不准确的问题,本文提出了一种融合语义信息与问题关键信息的多阶段注意力答案选取模型.该方法首先利用双向 LSTM 模型分别对问题和候选答案进行语义表示;然后采用问题的关键信息,包括问题类型和问题中心词,利用注意力机制对候选答案集合进行信息增强,筛选 Top K 个候选答案;然后采用问题的语义信息,再次利用注意力机制对 Top K 个候选答案集合进行信息增强,筛选出最佳答案.通过分阶段地将问题的关键信息和语义信息与候选答案的语义表示相结合,有效提高了对候选答案关键信息的捕获能力,从而提升了答案选取系统的性能.在三个数据集上对本文所提出的模型进行验证,相较已知同类最好模型,最高性能提升达 1.95%.

关键词 答案选取;语义信息;关键信息;相似度计算;多阶段注意力机制

中图法分类号 TP391 **DOI 号** 10.11897/SP.J.1016.2021.00491

An Answer Selection Model Based on Multi-Stage Attention Mechanism with Combination of Semantic Information and Key Information of the Question

ZHANG Yang-Sen[1)] WANG Sheng[1)] WEI Wen-Jie[1)] PENG Yuan-Yuan[2)] ZHENG Jia[2)]

[1)](*Institute of Intelligent Information Processing, Beijing Information Science and Technology University, Beijing* 100101)

[2)](*Institute of Software, Chinese Academy of Science, Beijing* 100190)

Abstract With the rapid development of Internet technology, the amount of text information in the network increases exponentially, hence people usually use some search engines to retrieve the required information from mass data. A search engine can be regarded as a special question answering system. When a question is given, the general processing flow of the automatic question answering system is as follows: first, the system analyzes the question to obtain its type, semantics and other relevant information; then, select a candidate answer set from the answer database according to the analysis results; finally, the system will rearrange the candidate set with various sorting techniques and select the best answer or the text with the best answer to return to the user. The flow shows that the selection effect of the best answer will directly affect the overall performance of the automatic question answering system. Traditional answer selection models usually use lexical or syntactic analysis and artificial constructing feature to select answers, which is difficult to capture the semantic association information between questions and candidate answers. With the development of deep learning technology, researchers applied the deep

收稿日期:2019-12-25;在线发布日期:2020-05-26.本课题得到国家自然科学基金(61772081)资助.**张仰森**,博士,教授,中国计算机学会(CCF)杰出会员,主要研究领域为自然语言处理、人工智能. E-mail: zhangyangsen@163.com. **王 胜**,硕士研究生,主要研究方向为自然语言处理. **魏文杰**,硕士研究生,主要研究方向为自然语言处理. **彭媛媛**,硕士,工程师,主要研究方向为自然语言处理. **郑 佳**,硕士,工程师,主要研究方向为自然语言处理.

learning framework into the answer selecting task, use the neural network model to obtain the semantic association information of the question and the candidate answer, and evaluate the matching association degree between them, then select the answer with the strongest matching relationship as the best answer. Because the selection of answers depends entirely on the information carried in the question, researchers often generate attention vector from the question semantic information to update the semantic representation of the candidate answers. Although this kind of attention model can strengthen the semantic relationship between the question and the candidate answer, it ignores the relationship of key information between them, therefore, the effectiveness of such models is affected. For different types of questions, the concerned content in best answers is often different. For example, when asking time-related questions, the best answer should be more focused on the key information of time or the information with strong time semantic association; when asking weather-related questions, the best answer should pay more attention to the key information related to weather. Also, the existing attention-based answer selection methods often establish the model of questions and answers at the same stage, which is not easy to capture the differences between the various candidate answers. To solve the problem that the answer key information capture is not accurate in the existing answer selection model, this paper proposes an answer selection model based on a multi-stage attention mechanism with a combination of semantic information and key information of the question. Firstly, this method uses a bi-directional LSTM model to represent questions and candidate answers semantically. Then the key information of the question, including the type of question and the headword of the question, is used to enhance the information of the candidate answer by attention mechanism, and the Top K candidate answers are selected. Finally, the attention mechanism with semantic information of the question is used again to enhance the information of the Top K candidate answer set to select the best answer. By combining the key information and semantic information of the question to enhance the semantic representation of the candidate answer in multi-stages, the ability to capture the key information of candidate answers is effectively improved, and the performance of the answer selection system is improved. The experimental results on three datasets show that the highest performance improvement is up to 1.95% compared with the other state of the art models.

Keywords answer selection; semantic information; key information; similarity computing; multi-stage attention mechanism

1 引 言

随着互联网技术的快速发展,网络中的文本信息量呈指数级增长,成为了人们获取信息的重要来源,因此,利用搜索引擎从海量信息中检索出所需的信息成为了人们获取信息的主要方式.然而,现有搜索引擎的检索策略大多是基于字符串匹配的,缺乏从语义角度挖掘知识的能力,导致搜索到的结果精度差,冗余度高[1],还需要用户从大规模搜索结果中进一步理解和筛选才能够获取到真正需要的信息,这与用户快速准确获得信息的需求还有一定的差距.随着文本处理与理解技术的快速发展和广泛应用,能够更好地满足用户需要的智能问答技术也逐步成熟,并催生了一批智能助手的问世,例如小米公司的小爱、苹果公司的 Siri、微软公司的小冰等.这些智能助手与传统的搜索引擎相比,更贴近用户的实际需求,他们都力求从语义层面分析用户的问题,精准定位用户的意图,从而快速、有效、准确地为用户提供所需的信息.

当给定一个问题时,自动问答系统一般的处理流程如下:首先,分析问题以获取问题的类型、语义

等相关信息；然后，依据分析结果在数据集中筛选出候选答案集合；最后在候选集合中采用各种排序技术进行重排，筛选出最佳答案或含有最佳答案的文本返回给用户. 因此，最佳答案的选取效果将直接影响到自动问答系统的整体性能，优化最佳答案的选取策略可以有效地提升自动问答系统为用户服务的能力. 本文将围绕该问题展开深入研究，以进一步提升最佳答案的选取效果.

传统的答案选取模型[2]大多利用词法或句法分析以及人工构造特征的方法来选取答案，这类方法较难捕捉到问题与候选答案之间的语义关联信息. 随着深度学习技术的发展，研究者们将深度学习框架引入到答案选取任务中来，利用神经网络模型获取问题和候选答案的语义关联信息，并对它们之间的匹配关联程度进行评估，进而选取匹配关系最强的答案作为最佳答案. 由于答案的选取完全依赖于问题所传递的信息，因此，在基于深度学习的答案选取模型中，研究者们往往会利用问题的语义信息生成注意力向量，以此来更新候选答案的语义表示，优化问题与候选答案之间匹配关系的评估效果. 这类引入注意力的模型虽然能够强化问题与候选答案之间语义关联的程度，但是在一定程度上忽略了两者之间关键信息的联系，从而影响其问题和答案的建模效果. 因为对于不同类型的问题，其最佳答案中关注的内容往往有所不同，例如询问时间相关的问题时，其最佳答案表示中应更注重于表示时间的关键信息或者与时间语义关联较强的信息；询问天气相关的问题时，其最佳答案应更注重于表示天气相关的关键信息或者与天气关联较强的信息. 另外，现有的基于注意力的答案选取模型往往将问题和答案的建模放在同一阶段进行，这对从多个候选答案中选取一个最佳答案的答案选取任务来说，不容易捕捉到答案相互之间的差异.

针对现有答案选取模型的以上问题，本文在语义注意力的基础上，提出了一种融合语义信息与问题关键信息的多阶段注意力答案选取模型(Multi-Stage Attention Answer Selection Model Combining Semantic Information and Key Information of the Question，MSAAS with KI-SI)，分阶段地将问题的关键信息和问题的语义信息以注意力机制的方式对候选答案进行信息增强，以增加对候选答案中的关键信息的捕获能力，解决在问题和答案的建模过程中，对候选答案关键信息捕获不足的问题，以此来提升答案的选取效果.

2　相关工作

2.1　答案选取相关工作

答案选取是自动问答技术的关键技术之一，其相关技术也可以用于文本理解、信息检索、智能服务等多个领域. 针对自动问答系统中的答案选取问题，以往的研究者们通常将其视为分类任务和相似度计算任务两种类型的问题进行解决. 基于分类的答案选取任务是依据问题与候选答案之间的关联关系，将候选答案分到正确或错误类别，将正确类别中的答案作为最佳答案. 基于相似度计算的答案选取任务是通过计算问题与候选答案之间的相似度，选取相似度最高的答案作为最佳答案. 为了能够有效提升答案选取的效果，大多研究学者都致力于研究问题与候选答案之间相关关系的表示，主要的研究工作可分为两个阶段：第一阶段是基于语言学知识和特征工程的答案选取方法，第二阶段是基于深度学习的答案选取方法.

基于语言学知识和特征工程的答案选取方法主要是结合外部资源对问题、候选答案进行词法、句法分析进而选取答案. 例如 Surdeanu 等人[3]提取了问题与候选答案的词频、词语之间的相似度等多种特征对候选答案进行排序，从而选出最佳答案. Yih 等人[4]利用 WordNet 来获取问题和候选答案的语义特征，以改进候选答案的排序效果. Tymoshenko 等人[5]对问题和答案的句法结构、语义结构进行分析，并利用 YAGO、DBpedia 和 WordNet 等知识库挖掘候选答案中与问题匹配的信息，最终实现答案段落的排序. 虽然这些答案选取方法都能捕捉到问题与候选答案之间的匹配关系，但是它们性能的好坏与提取特征的质量、采用的外部资源有很大关系，同时在实际的运用过程中，也需要一定的领域知识和较高的人工成本.

随着深度学习的发展，神经网络模型逐渐被引入到答案选取任务中并成为主流方法. 例如 Feng 等人[6]利用 CNN 模型分别对问题和候选答案进行语义表示，然后采用余弦相似度、Geometric mean of Euclidean and Sigmoid Dot product(GESD)和 Arithmetic mean of Euclidean and Sigmoid Dot product(AESD)三种方法对问题和候选答案的语义表示向量进行相似度计算，最后选取相似度最高的答案作为最佳答案，实验表明利用 GESD 的相似度计算方法取得了最好的效果. Guo 等人[7]利用余弦

相似度的计算方法对问题与候选答案中词语之间的相似度进行评估,然后将词语之间的相似程度和词语的词向量一同输入到 Skip-CNN 模型中,分别获取问题和候选答案的语义表示向量,最后将二者的语义表示向量进行拼接,利用 Softmax 对候选答案进行分类以选取问题的最佳答案. Tan 等人[8]采用 BiLSTM 对问题和候选答案进行语义编码,然后将问题的语义作为注意力对候选答案的编码进行加权更新,最后取相似度最高的候选答案作为最佳答案. 相比于基于语言学知识和特征工程的答案选取方法,基于深度学习的方法减少了对领域知识和外部因素的依赖,具有较强的通用性. 此外,这类方法能够在语义层面学习问题和候选答案之间的语义匹配关系,使得答案选取效果有了明显的提升.

在上述研究中,虽然已有的方法将词频、词语相似度等词级别的特征引入到了候选答案的语义表示中,但是对候选答案中的关键信息以及问题与候选答案之间的关联关系的捕捉能力有限. 因此本文在语义信息的基础上,试图融入问题关键信息,提出了一种融合语义信息与问题关键信息的多阶段注意力答案选取模型,提升候选答案中关键信息的捕获能力,优化候选答案的语义表示,从而更加全面地捕捉问题与候选答案之间的关联关系,以此来提升答案选取的准确率.

2.2 注意力机制相关工作

注意力机制[9]可以抽象为针对性地提高数据中特定位置的关注度,注意力机制最早被应用于图像领域,用以关注重点区域的重点信息. Bahdanau 等人[10]最早将注意力机制引入到 NLP 任务中,尝试在机器翻译过程中将目标端的输出与源端的输入进行对齐,从而提升机器翻译的效果[11-12]. 随后根据不同任务提出了各种注意力机制,例如 Cheng 等人[13]在机器阅读任务中提出了单向的自注意力机制,用以学习当前词语与句中前面部分词语之间的相关性;Vaswani 等人[14]对注意力机制进行了改进,抛弃了传统的 RNN 结构并提出完全基于自注意力机制的 Transformer 模型,解决了数据计算无法并行化的问题,极大地提高了计算效率;He 等人[15]和 Yu 等人[16]发现在推荐任务中,注意力机制可以有效地捕捉用户长期兴趣与短期兴趣,提高推荐系统的准确性.

在问答系统以及答案选取任务中,Tan 等人[8]基于 BiLSTM+CNN 的架构,采用注意力机制分别对问题和候选答案进行语义表示,并采用余弦相似度进行融合,证明了仅引入字级别的自注意力机制的模型就能起到很好的效果;Bachrach 等人[17]提出了一种针对答案选取任务的新注意力机制,该方法将问题语义和候选答案词频特征相结合共同加强候选答案中关键词在语义表示中的权重,使候选答案的语义表示向量更加准确,从而提升了答案选取任务的性能;Xu 等人[18]提出了一种基于门组自注意力(Gated Group Self Attention,GGSA)的答案选取模型,该模型很好地解决了全局注意力和局部注意力不能被很好区分的问题.

现有注意力机制在答案选取任务中的运用大多采用问题的信息对答案进行注意力增强,从而将问题和答案的建模放在同一阶段进行,这不利于从多个维度对候选答案的关键信息进行捕获,从而导致对于多个候选答案之间差异性的捕获能力有限. 为提升多维度信息的捕获能力,Chen 等人[19]在阅读理解任务中提出了一种两阶段的通用框架,首先使用经过 tf-idf 与 bigram 结合的检索方法找到与问题相关的文章,其次通过特征工程对段落及问题进行编码,构建阅读理解模型从文章段落中找到对应的答案,最终在多任务集上使用远程监督的方法提高了计算性能. Hao 等人[20]在问答任务中提出了一种基于端到端的问答网络模型,主要利用交叉注意力机制对问题和答案进行互相关注. 一方面利用答案信息强化问题的语义表示;另一方面利用问题信息对答案进行不同的关注. 同时,通过将外部知识库信息引入到 Embedding 中,缓解了未登录词的问题,使模型更有效地表示了问题和答案,提高了端到端模型的实验性能. 因此,本文将问题语义信息和问题关键信息分为多个阶段对候选答案进行信息增强,以此来加强模型对候选答案关键信息的捕获能力,提升对类似答案之间差异的判断能力.

3 方 法

自动问答系统的答案选取过程可以形式化为如下形式:给定问题 Q,在相应的候选答案集合 $\{A_1, A_2, \cdots, A_z\}$ 中寻找与问题 Q 最匹配的答案即最佳答案 $\{A_{\text{best}} \mid \text{best} \in \{1, 2, \cdots, z\}\}$,其中,$z$ 为候选答案的个数. 本文将答案选取任务分为两个部分:问题与候选答案的相关度计算和最佳答案的选取. 对于问题与候选答案的相关度计算部分,在问题语义信息关联关系的基础上引入问题关键信息,包括问题类型

和问题中心词两个维度，构建了一种融合语义信息与问题关键信息的多阶段注意力答案选取模型；对于问题最佳答案的选取，利用问题与候选答案之间的相关度，选出相关度最高的答案作为最佳答案，其中相关度采用问题和候选答案的语义编码向量的余弦相似度进行计算.

3.1　答案选取的基础模型

答案选取的基础模型的主要架构如图 1 所示，主要由问题与候选答案的语义表示层、语义抽象层和相关度计算层组成.

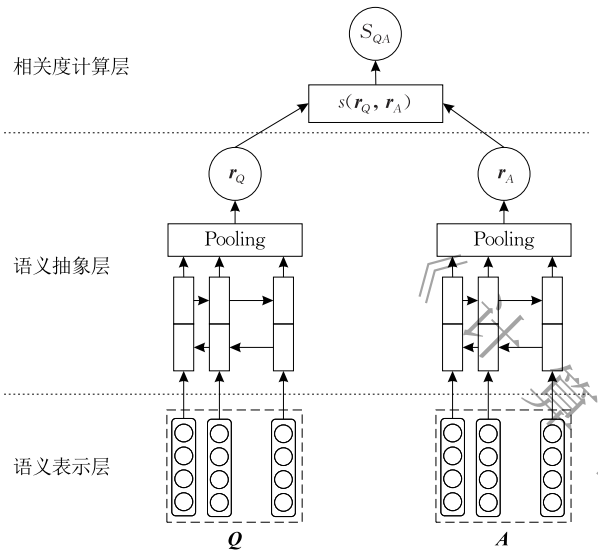

图 1　问题与候选答案相关度计算基础模型

(1) 语义表示层. 利用问题和候选答案所包含词语信息的词向量，分别对问题和候选答案进行语义表示，得到问题的语义表示 $\boldsymbol{Q}(\boldsymbol{w}_{q_1}, \boldsymbol{w}_{q_2}, \cdots, \boldsymbol{w}_{q_{n_Q}})$ 和候选答案的语义表示 $\boldsymbol{A}(\boldsymbol{w}_{a_1}, \boldsymbol{w}_{a_2}, \cdots, \boldsymbol{w}_{a_{n_A}})$，其中 n_Q 和 n_A 分别为问题和候选答案的词语个数，$\boldsymbol{w}_{q_x}, \boldsymbol{w}_{a_y} \in R^{1\times d}(1\leqslant x\leqslant n_Q, 1\leqslant y\leqslant n_A)$ 分别为问题的第 x 个词语的词向量和候选答案的第 y 个词语的词向量，且词向量的维度为 d.

(2) 语义抽象层. 采用 Bi-LSTM＋Pooling 对输入的问题和候选答案语义表示的上下文进行语义编码，分别得到问题和候选答案的语义表示 $\boldsymbol{r}_Q$ 和 $\boldsymbol{r}_A$.

(3) 相关度计算层. 利用余弦相似度计算问题和答案的语义表示 $\boldsymbol{r}_Q$ 和 $\boldsymbol{r}_A$ 之间的相似度 S_{QA} 作为问题和答案的相关程度的度量.

这一基础模型只是将问题和候选答案之间的语义信息进行相似度计算，但是对于问题而言，它在与候选答案进行相似度计算时，更希望候选答案中与问题相关的部分占有更高的权重，与问题不相关部分占有较低的权重.

3.2　基于问题语义信息注意力的信息增强模型

通过语义表示层和语义抽象层，可获问题的语义表示向量 $\boldsymbol{r}_Q$，这一向量较全面地包含了问题的上下文语义信息，利用问题的语义表示，采用注意力机制，对候选答案的语义信息进行增强，使得候选答案中与问题相关度较高的部分所占权重更高，以此来构建候选答案针对当前问题语义信息的表示，进而提升候选答案与问题语义的相关性. 基于问题语义信息注意力的信息增强模型的框架如图 2 所示.

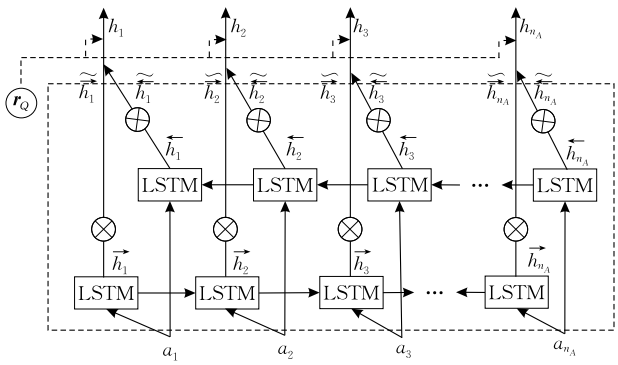

图 2　基于问题语义信息注意力的信息增强框架

基于问题语义信息注意力的信息增强主要利用问题的语义信息 $\boldsymbol{r}_Q$ 对候选答案的 LSTM 输出进行注意力加权更新，强化候选答案中与问题有关的部分. 在 LSTM 中，对每一时刻节点的正向输出 $\overrightarrow{h_{q_i}}$ 与反向输出 $\overleftarrow{h_{q_i}}$ 进行拼接，得到语义编码 h_{q_i}，h_{q_i} 同时包含当前时刻的上文信息与下文信息. 组合 LSTM 各个时刻的输出，得到问题的语义编码矩阵 $\boldsymbol{M}_Q = [h_{q_1}, h_{q_2}, \cdots, h_{q_n}]^{\mathrm{T}} = [d_{q_1}, d_{q_2}, \cdots, d_{q_m}]$. 其中 n 为 LSTM 展开的时间步数，m 为 LSTM 隐藏单元个数的 2 倍. 对问题的语义编码矩阵进行压缩，得到问题的语义信息 $\boldsymbol{r}_Q$，如式(1)所示.

$$\boldsymbol{r}_Q = [\max(d_{q_1}), \max(d_{q_2}), \cdots, \max(d_{q_m})] \tag{1}$$

同理将答案的每一时刻的 LSTM 正向和反向输出拼接得到每一时刻的候选答案的语义编码 h_{a_i}，将 $\boldsymbol{r}_Q$ 与 h_{a_i} 进行余弦相似度计算，将计算结果作为 $\boldsymbol{r}_Q$ 对 h_{a_i} 的关注权重 S_{weight_i}，计算公式如式(2)所示.

$$S_{weight_i} = \frac{\boldsymbol{r}_Q \cdot h_{a_i}}{|\boldsymbol{r}_Q||h_{a_i}|} \tag{2}$$

利用 S_{weight_i} 对 LSTM 每一时刻隐藏单元的输出 h_{a_i} 进行加权更新，计算公式如式(3)所示.

$$h'_{a_i} = S_{weight_i} h_{a_i} \tag{3}$$

将加权后的 h'_{a_i} 作为最终每个时刻的输出.

3.3　基于问题关键信息注意力的信息增强模型

本文采用问题类型和中心词作为问题的关键信息，利用注意力机制对候选答案进行信息增强.

3.3.1 基于问题类型的关键信息注意力

问题类型对候选答案的选取有十分重要的指导作用,对待同一个候选答案,不同类型的问题对候选答案中的关注点有所不同.例如对于表1所示的候选答案,当提问"When do an auto insurance premium go up?"时,候选答案中希望更加关注于"next renewal period"和"monthly quarterly semiannually annually"等表示时间的词语;当提问"Which factors affect the auto insurance premium?"时,候选答案会更希望关注于"activity or claim ticket and accident"等表示实物的词语.

表1 候选答案示例

your auto insurance premium will typically not change until your next renewal period depending on your payment term this typically can be monthly quarterly semiannually annually your premium be affect many factor the primary factor be your activity or claim ticket and accident be the thing that may cause your rate increase your rate can also be affect many other thing as insurance rate be typically determine the amount of risk the insurance company be bear in that market find out more contact your local agent and discuss your question about rate with them as each state, company and policy can vary greatly.

因此,我们对数据集中问题的类型和其最佳答案进行了分析,总结了7种问题的类型以及该类型问题的特征和常见的最佳答案类型,如表2所示.

表2 问题的类型、特征及答案常见类型

问题类型	问题特征	常见答案类型
人物问句	以"who、whom、whose"开头	多为与人物相关的信息
地点问句	以"where"开头	多为与地点相关的信息
时间问句	以"when"开头	多为与时间相关的信息
实物问句	以"what、which"开头	多集中于表达实物类信息的部分
数量问句	以"how much、how many、how long、how old、how far"开头	多集中于表达量词信息的部分
原因问句	以"why,how"开头	多集中于描述原因及动作的部分
其他问句	除以上六类以外	多集中于陈述事实类的部分

不同类型的问题对候选答案中关注的部分有所不同,参照语义信息增强的方法,提取问题的类型,构建类型的表示,作为一种注意力向量,引入到候选答案的语义信息表示中.具体来说,在模型初始化时,为每一种问题类型分别设定一个表示向量 $\boldsymbol{v}_{QT}$,利用 $\boldsymbol{v}_{QT}$ 对候选答案的LSTM输出进行注意力加权更新,强化候选答案中与问题类型有关的部分.计算流程为,将答案的每一时刻的LSTM正向和反向输出拼接得到每一时刻的候选答案的语义编码 h_{a_i},利用与式(2)相同的方法将 $\boldsymbol{v}_{QT}$ 与 h_{a_i} 进行相似度计算,得到 $\boldsymbol{v}_{QT}$ 对 h_{a_i} 的关注权重 T_{weight_i}.再次通过与式(3)相同的方法,利用 T_{weight_i} 对LSTM每一时刻隐藏单元的输出 h_{a_i} 进行加权更新,即可得到最终每一时刻的输出.随着模型的迭代训练,即可获得问题类型对应的语义信息,进而强化候选答案中与问题类型有关部分的权重.

3.3.2 基于问题中心词的关键信息注意力

当候选答案中存在多个与问题类型相关的部分时,仅采用问题类型进行信息增强很难进行区分,例如当问题类型为时间疑问句时,候选答案中有多个表达时间信息的部分,问题类型对于候选答案的注意力将会分散到多个与时间相关的部分上;当问题类型为原因疑问句或判断疑问句时,答案往往是一段话,只利用问题类型无法很好地加强候选答案对问题关键信息的捕获能力.针对上述问题,本文通过引入问题中心词的概念,以此来加大候选答案文本中与问题主题相关的词语所占的权重,同时减小不相关的词语所占的权重.

本文将问句中能够反映句子主要信息的名词或动词作为问题的中心词[21-22].例如问句"Does life insutance require a credit check?",它所表达的信息主要由"require"、" life insurance"和"credit check"体现;问句"When do an auto insurance premium go up?",它所表达的信息则主要由"go up"、" auto insurance premium"体现.

对于问题的中心词,利用依存句法分析来获取,如问句"How do I apply for Medicare in Texas?",通过依存句法分析,可得到如图3所示的结果.

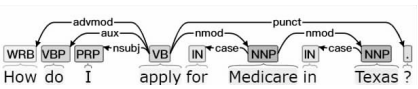

图3 问句的依存句法分析

其中,"apply"为主要动词,则提取其作为问题的主要动词 $word_V$.如果 $word_V$ 的主语或宾语为名词或名词短语,提取主语和宾语作为主要名词 $word_N$,再依次提取主要名词的修辞成分和并列成分中的名词添加到主要名词 $word_N$ 中,主要动词和名词构成问题的中心词,下文用head表示.另外,若句法分析无法提取其主要动词,则直接通过词性,过滤停用词后提取其中心词.因此,$word_N$ 的个数可能为多个.如在图3中,"apply"的主语为"I"、宾语为"Medicare",因为主语"I"为人称代词,不是名词或名

词短语，故不将其作为主要名词，而宾语“Medicare”为名词，故将其作为主要名词，同时“Texas”又作为名词修饰“Medicare”，因此，“Texas”也作为主要名词. 所以，图 3 问句中的中心词集合为{apply，Medicare，Texas}，其中，中心动词为 apply，中心名词为{Medicare，Texas}.

在得到问题的中心词后，将中心词对应的词向量集合的向量表示作为中心词的注意力向量 $\boldsymbol{v}_{QW}=(\boldsymbol{s}_1,\boldsymbol{s}_2,\cdots,\boldsymbol{s}_l)$，其中，$l$ 为问句中心词的个数，采用 $\boldsymbol{v}_{QW}$ 对候选答案正向 LSTM 的输出 $\overrightarrow{h_t}$ 和反向 LSTM 的输出 $\overleftarrow{h_t}$ 拼接后的输出 h_t 进行加权更新. 具体来说，将集合 $\boldsymbol{v}_{QW}$ 中的每个词向量分别和 h_t 进行相似度计算，然后将其中的最大值作为问题中心词的注意力向量在 h_t 上的权重表示 v_t，计算方法如式(4)所示.

$$v_t=\max\{\text{cossin}(h_t,\boldsymbol{s}_j)\}\{1\leq j\leq l\} \tag{4}$$

利用 v_t，采用类似于式(3)的方式对 h_t 进行加权更新，得到 t 时刻 h_t 的表示 h_t'. 依次采用同样的方式对候选答案每一时刻的表示进行加权更新，即得到基于问题中心词注意力的信息增强表示.

3.4　融合语义信息与问题关键信息的多阶段注意力答案选取模型

为了充分利用问题的语义信息和关键信息对候选答案进行信息增强，本文构建了融合语义信息与问题关键信息的多阶段注意力答案选取模型，具体来说，主要利用问题的相关信息，采用注意力机制，分为两个阶段对候选答案进行信息增强. 例如问题“How do I apply for Medicare in Texas?”，其候选答案集合为$\{\boldsymbol{A}_1,\boldsymbol{A}_2,\cdots,\boldsymbol{A}_n\}$. 首先使用式(1)计算得到问题的语义表示 $\boldsymbol{r}_Q$. 其次开始抽取问题的关键信息，该问题以“How”开头，因此问题类型为原因类型，同时提取问句的中心词集合{apply，Medicare，Texas}. 其中，问题类型注意力 $\boldsymbol{v}_{QT}$ 为模型初始化时为每种类型随机设定的向量，问题中心词注意力集合为 $\boldsymbol{v}_{QW}=\{\boldsymbol{s}_1,\boldsymbol{s}_2,\boldsymbol{s}_3\}$，$\boldsymbol{s}_1$、$\boldsymbol{s}_2$、$\boldsymbol{s}_3$ 分别为“apply”、“Medicare”、“Texas”对应的语义向量. 采用 3.3 节所述的方法，利用注意力机制对候选答案的语义表示进行问题关键信息增强，构建候选答案针对当前问题关键信息的语义表示 $r_{A_i}(i=1,2,\cdots,n)$，并与问题的语义表示 $\boldsymbol{r}_Q$ 进行相关度计算，依据相关度排序后，选出前 k 项$\{\boldsymbol{A}_{i_1},\boldsymbol{A}_{i_2},\cdots,\boldsymbol{A}_{i_k}\}$作为当前候选答案集合. 最后，将问题的语义信息 $\boldsymbol{r}_Q$ 作为注意力向量，采用 3.2 节所述方法，再次利用注意力机制对筛选出的候选答案集合$\{\boldsymbol{A}_{i_1},\boldsymbol{A}_{i_2},\cdots,\boldsymbol{A}_{i_k}\}$进行语义信息增强，构建当前候选答案针对问题语义信息的语义表示 $r'_{A_i}(i=i_1,i_2,\cdots,i_k)$，与问题的语义表示 $\boldsymbol{r}_Q$ 进行相关度计算，依据相关度排序后，即得到最优的候选答案$\{\boldsymbol{A}_{\text{best}}\mid \text{best}\in(i_1,i_2,\cdots,i_k)\}$. 具体的模型框架图如图 4 所示.

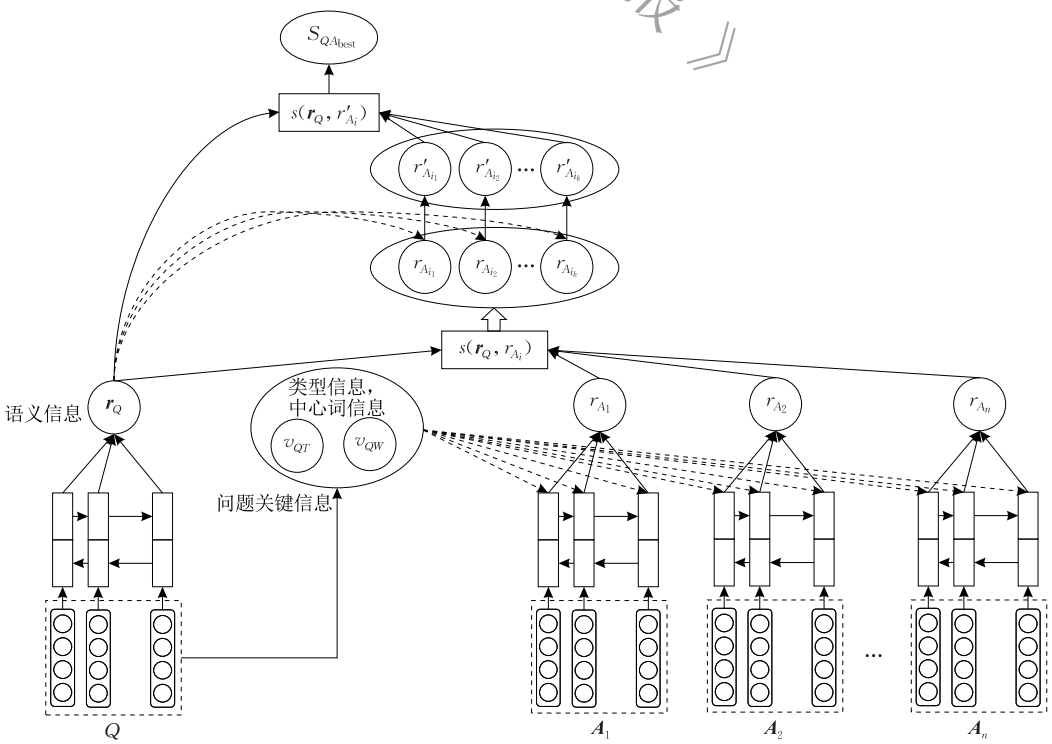

图 4　融合语义信息与问题关键信息的多阶段注意力答案选取模型

在对问题的语义表示和候选答案的语义表示进行相关度计算时，采用如式(2)余弦相似度的方法计算两者之间的相关度 S_{QA}.

答案选取模型期望达到的效果是：当模型的输入为问题的最佳答案时，S_{QA} 应该尽可能大；当模型输入为问题的非最佳答案时，S_{QA} 应该尽可能小. 因此，在对模型训练过程中，每一轮同时输入问题 $\boldsymbol{Q}$、最佳答案 $\boldsymbol{A}^+$ 和非最佳答案 $\boldsymbol{A}^-$，然后分别计算问题与最佳答案和非最佳答案的相关度 S_{QA^+} 和 S_{QA^-}，再采用式(5)所示的 Hinge Loss 函数作为损失函数对模型进行训练.

$$loss=\max\{0,mar-(S_{QA^+}-S_{QA^-})\} \tag{5}$$

其中，当 $S_{QA^+}-S_{QA^-}\geq mar$ 时，说明模型能够很好地区分最佳答案和非最佳答案，当 $S_{QA^+}-S_{QA^-}<mar$ 时，此时模型不能很好地区分正确答案与错误答案，需要调整模型参数进行迭代计算. mar 具体的取值将在 4.3 节实验参数设置部分进行说明.

对于非最佳答案 $\boldsymbol{A}^-$ 的选取，为了提升模型的学习能力，在训练的过程中，选取全部问题的候选答案中除最佳答案 $\boldsymbol{A}^+$ 之外的最佳答案作为 $\boldsymbol{A}^-$ 的值，具体如式(6)所示.

$$\boldsymbol{A}^-=\arg\max(S_{QA_i})(\boldsymbol{A}_i\neq\boldsymbol{A}^+,0\leq i\leq z_{sum}) \tag{6}$$

其中，z_{sum} 为训练数据集合中所有问题候选答案的总数.

4 实验与分析

4.1 实验数据集

为了验证本文提出模型的有效性，本文选择在 InsuranceQA 数据集、TREC-QA 数据集和 Wiki-QA 数据集上设计实验并分析，以验证本文模型的有效性.

4.1.1 InsuranceQA 数据集

InsuranceQA 数据集是一个来自保险领域的专业数据集，由 Feng 等人[6]构建，数据集中的所有问题都是来自现实世界真实用户的提问，问题的答案一般比较长. 数据集共包括四部分，分别为训练集、验证集、测试集 1、测试集 2，共有 17 487 个问题和 24 981 个答案，数据集的详细数量信息如表 3 所示，其中，Q-A 为问题的平均长度，A-A 为答案的平均长度. InsuranceQA 数据集的评价指标采用最佳答案的准确率 $P@1$ 进行评价.

表 3 InsuranceQA 问题与答案数量分布

	训练集	验证集	测试集 1	测试集 2
问题	12887	1000	1800	1800
答案	18540	1454	2616	2593
Q-A	7.15	7.16	7.16	7.17
A-A	95.61	95.54	95.54	95.54

除此之外，本文还对数据集的问题类型分布进行统计，统计结果如图 5 所示. 从图中可以看出，在训练集、验证集、测试集 1 和测试集 2 中各类问题的问题类型分布基本一致，其中占比最高的为其他问句，实物问句的所占比例也明显较高，占比最少的为地点问句.

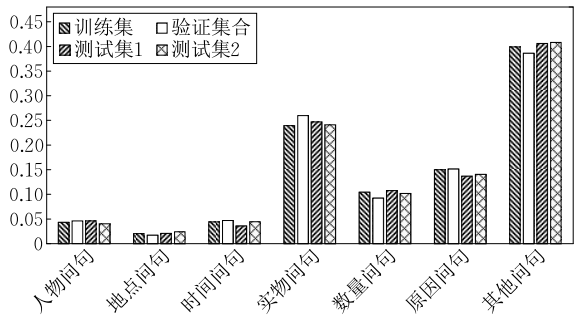

图 5 InsuranceQA 问题类型分布情况

4.1.2 TREC-QA 数据集

TREC-QA 数据集起源于国际文本检索会议(TREC)的问答任务，任务面向开放领域，且多为基于事实的小文本片段. 该数据集的训练集 TRAIN 为原始标注数据，每年发布一版，Wang 等人[23]清理所有的训练集后，得到了 TRAIN-ALL 训练集，达到了较高的数据质量，后来学者对验证集与测试集也进行了清理，得到了 CLEAN-DEV 与 CLEAN-TEST，在本文实验中，采用 TRAIN-ALL、CLEAN-DEV 与 CLEAN-TEST 进行模型的训练和验证. 数据集的具体信息如表 4 所示，其中 Question 为问题个数，Pairs 为问题-答案对的个数，Q-A 为问题的平均长度，A-A 为答案的平均长度.

表 4 TREC-QA 问题与答案数量分布

	Question	Pairs	Q-A	A-A
TRAIN	94	4718	11.3	24.6
TRAIN-ALL	1229	53417	8.3	27.7
CLEAN-DEV	65	1117	8.0	24.9
CLEAN-TEST	68	1442	8.6	25.6

同样，本文还对该数据集的问题类型分布进行统计分析，统计结果如图 6 所示，从图中可以发现在 TRAIN-ALL、CLEAN-DEV 和 CLEAN-TEST 中，

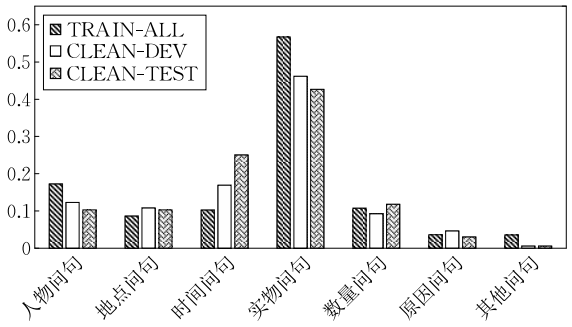

图 6　TREC-QA 问题类型分布情况

各类问题的问题类型分布基本一致，其中占比最高的为实物问句，占比最少的为其他问句.

在该数据集中，一个问题通常对应多个正确答案，需要尽可能将正确答案排名靠前. 因此，该数据集的性能评价指标采用 *MAP* 与 *MRR*，其中 *MAP* 表示所有正确答案的平均得分，如式(7)所示.

$$MAP=\frac{1}{N_{Ques}}\sum_{q\in Ques}ave(P(q)) \tag{7}$$

其中，*Ques* 表示问题集合，N_{Ques} 表示问题的总数，$P(q)$表示正确答案排序位置的得分，$ave(P(q))$表示该问题对应所有正确答案排序位置的平均得分，*MAP* 得分越高，则全部正确答案的排名越靠前.

MRR 表示问题对应的第一个正确答案的平均得分，其计算公式如式(8)所示.

$$MRR=\frac{1}{N_{Ques}}\sum_{q\in Ques}\frac{1}{rank_q} \tag{8}$$

其中，*Ques* 表示问题集合，N_{Ques} 表示问题的总数，$rank_q$表示第一个正确答案的排名，*MRR* 得分越高，则第一个结果越可能为正确答案.

4.1.3　Wiki-QA 数据集

Wiki-QA 是一个开放域问题回答的数据集，采用 Bing 查询日志作为问题源，每个问题都链接到一个可能有答案的维基百科页面，采用维基百科页面的摘要作为候选答案，然后采用众包的方式进行数据标注. 数据集的具体信息如表 5 所示，其中 Question 为问题个数，Answer 为答案个数，Q-A 为问题的平均长度，A-A 为答案的平均长度. Wiki-QA 数据集也是一个问题对应多个正确答案，因此同样采用 *MAP* 与 *MRR* 作为性能评价指标.

表 5　Wiki-QA 问题与答案数量分布

	Question	Answer	Q-A	A-A
Train	873	18821	6.36	25.51
Dev	126	1119	6.72	24.59
Test	243	2309	6.42	25.33

同样，本文还对数据集的问题的类型分布进行统计，统计结果如图 7 所示，从图中可以发现在 Train、Dev 和 Test 中的问题类型分布基本一致，其中占比最高为实物问句，占比最少为原因问句.

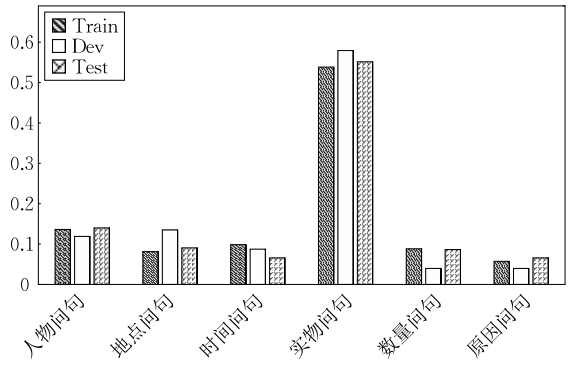

图 7　Wiki-QA 问题类型分布情况

4.2　实验对比模型

本文的主要对比模型如下：

Bag-of-Word[6]. 该模型采用问题和候选答案词语的 IDF 权重对词语的词向量进行加权求和，构建问题和候选答案的特征向量表示，采用余弦相似度计算问题和答案特征向量的相似度. 该模型是采用传统方式进行答案选择的代表模型.

Attention based Bi-LSTM[8]. 该模型使用 Bi-LSTM 对问题和候选答案进行语义编码，将问题的语义作为注意力对候选答案的编码进行更新，最后使用余弦相似度进行相似度计算. 该模型是较早将 Attention 机制引入到答案选择的方法.

IARNN-Gate[24]. 该模型将注意力信息加入到 GRU 的每个门函数中，构建了基于 RNN 的门控注意力单元，以此来构建问题和候选答案的特征向量表示，采用 GESD 进行相似度计算.

Multihop-Sequential-LSTM[25]. 该模型采用动态记忆网络(DMNS)对问题和答案进行建模，采用了多种注意力机制，进行迭代的注意力操作，构建问题和候选答案的特征向量表示，采用余弦相似度进行相似度计算.

Transformer with Hard Negatives[26]. 该模型采用 Transformer 对问题和答案进行建模，并利用 Hard Negatives 的方式选取负例样本，采用余弦相似度进行相似度计算.

BERT-Attention[27]. 该模型采用 BERT 模型对问题和答案进行建模，并构建了基于问题语义的注意力机制，采用余弦相似度进行相似度计算.

HAS[27]. 该模型的架构与 BERT-Attention 类似，但是采用了 Hashing 机制对候选答案的编码进行存储，避免实时在线计算，有效降低了计算时间与

计算资源.

Multi-Cast Attention Networks[28]. 该模型采用多种 Attention 和 Pooling 机制对问题和候选答案进行编码和交互,采用分类方法判断候选答案是否为最佳答案.

Question Classification-Deep Learning[29]. 该模型融合问题分类、实体识别、实体强化和深度学习的方法对问题和候选答案进行编码和交互,实现最佳答案的选择.

RE2[30]. 该模型主要研究序列间对齐的关键特性的选取,构建了原始点对齐特性、先前对齐特性和上下文特性,对问题和候选答案编码和交互,实现最佳答案的选择.

Comp-Clip+LM+LC[31]. 该模型通过潜在聚类的方式挖掘文本中的附加信息,实现文本中的信息聚合,从而增强对问题和答案的编码效果,实现最佳答案的选取.

4.3 实验参数设置

本文采用深度学习框架 PyTorch 对相关模型进行编码实现,并在 Ubuntu16.04 系统上采用 GPU(Tesla P100)进行模型的训练和调试. 在实验过程中,采用词向量的维度大小设置为 300,对于模型中各个参数的设置,本文采用 Hyperopt 库进行分布式参数调节,获取模型的最优参数集合,具体的选取结果为隐藏层的维度为 300,mini-batch 的大小设置为 16,优化函数采用 Adam,学习率 lr 设置为 0.001.

针对损失函数中 mar 值的选取,在各个数据集的验证集上性能随其取值变化如图 8 所示.

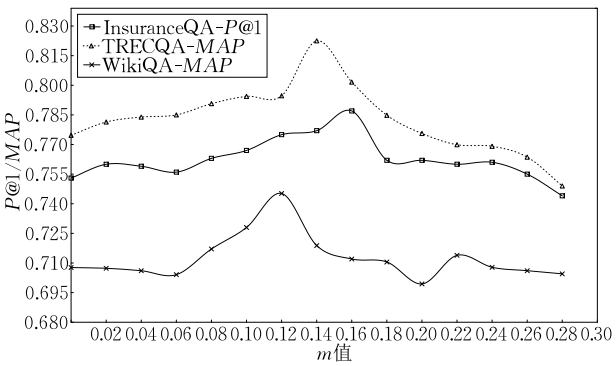

图 8 性能随 mar 值变化趋势图

我们发现 mar 取值过小和过大都会对模型在对候选答案的正负例的判断能力产生影响,进而影响最终候选答案的选取能力,最终我们选取验证集上性能最好的 mar 值作为最终的取值,具体来说,在 InsuranceQA 数据集、TREC-QA 数据集和 Wiki-QA 数据集上的取值分别为 0.18、0.16 和 0.14.

由于本文采用的是多阶段的模型,第一阶段的选择个数 k 对于模型的性能有着一定的影响,在三个数据集的验证集上,性能随 k 值的变化趋势如图 9 所示.

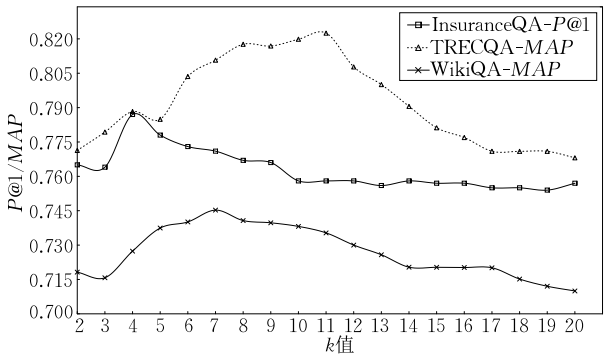

图 9 性能随 k 值变化趋势图

我们可以发现,在不同的数据集中,随着平均正确答案个数的增加,k 的最佳取值明显增大,其中 InsuranceQA 验证集的正确答案个数为 1 个,Wiki-QA 验证集上平均正确答案个数为 2.03,TREC-QA 验证集上平均正确答案个数为 3.153. 同样选取验证集上性能最好的 k 值作为最终的取值,具体来说,在 InsuranceQA 数据集、Wiki-QA 数据集和 TREC-QA 数据集上的取值分别为 4、7 和 11.

针对中心词的抽取策略,本文探究了否定副词、方位介词的抽取对性能的影响. 数据集中否定副词及方位介词的分布如表 6 所示. 其中,PP 代表问题中存在方位介词的句子个数,NA 表示问题中存在否定副词的句子个数,Question 表示问题的总个数. 从表中可以看出,否定副词在问题中出现的次数普遍较低,方位介词占比则较多. 以 TREC-QA 数据集为例做了实验对比,实验效果如表 7 所示,其中,SAAS with KI(head)表示单独添加问题中心词注意力模型,SAAS with KI(head)+PP+NA 表示在上述模型基础上增加了方位介词与否定副词的抽取.

表 6 方位介词与否定副词的数量分布

	PP	NA	Question
InsuranceQA	2260	8	17487
TREC-QA	326	4	1362
Wiki-QA	260	2	1242

表 7　方位介词与否定副词的实验对比结果

model	*ACC*	*MAP*	*MRR*
SAAS with KI(head)	79.41	76.24	86.09
SAAS with KI(head)＋PP＋NA	79.41	76.31	85.63

由实验可知，中心词选取过程中增加否定副词及方位介词对实验性能的影响甚微. 因此，本文将问句中能够反映句子主要信息的名词或动词作为问题的中心词.

4.4　实验结果及分析

按照相关数据集的实验流程和评测指标，本文分别对 InsuranceQA 数据集、TREC-QA 数据集和 Wiki-QA 数据集进行了实验分析，具体实验结果如表 8、表 9 和表 10 所示，由于本文实验的数据集划分和实验流程完全按照各个数据的规范进行实验，因此，表中显示的实验结果均来自于相关论文中报告的结果.

表 8　InsuranceQA 数据集实验对比结果

model	Dev	Test1	Test2
Bag-of-Word[6]	31.90	32.10	32.20
Attention based Bi-LSTM[8]	68.90	69.00	64.80
IARNN-Gate[24]	70.00	70.10	62.80
Multihop-Sequential-LSTM[25]	—	70.50	66.90
Transformer with Hard Negatives[26]	75.70	75.60	73.40
BERT-Attention[27]	—	76.12	74.12
HAS[27]	—	76.38	73.71
SAAS with KISI	76.00	75.28	73.83
MSAAS with KI-SI(type)	78.60	78.06**	74.56*
MSAAS with KI-SI(head)	78.30	**78.33****	**75.06****
MSAAS with KI-SI(head＋type)	**78.70**	77.78**	74.72*

表 9　TREC-QA 数据集实验对比结果

Model	*MAP*	*MRR*
Attention based Bi-LSTM[8]	75.30	83.00
IARNN-Gate[24]	73.70	82.10
Multihop-Sequential-LSTM[25]	81.30	89.30
Multi-Cast Attention Networks[28]	83.80	90.39
Question Classification-Deep Learning[29]	**86.47**	90.39
SAAS with KISI	76.31	88.22
MSAAS with KI-SI(type)	82.78	91.29**
MSAAS with KI-SI(head)	80.89	90.61*
MSAAS with KI-SI(head＋type)	81.97	**91.58****

表 10　Wiki-QA 数据集实验对比结果

Model	*MAP*	*MRR*
IARNN-Gate[22]	72.60	74.00
Multihop-Sequential-LSTM[25]	72.20	73.80
RE2[30]	74.52	76.18
Comp-Clip＋LM＋LC[31]	76.40	78.40
BERT-Attention[27]	80.65	81.83
HAS[27]	81.01	82.22
SAAS with KISI	72.57	73.91
MSAAS with KI-SI(type)	**76.90****	**78.51****
MSAAS with KI-SI(head)	74.40*	75.73
MSAAS with KI-SI(head＋type)	75.49**	77.00**

其中，SAAS with KISI 模型表示将问题的语义信息注意力和问题的关键信息注意力都添加在模型的第一阶段，构建候选答案的三个语义表示，然后对三个语义表示结果进行融合，构建候选答案的语义表示与问题的语义表示，并将二者进行交互选出最佳答案；MSAAS with KI-SI 表示本文的融合问题关键信息和问题语义信息的多阶段注意力答案选取模型，其中，MSAAS with KI-SI(type)表示第一阶段只采用问题类型作为问题关键信息进行信息增强，MSAAS with KI-SI(head)表示第一阶段只采用问题中心词作为问题关键信息进行信息增强，MSAAS with KI-SI(head＋type)表示第一阶段同时采用问题类型和问题中心词作为问题关键信息进行信息增强.

在表 8、表 9 和表 10 中，“*”表示显著性水平 $p<0.05$，“**”表示显著性水平 $p<0.01$，本文显著性验证参考文献[32]中的方法，在测试集上，采用 1000 次有放回的抽样进行评估. 具体来说，InsuranceQA 数据集是针对 HAS 模型进行显著性检验的，TREC-QA 数据集是针对 Question Classification-Deep Learning 模型进行了显著性检验的，Wiki-QA 数据集是针对 Comp-Clip＋LM＋LC 模型进行了显著性检验的. 从表可以看出，相较于表中的对比模型，本文模型的多项指标都有显著性提高.

根据表 8、表 9 和表 10 中的结果，从对问题和答案的编码方式来看，可以发现 Bag-of-Word 模型远不如采用深度学习的编码方式，这是由于 Bag-of-Word 模型单纯地从词的角度分析，未考虑文本的内容特征和其他关联特征. 从注意力的增加来看，添加了注意力机制的模型效果要明显优于不添加注意力机制模型的效果，这是由于注意力机制加强了问题和答案的交互能力；从注意力机制的添加方式来看，采用 self-attention 或者 multi-head self-attention 的模型(Multihop-Sequential-LSTM、Transformer with Hard Negatives 模型)效果也要优于其他注意力添加方式；另外，基于 BERT 的模型(BERT-Attention、HAS)，相较于以往的模型，取得了最佳的效果.

在 InsuranceQA 数据集中，相比于以往单个维度注意力的添加，本文 MSAAS with SIKI 模型分阶段地融合了语义信息和问题关键信息两个维度的注意力，取得了最好的效果，证明了本文模型的有效性. 具体来说，除了基于 BERT 的模型，本文的

SAAS with KISI 模型就表现出了明显的优势，说明本文问题的语义信息和问题关键信息的信息增强是有效果的；在进一步将问题关键信息和问题语义信息分阶段地进行信息增强以后，MSAAS with KI-SI 模型的性能也超过了基于 BERT 的模型，表现出了最优的性能，说明了本文构建的分阶段的信息增强方式是有效的.

在 TREC-QA 数据集中，本文提出的 MSAAS with KI-SI(head＋type)模型在 *MRR* 指标上取得的结果明显好于其他模型，在 *MAP* 指标上虽然没有达到最优，但也维持在比较高的性能. 同时，在添加多阶段的注意力机制以后，本文 MSAAS with KI-SI 模型的性能都是有所提升的，也说明了本文多阶段模型的有效性. 对于 *MAP* 指标稍微偏低的原因可能是由于在 TREC-QA 数据集中，有少量问题的正确答案个数比较多(在训练集、验证集和测试集上，最多的一个问题的正确答案个数分别为 51、17 和 14，平均个数为 5.218、3.153 和 3.647)，本文模型在进行分阶段筛选时，若正确答案的个数超过了筛选的数量，将有部分正确答案不能筛选到，则在计算 *MAP* 指标时作为较低的得分处理，从而导致 *MAP* 指标中的正确答案的平均得分普遍偏低.

在 Wiki-QA 数据集中，本文的 MSAAS with KI-SI(head＋type)模型的性能虽不如基于 BERT 的 BAER-Attention 和 HAS 模型，但是也明显好于其他模型，也说明了本文模型的有效性. 对于本文模型性能不如文献[27]的两个模型的性能，我们通过分析发现，由于 Wiki-QA 数据集的问句采用 Bing 的搜索日志构建，相较于 InsuranceQA 数据集和 TREC-QA 数据集而言，显得更加的随意，其句法结构和语义结构也不够完善，由于 BERT 模型采用了大规模语料进行预训练，对于非正式语言的编码能力要比本文模型强，导致本文模型对于问题的编码不如基于 BERT 的模型效果好，从而在最终结果上要稍差一些.

4.4.1 问题语义和关键信息注意力性能分析

为了验证本文模型中问题语义注意力和问题关键信息注意力的引入对模型的性能的影响，本文在三个数据集上分别设置了六组对照实验，分别是 3.1 节叙述的基础模型(AS)、只采用问题关键信息对候选答案进行第一阶段注意力增强选出最佳答案的模型(SAAS with KI)、只采用问题语义信息对候选答案进行第一阶段注意力增强选出最佳答案的模型(SAAS with SI)和 MSAAS with KI-SI 模型. 其中，SAAS with KI 模型包括 SAAS with KI(t)、SAAS with KI(h)和 SAAS with KI(t&h)，分别对应问题的关键信息单独采用问题类型、单独使用问题中心词以及同时采用问题类型和中心词进行关键信息增强的模型. 具体的实验结果如图 10、图 11 和图 12 所示.

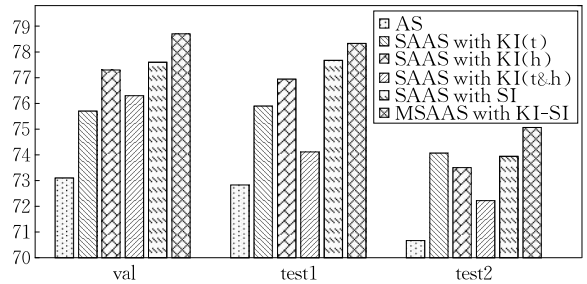

图 10 InsuranceQA 问题语义和关键信息注意力性能对比图

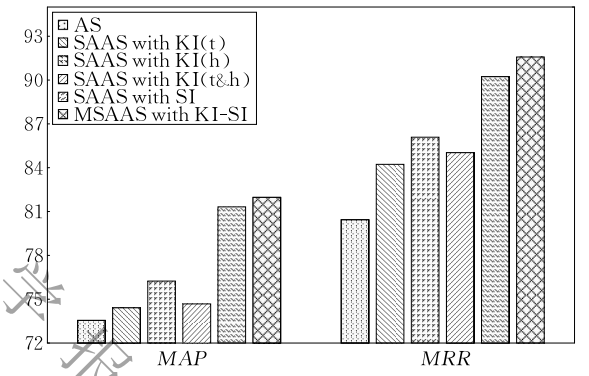

图 11 TREC-QA 问题语义和关键信息注意力性能对比图

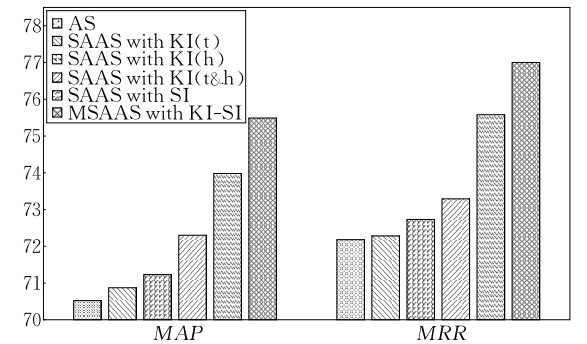

图 12 Wiki-QA 问题语义和关键信息注意力性能对比图

从图 10、图 11 和图 12 可以看出，对于三个数据集，在基础模型上单独添加问题语义信息和问题关键信息的注意力对候选答案进行信息增强，相较于基础模型都有不同程度的性能提升，问题语义信息的注意力信息增强性能提升的程度要大问题关键信息；针对问题关键信息，添加问题中心词注意力对性能的提升优于问题类型注意力；同时，在第一阶段添加问题关键信息的基础上，在第二阶段再次添加

问题语义信息，性能也有一定程度的提升.这说明本文所构建的问题语义信息和问题关键信息均对模型性能的提升是有帮助的.

另外，单独对比问题类型、问题中心词、问题语义三种注意力对模型性能的影响（见模型 SAAS with KI(t)，SAAS with KI(h)，SAAS with SI 的效果），可以发现，单独添加问题语义注意力对模型效果的提升最为明显，可能是问题的语义信息在一定程度上也包含了问题类型信息和问题的中心词信息以及一些其他信息，这也是在我们多阶段的模型中将语义信息添加在第二阶段的原因之一.

4.4.2　问题语义和关键信息注意力可视化分析

为了更清楚地说明本文问题语义信息和关键信息对于模型性能的影响，我们从数据集中选取了一些问题和其候选答案，输出了其各个词语在各个阶段的权重表示，并进行了可视化分析，如在 InsuranceQA 数据集中，对于问题："When be the first Life Insurance policy issue?"，首先进行第一阶段信息增强.该问题为时间类型的问句，抽取出的中心词集合为{first，Life Insurance policy，issue}，其最佳答案与排名第一和任一其他的非最佳答案的语义表示，在经过问题关键信息注意力增强后的语义表示可视化为图 13、图 14 和图 15.

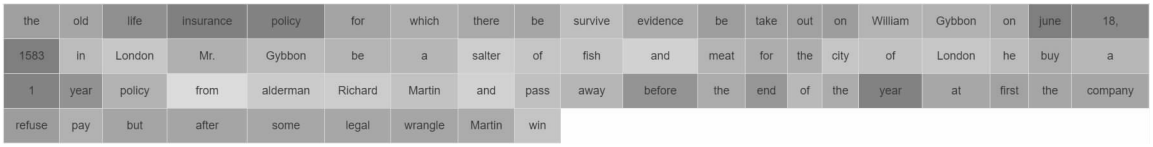

图 13　最佳候选答案添加问题关键信息的语义表示可视化

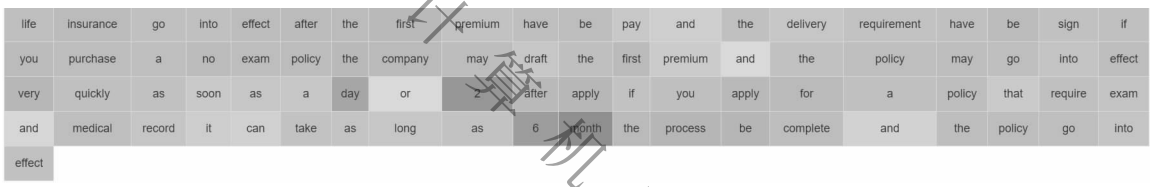

图 14　排名第一的非最佳候选答案添加问题关键信息的语义表示可视化

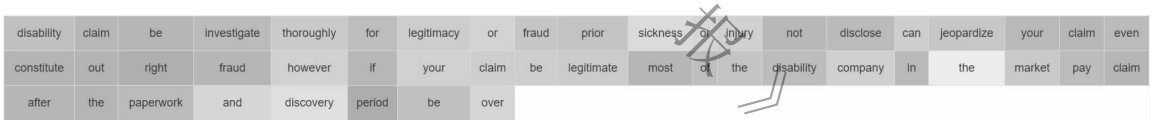

图 15　其他非最佳候选答案添加问题关键信息的语义表示可视化

其中，最佳答案、排名第一的非最佳答案和任一其他非最佳答案与问题的相似度得分分别为 0.4253、0.3083 和 −0.2554.我们可以发现，最佳答案直接对问题所对应的产生时间及背景进行了阐述；而排名第一的非最佳答案虽然提到了时间信息，但在语义方面，讲述的是保险生效时间，与问题语义不符.

同时我们还可以发现，对于最佳候选答案，在其语义表示中，"June 18，1583"、"1 year"、"end"、"before"等与时间相关的词语和"life"、"insurance"、"policy"等与问题中心词语相关的词语的权重要明显高于其他词语的权重；而对于非最佳答案，其权重分布相对比较分散，说明了添加问题的关键信息，可以让候选答案中与问题关键信息相关的词语权重加大，更容易捕获候选答案中的关键信息，从而建立候选答案与问题的联系，证明了本文问题关键信息注意力的有效性.

接着，在第一阶段关键信息增强的基础上进行第二阶段的信息增强，将添加了问题语义信息的结果进行可视化，其结果如图 16、图 17 和图 18 所示.

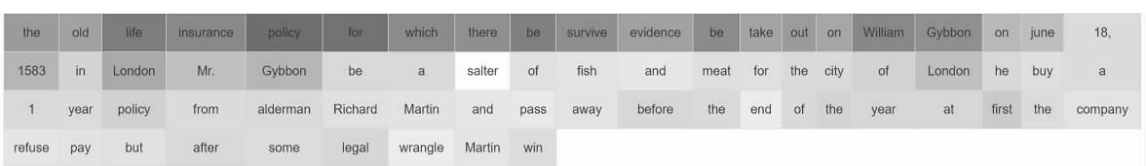

图 16　最佳候选答案添加问题语义信息的语义表示可视化

life	insurance	go	into	effect	after	the	first	premium	have	be	pay	and	the	delivery	requirement	have	be	sign	if
you	purchase	a	no	exam	policy	the	company	may	draft	the	first	premium	and	the	policy	may	go	into	effect
very	quickly	as	soon	as	a	day	or	2	after	apply	if	you	apply	for	a	policy	that	require	exam
and	medical	record	it	can	teke	as	long	as	6	months	the	process	be	complete	and	the	policy	go	into
effect																			

图 17 排名第一的非最佳候选答案添加问题语义信息的语义表示可视化

disability	claim	be	investigate	thoroughly	for	legitimacy	or	fraud	prior	sickness	or	injury	not	disclose	can	jeopardize	your	claim	even
constitute	out	right	fraud	however	if	your	claim	be	legitimate	most	of	the	disability	company	in	the	market	pay	claim
after	the	paperwork	and	discovery	period	be	over												

图 18 其他非最佳候选答案添加问题语义信息的语义表示可视化

其中，最佳答案、排名第一的非最佳答案和任一其他非最佳答案与问题的相似度得分分别为 0.4613、0.2314 和 −0.0097. 对于最佳候选答案，在其语义表示中，与问题语义相关的词语或者句子的权重要明显高于其他词语的权重，如首句“the old life insurance policy for …”；对于排名第一的非最佳答案，其主要权重也集中在与问题语义相关的开头，如“life insurance go into effect after the first …”；而对于图 18 中的非最佳答案，其权重的分布相对比较分散，虽然也有一些词语权重较高，但是也都不是非常明显，且与问题的语义关联性不是太高. 进一步证明了本文问题语义信息注意力的有效性.

4.4.3 多阶段注意力引入性能分析

为了验证模型将问题语义注意力和关键信息注意力分多个阶段引入对模型性能的影响，本文在三个数据集上设置了六组对照实验，分别是 3.1 节叙述的基础模型(AS)、SAAS with KISI、第一、二阶段分别采用问题语义信息和问题关键信息进行注意力增强选出最佳答案的模型(MSAAS with SI-KI)和 MSAAS with KI-SI 模型. 其中，MSAAS with SI-KI 模型同样包括 MSAAS with SI-KI(t)、MSAAS with SI-KI(h)和 MSAAS with SI-KI(t&h)，分别对应问题的关键信息采用问题类型、问题中心词、同时采用问题类型和中心词. 实验结果如图 19、图 20 和图 21 所示.

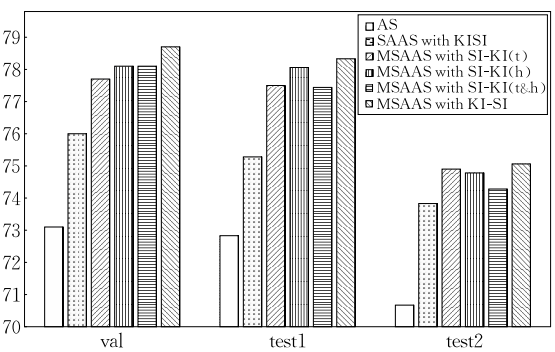

图 19 InsuranceQA 多阶段注意力引入性能对比图

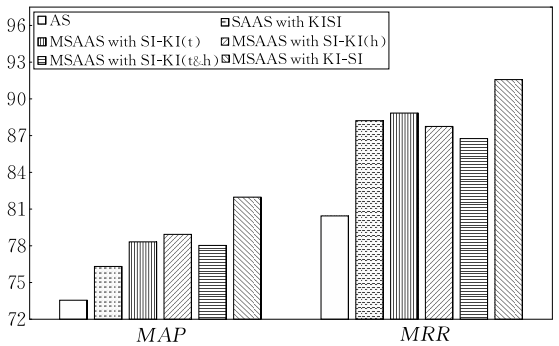

图 20 TREC-QA 多阶段注意力引入性能对比图

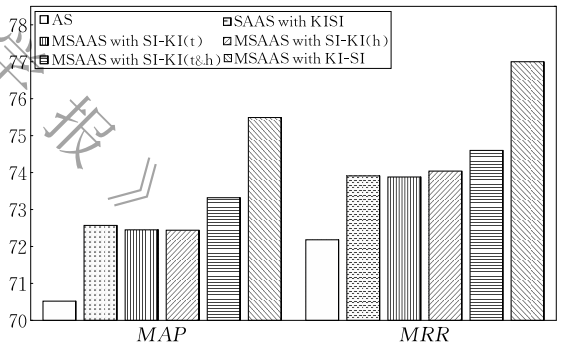

图 21 Wiki-QA 多阶段注意力引入性能对比图

从图 19、图 20 和图 21 可以看出，在三个数据集上相比于在同一阶段加入多种注意力(SAAS with SIKI)以及交换问题语义注意力和问题关键信息注意力的添加顺序(MSAAS with SI-KI)，本文的 MSAAS with KI-SI 模型性能均达到了最优效果，说明了本文提出的分阶段注意力的方法的有效性.

本文的多阶段注意力机制跟人在做答案选择任务时的思维方式是相似的，当人在做答案选取任务时，一般来说会首先阅读问题，然后以问题中的一些关键信息对候选答案进行初步地筛选，接着，以问题中的详细信息与初步筛选出来的答案进行进一步地对比，从而选出最佳答案. 人类以关键信息进行初步筛选的过程就可以看作是 MSAAS with KI-SI 模型

第一阶段以问题关键信息进行信息增强筛选答案的过程；人类以问题中的详细信息进行进一步对比的过程就可以看作是 MSAAS with KI-SI 模型第二阶段以问题语义信息进行信息增强筛选答案的过程，因此本文的模型与人进行该任务的步骤是大致吻合.

5 总　结

本文提出了一种融合语义信息与问题关键信息的多阶段注意力答案选取模型，分阶段地将问题的语义信息和关键信息通过注意力机制的方式对候选答案的语义表示进行信息增强，加强了对候选答案中与问题相关的信息的建模能力，增强了模型对候选答案关键信息的捕获能力，从而有效提升了答案选取任务的性能；同时，在模型的训练过程中，对于负样本的选取，实时选取出最佳答案以外的最优答案作为负样本，以对模型进行优化，增强了模型的学习能力. 通过在 InsuranceQA、TREC-QA 和 Wiki-QA 数据集上的相关实验，本文的模型都表现出优越的性能，并在不使用大规模辅助语料的基础上，在多个指标中超过了已知最好的同类模型.

不过，在以上的研究过程中，我们主要集中在英文数据集上，在未来的工作中我们将尝试对中文语料进行处理，验证该模型是否具有普适性；同时，答案选取任务在具体的使用过程中与搜索引擎类似，一般需要进行实时在线计算，对模型的时间性能要求较高，在后续的工作中，我们也将进一步优化模型的执行效率；另外，本文所提模型在最优答案较少的数据集（如 InsuranceQA）上的效果要明显好于有多个答案的数据集，同时也表现在 TREC-QA 和 Wiki-QA 数据集上的 *MAP* 性能略低，这也是我们后期对答案选择模型进一步优化的研究重点. 另外，随着 ELMo、Bert、GPT 等预训练模型的兴起和迁移学习技术的发展，大规模预训练＋微调的方式正在成为一种新的思路[33-34]，因此在后续的研究中，如何利用大规模数据来提升答案选取任务的效果将是我们的重点研究方向.

参　考　文　献

[1] Zhao Yi-Ping. Comparative Study on Common and Semantic Search Engines [M. S. dissertation]. Jilin University, Changchun, 2009(in Chinese)
(赵夷平. 传统搜索引擎与语义搜索引擎比较研究[硕士学位论文]. 吉林大学，长春，2009)

[2] Heilman M, Smith N A. Tree edit models for recognizing textual entailments, paraphrases, and answers to questions //Proceedings of the Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Los Angeles, USA, 2010: 1011-1019

[3] Surdeanu M, Ciaramita M, Zaragoza H. Learning to rank answers to non-factoid questions from Web collections. Computational Linguistics, 2012, 37(2): 351-383

[4] Yih W T, Chang M W, Meek C, et al. Question answering using enhanced lexical semantic models//Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. Sofia, Bulgaria, 2013: 1744-1753

[5] Tymoshenko K, Moschitti A. Assessing the impact of syntactic and semantic structures for answer passages reranking// Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. Melbourne, Australia, 2015: 1451-1460

[6] Feng M, Xiang B, Glass M R, et al. Applying deep learning to answer selection: A study and an open task//Proceedings of the 2015 IEEE Workshop on Automatic Speech Recognition and Understanding. Scottsdale, USA, 2016: 813-820

[7] Guo J, Yue B, Xu G, et al. An enhanced convolutional neural network model for answer selection//Proceedings of the 26th International Conference on World Wide Web Companion. Perth, Australia, 2017: 789-790

[8] Tan M, Santos C D, Xiang B, et al. Improved representation learning for question answer matching//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin, Germany, 2016: 464-473

[9] Denil M, Bazzani L, Larochelle H, et al. Learning where to attend with deep architectures for image tracking. Neural Computation, 2011, 24(8): 2151-2184

[10] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473, 2014

[11] Britz D, Goldie A, Luong M T, et al. Massive exploration of neural machine translation architectures//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark, 2017: 1442-1451

[12] Tang G, Müller M, Rios A, et al. Why self-attention?: A targeted evaluation of neural machine translation architectures //Proceedings of the Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium, 2018: 4263-4272

[13] Cheng J, Dong L, Lapata M. Long Short-Term Memory-Networks for Machine Reading//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Austin, USA, 2016: 551-561

[14] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need//Proceedings of the Advances in Neural Information Processing Systems. Los Angeles, USA, 2017: 5998-6008

[15] He X, He Z, Song J, et al. NAIS: Neural attentive item similarity model for recommendation. IEEE Transactions on Knowledge and Data Engineering, 2018, 30(12): 2354-2366

[16] Yu S, Wang Y, Yang M, et al. NAIRS: A neural attentive interpretable recommendation system//Proceedings of the 12th ACM International Conference on Web Search and Data Mining. Melbourne, Australia, 2019: 790-793

[17] Bachrach Y, Zukovgregoric A, Coope S, et al. An attention mechanism for neural answer selection using a combined global and local view//Proceedings of the 2017 IEEE 29th International Conference on Tools with Artificial Intelligence. Boston, USA, 2017: 425-432

[18] Xu D, Ji J, Huang H, et al. Gated group self-attention for answer selection. arXiv preprint arXiv:1905.10720, 2019

[19] Chen D, Fisch A, Weston J, et al. Reading Wikipedia to answer open-domain questions//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver, Canada, 2017: 1870-1879

[20] Hao Y, Zhang Y, Liu K, et al. An end-to-end model for question answering over knowledge base with cross-attention combining global knowledge//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver, Canada, 2017: 221-231

[21] Huang Z, Thint M, Qin Z. Question classification using head words and their hypernyms//Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing. Honolulu, USA, 2008: 927-936

[22] Li X, Roth D. Learning question classifiers: The role of semantic information. Natural Language Engineering, 2006, 12(3): 229-249

[23] Wang M, Smith N A, Mitamura T. What is the Jeopardy model? A quasi-synchronous grammar for QA//Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Prague, Czech Republic, 2007: 22-32

[24] Wang B, Liu K, Zhao J. Inner attention based recurrent neural networks for answer selection//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin, Germany, 2016: 1288-1297

[25] Tran N K, Niedereée C. Multihop attention networks for question answer matching//Proceedings of the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval. Ann Arbor. Michigan, USA, 2018: 325-334

[26] Kumar S, Mehta K, Rasiwasia N. Improving answer selection and answer triggering using hard negatives//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Hong Kong, China, 2019: 5913-5919

[27] Xu D, Li W J. Hashing based answer selection. arXiv preprint arXiv:1905.10718, 2019

[28] Tay Y, Tuan L A, Hui S C. Multi-cast attention networks for retrieval-based question answering and response prediction //Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. London, UK, 2018: 2299-2308

[29] Madabushi H T, Lee M, Barnden J. Integrating question classification and deep learning for improved answer selection //Proceedings of the 27th International Conference on Computational Linguistics. Santa Fe, USA, 2018: 3283-3294

[30] Yang R, Zhang J, Gao X, et al. Simple and effective text matching with richer alignment features//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy, 2019: 4699-4709

[31] Yoon, S, Dernoncourt F, Kim D S, et al. A compare-aggregate model with latent clustering for answer selection// Proceedings of the 28th ACM International Conference on Information and Knowledge Management. Beijing, China, 2019: 2093-2096

[32] Koehn P. Statistical significance tests for machine translation evaluation//Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing. Barcelona, Spain, 2004: 388-395

[33] Garg S, Vu T, Moschitti A. TANDA: Transfer and adapt pre-trained transformer models for answer sentence selection. arXiv preprint arXiv:1911.04118, 2019

[34] Lai T, Tran Q H, Bui T, et al. A gated self-attention memory network for answer selection//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Hong Kong, China, 2019: 5955-5961

ZHANG Yang-Sen, Ph. D., professor. His major research interests include natural language processing and artificial intelligence.

WANG Sheng, M. S. candidate. His major research interest is natural language processing.

WEI Wen-Jie, M. S. candidate. His major research interest is natural language processing.

PENG Yuan-Yuan, M. S., engineer. Her major research interest is natural language processing.

ZHENG Jia, M. S., engineer. His major research interest is natural language processing.

Background

The problems studied in this article are very relevant to the automatic question answering system, one of the current research hotspots. In recent years, with the continuous development of artificial intelligence technology, various automatic question answering systems have come out one after another. In these systems, answer selection is a key step, which directly affects the performance of these systems.

Aiming at the problem of inaccurate capture of key information in the answer, this paper proposes a multi-stage attention answer selection model that combines semantic information and key information of the question. By combining the semantic information and the key information of the question with the semantic representation of the candidate answers in stages, the system performance is improved effectively. Aiming at the problem that the candidate answers are difficult to sort, this paper refers to the process of human thinking, proposes a strategy of answer selection by layers, which improves the corresponding evaluation indexes of answer selection such as accuracy and mean reciprocal rank.

The authors and laboratory of this article have a lot of research in the field of natural language processing. For example, they have proposed a semantic error correction model in text error correction, and a reading comprehension model in semantic understanding.

Our work is supported by the National Natural Science Foundation of China (Grant No. 61772081).