

基于深度语义的三阶段式问题检索模型^①

赵 钊, 尚爱国, 焦一凯, 朱欣娟

(西安工程大学 计算机科学学院, 西安 710600)

通信作者: 朱欣娟, E-mail: zhuxinjuan@xpu.edu.cn



摘 要: 随着检索式问答技术的日趋成熟, 如何有效利用现有的模型和检索工具, 达到问答系统的整体优化, 是亟待研究的现实问题. 提出了一种基于深度语义的三阶段式问题检索模型 (TSFR-RM), 用于构筑智能客服问答系统. 首先基于深度学习计算方法计算用户问题和知识库问题的文本表征相似度, 锁定 top-k 候选答案集, 同时赋予模型泛化检索的能力; 其次针对用户问题与知识库问题答案对, 构造多角度语义特征, 进行精确比对计算; 最后构造状态预测模型返回问题检索精准答案. 通过真实文旅机构客服问答数据集实验及实际应用效果表明, 该模型相较于其他基于特征和表征的问题检索模型, 在精确率 (precision) 性能指标上最高提升 9.3 个百分点, 提升优化了智能客服检索系统的准确性.

关键词: 智能客服系统; 问题精确检索; 多阶段任务学习; 多角度语义特征; 问题状态判断; 语义分割; 数字经济; 问答系统

引用格式: 赵钊, 尚爱国, 焦一凯, 朱欣娟. 基于深度语义的三阶段式问题检索模型. 计算机系统应用, 2023, 32(5): 244-252. <http://www.c-s-a.org.cn/1003-3254/9081.html>

Three-stage Problem Retrieval Model Based on Deep Semantics

ZHAO Zhao, SHANG Ai-Guo, JIAO Yi-Kai, ZHU Xin-Juan

(School of Computer Science, Xi'an Polytechnic University, Xi'an 710600, China)

Abstract: As the retrieval-based question and answer (Q&A) technology becomes increasingly mature, determining how to effectively use existing models and retrieval tools to achieve overall optimization of Q&A systems is a practical problem that needs to be studied. The study proposes a three-stage and a fusion of feature and representation problem retrieval model (TSFR-RM) for constructing intelligent customer service Q&A systems. Firstly, the similarity between the text representations of users' questions and questions in knowledge bases is calculated by deep learning methods to target the top-k candidate answer set and give the model the ability of generalized retrieval. Secondly, multi-angle semantic features are constructed for pairs of answers to users' questions and questions in knowledge bases to perform accurate comparison calculations. Finally, a state prediction model is built to return accurate answers to question retrieval. The experimental and practical application results show that the model improves the accuracy of the intelligent customer service retrieval system for a cultural tourism institution by up to 9.3 percentage points in the performance index of precision compared with other feature- and representation-based question retrieval models.

Key words: intelligent customer service system; accurate question retrieval; multi-stage task learning; multi-angle semantic features; problem status judgment; semantic segmentation; digital economy; question and answer (Q&A) system

① 基金项目: 国家重点研发计划 (2019YFC1521405)

收稿时间: 2022-10-21; 修改时间: 2022-11-30; 采用时间: 2022-12-11; csa 在线出版时间: 2023-02-28

CNKI 网络首发时间: 2023-03-01

随着数字经济的兴起,客户服务领域的数字化、智能化转型,已经成为我国各行业开展转型落地工作的重要切入点之一。相比传统客服,智能客服利用自然语言处理、人工智能、语音识别等技术提升了文字、语言的处理能力,在接入渠道、响应效率、数据管理分析等方面具有突出优势,极大地提高了客服的工作效率。目前自动问答技术大体分为检索式问答技术和生成式问答技术两大类^[1]。相比于生成式问答生成答案的开放性,检索式问答中机器人的回答是预先设置好的,不会产生不可控制的答案,在智能客服系统中得到广泛应用。目前检索式自动问答技术发展日趋成熟,利用现有的模型和检索工具,用户可以用自然语言作为查询的问题从大量的语料中寻找最恰当的内容作为回复,获得更符合用户需求的信息资源^[2,3]。

检索式自动问答模型的构建主要采用3种方法:基于规则模板匹配的方法、基于统计或机器学习的方法和基于神经网络的深度学习的方法^[1,4]。基于规则模板匹配的方法适用于构建垂直领域的问答模型,满足于垂直领域或者某个特定范围的用户问题,但需要人工制定规则,缺点是耗费时间且模型扩展性欠佳。基于统计或机器学习的方法又可以看作基于特征的方法。Clark等^[5]融合了基于模式匹配、基于统计与基于知识库的推理方法,首先统计了经过语料库检索后得出的语料,然后使用支持向量机评估语义相似程度,最后由简单的知识推理得到问题的最终答案。Fan等^[6]使用多维特征组合增强了社区问答系统CQA中的答案选择,首先使用基于相似度排序的传统问答方法以及问题答案的信息来寻找问题答案之间的相似性,然后使用答案的文本描述信息来判断答案的合理性,该方法提高了正确答案的整体排名。Aithal等^[7]在问答系统中引入了一个问题相似机制,该机制通过计算生成的问题与提出的问题之间的余弦相似度得到问题相似得分,之后基于阈值识别过滤掉问答过程中不可回答、无关和可回答的问题,提高了问答系统答案的准确性。基于统计或机器学习的方法可解释性较好,依赖人工寻找的特征更加精准,在检索时可以定位到问题的关键信息使得检索结果更准确,但泛化能力一般且模型的效果在很大程度上取决于特征的选择。基于神经网络的深度学习的方法又可以看作基于表征的方法。Gao等^[8]将问题和答案之间的相似性矩阵与词向量拼

接,再使用卷积神经网络提取特征,通过最大化问题答案对的匹配分数训练模型。Humeau等^[9]集成了Bi-encoder和Cross-encoder的优点并避免了缺点,提出了新的检索架构Poly-encoder,在一定程度上保证了检索质量和检索速度。Khattab等^[10]为了兼顾匹配的效率和doc中的上下文信息,提出了一种新的排序模型ColBERT,ColBERT提出了一种新颖的后期交互范式,用于估计查询query和文档doc的相关性,使用两个encoder分别对query和doc进行编码,可预先计算doc的编码过程并缓存,保证了模型推理的高效率;使用后期交互策略也使得query和doc能够进行交互。Lu等^[11]提出了一种生成高质量文本表示的预训练语言模型SEED-Encoder,使用一个弱解码器对自动编码器进行预训练,通过限制解码器的容量和注意力灵活性,来促进编码器产生更好的文本表示,实验证明该模型显著提高了密集检索模型的有效性。Gao等^[12]提出了一个对比学习框架SimCSE,它极大地改善了语义文本相似性任务中的句子嵌入,在该方法的基础上以任何随机句子作为查询来检索相似句子,与BERT检索的句子相比,SimCSE检索的句子具有更高的质量。基于神经网络的深度学习的方法泛化能力强,能够捕捉到问题与历史问题之间的语义关系,但依赖于通过模型产生的文本表征,可能导致模型自身导出的句向量质量较低,难以反映出两个句子的语义相似度。

综上,基于统计或机器学习的方法以及基于神经网络的方法在结构上都采用端到端的模型,过分依赖大量训练数据且训练任务单一,使得整个检索过程存在一定的局限性。基于统计或机器学习的方法人工参与寻找的特征更加精准且可解释性好,而基于神经网络的深度学习的方法泛化能力强,两者各有所长。智能客服系统的关键在于提高问题检索召回的精度和效率,如何合理利用现有模型和方法,达到检索系统的整体优化和性能提升,是亟待研究的问题。针对以上分析,本文提出了一种基于深度语义的三阶段式问题检索模型(three-stage and a fusion of feature and representation problem retrieval model, TSFR-RM)。该模型融合了用户问题与知识库问题的特征与表征,在结构上将问题的检索流程划分为3个阶段:首先基于深度学习方法计算问题和知识库问题的文本表征相似度,锁定top-k候选答案集,同时赋予模型泛化检索的能力;其次针对用户问题与知识库问题答案对,构造多角度

语义特征,进行精确比对计算;最后构造状态预测模型返回问题检索精准答案.整个检索流程体现了先找准范围,后精确锁定的思想,提高了系统整体检索质量.

1 基于深度语义的问题检索优化模型

基于深度语义的问题检索优化模型如图1所示,模型从整体结构上分为3部分,分别是文本语义表征比对、精确检索特征计算和候选问题比对状态判断.在文本语义表征比对阶段,对案例库CB中存放的搜集到的用户问答数据对进行整理合并形成知识库KB.计算由文本表征模型得到的用户问题query的文本表征query embedding和由相似向量检索库Faiss处理得到的知识库问题KB_Query的文本表征KB_Query embedding的相似度,得到深度语义得分Deep_Score,排序后为用户问题缩小检索范围,取前top-k生成待精确检索问答集Question-Answer.由于客服用户问题大

都为短文本问题,短文本问题独特的结构使其包含的有效信息较少,造成样本的特征稀疏且特征集的维数高^[13].因此,很难从用户问题中抽取到关键、准确、足够的样本特征用于相似问题的检索.针对短文本的这一特点,在精确检索特征计算阶段,基于第1阶段得到的待精确检索问答对,为其构造5种外部特征,分别是 $Sim(Q, KBQ)$ 、 $Sim(QE, KBQE)$ 、 $Sim(QR, KBQR)$ 、 $Sim(Q, KBA)$ 和 $Deep_Score$,全方位辅助问题检索,5种特征的详细阐述参阅第2.2节和第3节.在候选问题比对状态判断阶段,将基于待检索问题构造的5个特征送入基于随机森林的问题状态预测模型进行预测,预测结果为true代表该候选问题与待检索问题相似,其答案可以作为待检索问题的答案,false代表该候选问题与待检索问题不相似,其答案不可以作为待检索问题的答案.通过对候选问题的状态进行判断可以得到用户问题的最终唯一确定答案Query answer.

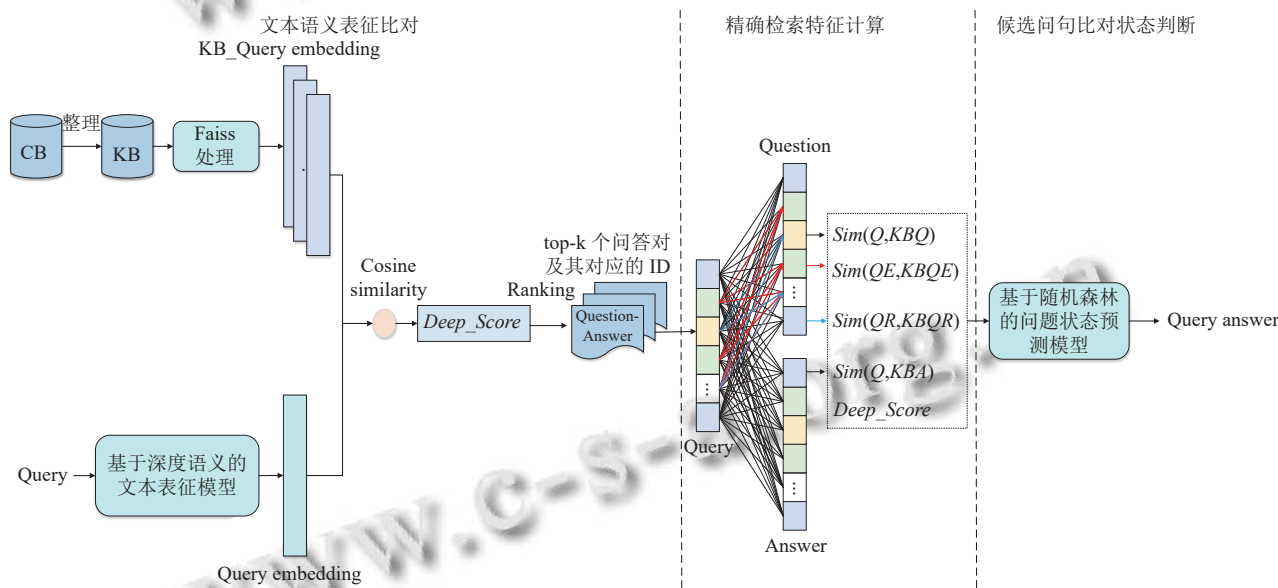


图1 基于深度语义的问题检索优化模型

2 文本语义表征比对

2.1 基于深度语义的文本表征模型

研究表明,在大规模文本数据集上预训练的模型有助于文本表征和其他自然语言处理任务,可以避免从头开始训练新模型.常见的预训练模型是词嵌入,典型代表是 Word2Vec 和 GloVe. BERT^[14]在文本特征提取上的显著优势在许多 NLP 任务上有很好的应用,包括问答、文本分类等. BERT 语句嵌入空间存在的词频

偏差和低频词分散稀疏的问题导致直接通过 BERT 模型产生的句子嵌入不能很好地捕捉句子的潜在含义,具体体现为用 BERT 对那些语义上完全无关的句子对编码后也可以获得较高的相似度分数.针对 BERT 模型的这一缺点,文献 [15] 提出了 Sentence-BERT,该方法对预训练的 BERT 进行修改,使用孪生网络和三胞胎网络来获得在语义上有足够意义的句子向量. Sentence-BERT 在保证准确性的同时又缩短了计算时间,使其更

适用于基于语义的信息检索. 由于 Sentence-BERT 存在训练和预测不一致的问题使得该模型的训练过程不稳定且很难根据具体情况适时调整参数. 因此, 文献 [16] 提到了 CoSENT, 该方法提出了一种排序的损失函数 CoSENT_loss, 有效优化了检索阶段用来计算文本相似度的余弦相似度值, 将此方法运用于 Sentence-BERT 的训练阶段可以使模型的训练过程更贴近预测, 模型的收敛速度和效果更好.

因此, 本文用 CoSENT 方法优化了 Sentence-BERT 的结构, 优化后的模型结构如图 2 所示. 为了使构筑的文本表征模型满足于不同领域的智能客服系统, 本文在模型的训练上利用迁移学习的思想, 先利用通用领域的的数据训练, 再使用垂直领域的的数据进行迁移, 逐步优化, 使模型具有普适性.

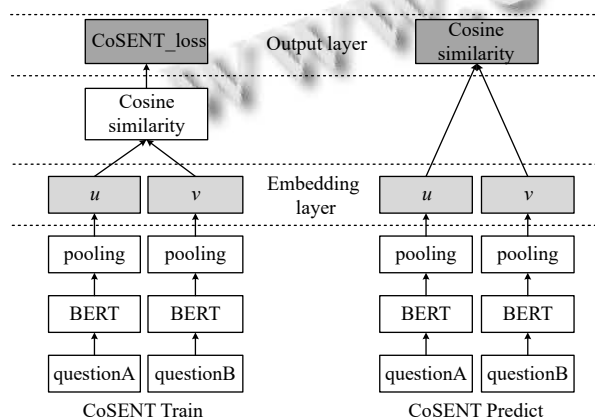


图 2 基于深度语义的文本表征模型

2.2 待检索问题与候选问题的相似度度量

为了判断待检索问题和知识库问题的语义相似度, 该条待检索问题需要和知识库中所有标准问题进行交互对比, 而交互会损耗大量的时间使得检索模型的响应速度变慢. 因此, 本文使用了应用级别的近似向量检索库 Faiss^[17] 来加速检索相似向量的计算过程. 利用第 2.1 节得到的文本表征模型, 可以事先计算好知识库 KB 中所有标准问题的特征表示向量, 获得集合 $S = \{s_1, s_2, \dots, s_n\}$ 放入库中, 当用户咨询产生一条待检索问题时, 通过文本表征模型获得待检索问题的语义向量 X , 然后通过计算向量 X 和集合 S 中所有标准问题特征表示向量的空间距离, 可以快速获得待检索问题和案例库中所有标准问题之间的深度语义得分, 并基于该深度语义得分对候选标准问题进行排序, 返回 top-k 条与待检索

问题语义最相似的检索候选集. 同时, 得到的 top-k 个深度语义得分就作为待检索问题与候选问题的相似度度量.

待检索问题与候选问题的相似度度量的计算公式如式 (1) 所示:

$$Deep_Score = Sim_{L2-norm}(CoSENT(Query), CoSENT(Question_i)), i \in [1, m] \quad (1)$$

其中, $Deep_Score = \{ds_1, ds_2, \dots, ds_m\}$ 表示待检索问题与候选问题的相似度度量集合, $CoSENT(Query)$ 为文本表征模型, 结果为通过文本表征模型生成的特征向量. $Sim_{L2-norm}$ 为用 cosine 计算归一化的两向量的相似度函数. 记 $CoSENT(Query)$ 的结果为 A , $CoSENT(Question_i)$ 的结果为 B . 假定 A 和 B 是两个向量, $Sim_{L2-norm}$ 的计算如式 (2) 所示:

$$Sim_{L2-norm}(A, B) = \cos(\theta) = \frac{A' \cdot B'}{\|A'\| \cdot \|B'\|} \quad (2)$$

其中, A' 、 B' 为归一化后的向量, $A' = \frac{A}{\|A\|}$, $B' = \frac{B}{\|B\|}$.

3 精确检索特征计算

3.1 待检索问题语义块与候选问题、答案语义块的相似度度量

第 2.2 节所提到的外部特征是从待检索问题的整体结构出发来构造的, 体现于待检索问题与候选问题的相似度度量. 在精确检索特征计算阶段兼顾客问语句的细节特征, 将所构造的外部特征聚焦于句子的最小单位—词. 句子由词构成. 同理, 句子的语义由词的语义构成, 每一个词可以看作是一个语义块, 语义块是句子的语义构成单位. 因此, 候选问题如果可以表示待检索问题, 那么候选问题的语义块与待检索问题的语义块必然存在相似之处. 因此, 可以使用待检索问题的语义块与候选问题的语义块的相似度度量来作为待检索问题的另一外部特征.

在进行问题与答案匹配时, 从概念上看是在衡量问题和答案在语义上的匹配程度. 从问题和答案中提取语义特征可以有效增强答案选择模型^[18]. 因此, 本文将待检索问题的语义块与候选问题答案的语义块的相似度度量作为又一外部特征来衡量问题与答案之间的语义关联度, 辅助问题检索.

定义待检索问题语义块与候选问题语义块、候选问题答案语义块的相似度度量计算公式分别如式 (3), 式 (4) 所示:

$$Sim(Q, KBQ) = \frac{\sum_{i=1}^m \max(\text{synonyms}(q_i, kq_j))}{m}, j \in [1, n] \quad (3)$$

$$Sim(Q, KBA) = \frac{\sum_{i=1}^m \max(\text{synonyms}(q_i, ka_j))}{m}, j \in [1, n] \quad (4)$$

其中, Q 表示待检索问题的语义块集合, 有 $Q = \{q_1, q_2, \dots, q_m\}$. KBQ 表示知识库中待检索问题的语义块集合, 有 $KBQ = \{kq_1, kq_2, \dots, kq_n\}$. KBA 表示待检索问题的候选问题答案的语义块集合, $KBA = \{ka_1, ka_2, \dots, ka_n\}$. $Sim(Q, KBQ)$ 、 $Sim(Q, KBA)$ 分别代表待检索问题的语义块与候选问题、答案的语义块的相似度度量. $\text{synonyms}(q_i, kq_j)$ 为相似函数, 其本质是一个开源的计算语义块相似度的工具包 Synonyms, 结果为两个词语的语义相似度. m, n 均为语义块的个数, \max 为最大值函数.

3.2 待检索问题实体、关系与候选问题实体、关系的相似度度量

语义块作为句子语义的构成单位, 按其语义功能分类, 可分为主语义块和辅语义块两大类; 主和辅是从句义表达的角度划分的, 主语义块是句义的“必不可少”的成分, 辅语义块是句义的“可有可无”的成分^[19]. 如用户问题“进场和出场的时间是否有严格限制?”, 对这个句子分析, 发现“进场”“出场”“时间”“是否”“有”“限制”这几个语义块完全可以表达句子的语义, 为主语义块; 而“和”“的”“严格”等语义块只起到修饰和补充的作用, 仅为了体现句子的完整性而存在, 为辅语义块. 因此, 忽略辅语义块对句子语义的影响, 基于句子的主语义块构造待检索问题的外部特征是我们研究的重点. 一个句子的主语义块可以表达两个方面的内容, 一是表达对象, 二是对象的表现, 前者是“什么”, 后者是“怎么样”. 因此一个句子的主语义块由实体和关系组成, 实体指的是客观存在并可相互区别的事物, 关系指的是实体与实体之间的联系. 在上述用户问题的主语义块中, “进场”“出场”和“时间”为实体, “是否”“有”和“限制”为关系. 因此, 问题一般可以概括为实体-关系或者是实体-关系-实体^[20]. 综上, 待检索问题实体、关系与候选问题实体、关系的相似度度量可以表示两者之间的语义关系, 可以作为待检索问题的外部特征辅助问题检索.

待检索问题实体、关系与候选问题实体、关系的

相似度度量的计算公式分别如式 (5)、式 (6) 所示:

$$Sim(QE, KBQE) = \sum_{i=1}^m \max(\text{synonyms}(qn_i, kqn_j)) / m, j \in [1, n] \quad (5)$$

$$Sim(QR, KBQR) = \max(\text{synonyms}(qv_i, kqv_j)), i \in [1, m], j \in [1, n] \quad (6)$$

其中, QE 表示待检索问题的实体集合, $QE = \{qn_1, qn_2, \dots, qn_m\}$. QR 表示待检索问题的关系集合, $QR = \{qv_1, qv_2, \dots, qv_m\}$. $KBQE$ 表示候选问题的实体集合, $KBQE = \{kqn_1, kqn_2, \dots, kqn_n\}$. $KBQR$ 表示候选问题的关系集合, $KBQR = \{kqv_1, kqv_2, \dots, kqv_n\}$. $Sim(QE, KBQE)$ 、 $Sim(QR, KBQR)$ 分别代表待检索问题与候选问题实体、关系的相似度度量. $\text{synonyms}(qv_i, kqv_j)$ 为相似函数, 其本质是在本研究中提出的基于文献 [21] 和近义词工具包 Synonyms 的词语相似度混合计算方法, 该方法在一定程度上既保证了词语相似度的准确性又避免了《同义词词林》未收录词对词语相似度的计算造成的影响, 提高了外部特征的准确度. 其结果为两个词语的语义相似度. 其他定义同上.

4 候选问题比对状态判断

由于待检索问题的一系列外部特征直接影响着检索模型能否精确返回待检索问题的唯一候选问题, 因此可以将检索模型中的精确检索问题看成是一个一系列特征下的分类预测问题. 随机森林 (RF) 算法具有良好的准确性和高效性, 被认为是性能最优的分类算法之一^[22], 大量研究表明: RF 具有较高的预测精度, 对数据中存在异常值和噪声等不敏感, 鲁棒性强^[23]. 随机抽样在 RF 算法中的良好应用使得 RF 算法拥有较强的泛化能力且实现简单. 此外, RF 训练速度快, 预测精度较高, 能够在一定程度上避免计算陷入过拟合, 且对各种数据集的运算均适用^[23]. 本研究采用 RF 算法, 以上述提到的 5 种特征量作为输入量, 构建基于随机森林的问题状态预测模型. 模型构建流程如图 3 所示.

5 实验与结果

5.1 实验数据

本文选用的文本表征模型的通用领域训练数据集为 Chinese-STS-B 数据集, 该数据集由英文原数据集 STS-B^[24]生成. 每一条数据由句子 1、句子 2 和两个句

子的相似性评分构成. 原始 Chinese-STS-B 数据集的相似度得分为 0 到 5, 数值越大表示两个句子越相似, 为了简化数据的处理过程, 本文设置了一个阈值将文本对相似性评分转化为 0 和 1, 0 表示文本对中的文本不匹配, 1 表示匹配.

为了从工程应用的角度验证本文问题检索模型的有效性, 采集了目前对智能客服系统有刚需的文旅机

构的问答数据构建文旅机构问答数据集 CTOQA (cultural and tourism organizations question answering dataset). 并基于该数据集构建 fine-tune 数据集对文本表征模型进行领域迁移, 构造格式同 Chinese-STS-B 数据集. 采集的问答数据由人工整理并标注的网爬文旅机构问答数据和从特定文旅机构收集的用户咨询问答对组成. STS-B 数据集和文旅机构问答数据集的具体数量如表 1 所示.

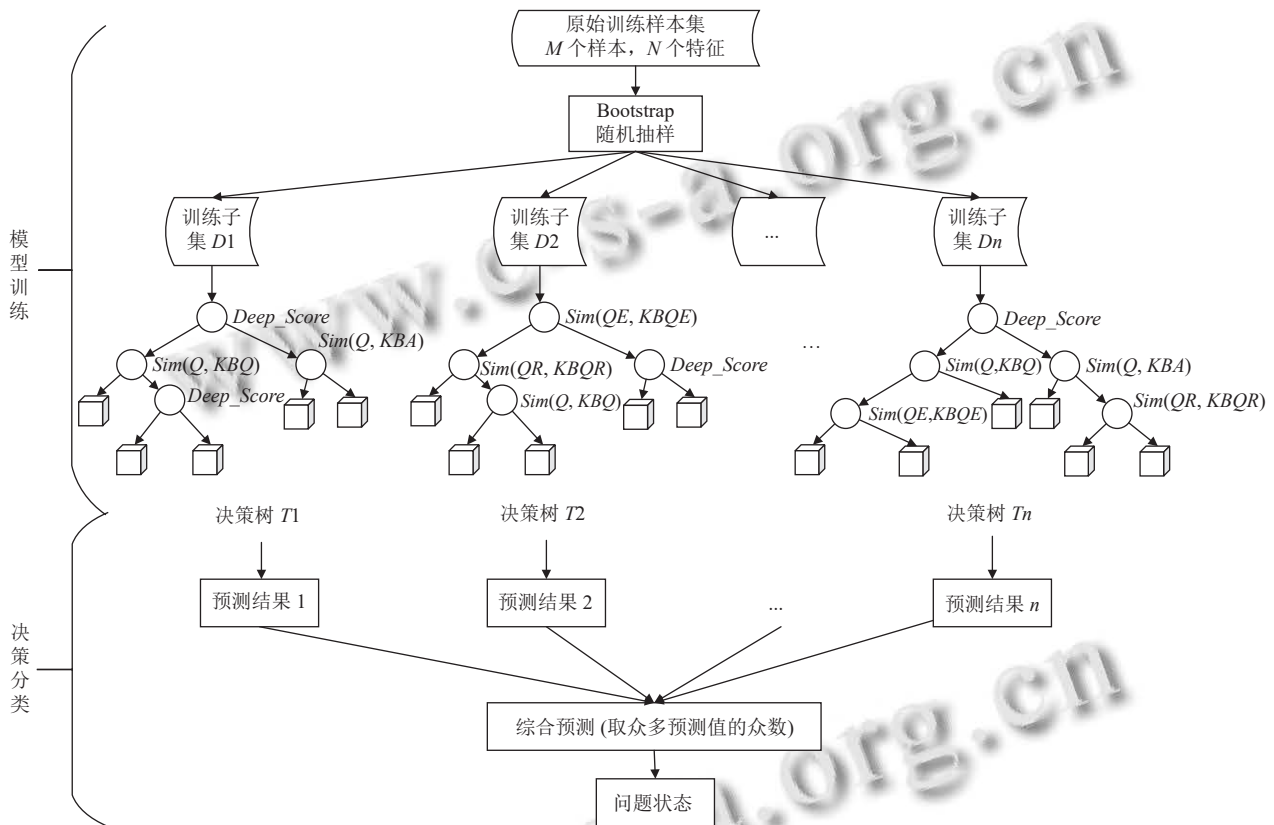


图 3 基于随机森林的问题状态预测模型

表 1 实验数据集

分类	Chinese-STS-B	CTOQA
Training	5231	2925
Validation	1458	816
Test	1361	759

随机森林模型的数据集为基于问答数据构造的 1 700 个特征样本, 按 7:3 的比例划分为训练集和测试集, 训练集样本和测试集样本分别为 1 190 和 510. 检索模型的测试数据集为已经被人工标记的 1 600 个用户问题.

5.2 度量标准

检索模型有效性需要通过检索结果排序的准确性

来验证^[2], 本文选用了信息检索任务中的 $P@N$ 、 $MAP@N$ 、 MRR 等指标来评测检索模型的性能, 下面详细阐述了各指标的概念及计算公式.

(1) $P@N$ (precision): 检索结果中排名前 N 的候选问题的准确率, 计算公式如式 (7) 所示:

$$P@N = \frac{1}{|C|} \sum_{q \in C} \frac{\sum_{i=1}^N rela(i)}{N} \quad (7)$$

其中, C 表示待检索问题的候选问题集, $rela(i)$ 表示第 i 个候选问题与待检索问题是否相似, 相似取值为 1, 不相似取值为 0. 为了验证检索模型是否能为用户返回唯一正确答案, 本文将 N 赋值为 1, 只关注检索结果中排

名第一的候选问题的准确率。

(2) $MAP@N$ (mean average precision): 检索结果中排名前 N 的候选问题的平均准确率, 计算公式如式 (8) 所示:

$$MAP@N = \frac{1}{|C|} \sum_{q \in C} \frac{\sum_{i=1}^N per(i)rela(i)}{N} \quad (8)$$

其中, $per(i)$ 表示前 i 个候选问题中与待检索问题相似的问题所占的比例, 其他含义同上。本文将 N 赋值为 10, 只关注检索结果中排名前 10 的候选问题的平均准确率。

(3) MRR (mean reciprocal rank): 检索结果中若第 1 个返回的正确答案排名为 n , 则 MRR 得分为 $1/n$, 对整个检索模型来说, 计算公式如式 (9) 所示:

$$MRR = \frac{1}{|C|} \sum_{i=1}^{|C|} \frac{1}{rank(i)} \quad (9)$$

其中, $rank(i)$ 表示在第 i 个待检索问题的检索结果中, 第 1 个与用户问题相似的问题返回的排名, 其他含义同上。

5.3 实验环境

本文实验使用的硬件环境是 Linux CentOS 7.8.2003 系统, 使用的 GPU 是 Nvidia GeForce RTX 3060, 使用的软件环境基于 Python 3.6, 框架是 TensorFlow 1.15.0。

5.4 实验参数设置

本文在训练文本表征模型时使用的预训练模型是谷歌中文预训练 BERT 模型, 该模型有 12 层, 768 个隐藏单元, 多头注意力头数是 12, 参数的总量是 110 M。基于 BERT 模型对文本表征模型进行训练, epoch 为 5, 最大序列长度为 64, 优化器是 Adam, 学习速率是 $2E-5$, 训练的 Batch_size 为 32。

5.5 实验结果与分析

5.5.1 与基线模型的对比及分析

为了验证本文问题检索模型的有效性, 本文选取了几种典型检索模型作为基线在 CTOQA 数据集上进行了实验对比, 即先用 Chinese-STS-B 数据集进行预训练, 然后用 CTOQA 数据集进行 fine-tune 和测试。实验结果如表 2 所示。

实验表明, 本文所提出的基于深度语义的三阶段问题检索模型 TSFR-RM 在各个指标上均优于其他检索模型。

相较于基于特征的检索模型 BM25、Jaccard, 本文

模型解决了词义局限与结构局限等特点, 可以捕捉到问题的潜在语义, 在一定程度上提高了文本相似度任务的准确率, 提高了检索模型的各项指标。

表 2 各模型在 MRR 、 $MAP@10$ 、 $P@1$ 上的对比

模型	$P@1$	$MAP@10$	MRR
BM25 ^[25]	0.7212	0.5716	0.7782
Jaccard ^[26]	0.7272	0.5696	0.7961
Sentence-BERT ^[15]	0.7172	0.5825	0.8047
Poly-Encoder ^[9]	0.7412	0.6088	0.8102
ColBERT ^[10]	0.7655	0.6350	0.8542
SEED-Encoder ^[11]	0.7882	0.6386	0.8452
SimCSE ^[12]	0.8194	0.6422	0.8322
TSFR-RM	0.9124	0.6562	0.8778

相较于基于表征的检索模型 Sentence-BERT, 本文模型融入了问题的外部语义特征提高了用户问题与候选问题之间有效的特征交互, 获得了较为准确的问题语义表征, 使得其检索指标优于该模型; 相较于 Poly-Encoder 和 ColBERT, 本文模型未采用端到端的模型, 增强了用户问题与候选问题的细粒度深层次交互, 显著增强了检索模型的表达能力, 提高了检索模型的各项指标; 相较于 SEED-Encoder 和 SimCSE, 本文模型使用 CoSENT 方法优化了检索阶段用来计算文本相似度的余弦相似度值, 有效地学习到了文本表征, 使得模型的语义表征能力得到极大的增强, 提高了检索模型的性能。

具体而言, 对于检索指标 $P@1$ 、 $MAP@10$ 、 MRR , 相较于表现最佳的相关模型, 分别有 9.3%、1.4%、4.56% 的相对提升。TSFR-RM 的优势是: 在结构上, 三阶段式的检索模型构建流程增加了模型多任务学习的能力, 改善了其他检索模型只针对单一特定任务给整个检索流程带来的局限性; 在理论上, 通过为用户问题构建外部特征, 大大增强了用户问题与候选问题的多角度细粒度的深层次交互, 进一步获取到了用户问题深层次的特征语义信息。因此, 本文模型采用三阶段式结构充分融合了用户问题的表征与特征, 优化了检索模型的整体构建流程既保证了检索模型的召回率也提高了检索模型的精准度, 相比于其他检索模型, 在指标上有较大的提升。

5.5.2 消融实验

为了验证本文所述 3 个阶段在模型整体中的贡献以及每个阶段对模型实验结果产生的影响, 在 CTOQA 数据集上进行了消融实验。设置实验如下: 剔除掉模型

的第2、3阶段(表3第1行),剔除掉模型的第3阶段(表3第2行),涉及模型的全部阶段(表3第3行).实验结果如表3.其中,第1阶段用TR(text representation)表示,第2阶段用PF(precise feature),第3阶段用RF(random forest)表示.

数据表明,将模型的第2、3阶段剔除后,与原模型相比,性能差距较大,在文本语义表征阶段可以找到答案,但不能找准答案;将模型第3阶段剔除后,可以使得模型的性能得到最大程度的保留,与原模型性能差距不大,在第1阶段的基础上,融入为问题构建的外部特征,可以大大增强模型找准答案的能力;保留模型的前两个阶段,加入第3个阶段,随机森林模型的加入减弱了外部特征中异常值的影响,大大提高了模型的整体性能.由此可见,3个阶段各自发挥的作用实现了检索模型的整体优化和性能提升.

表3 模型消融实验结果

模型	P@1	MAP@10	MRR
TR	0.7878	0.6390	0.8366
TR+PF	0.8666	0.6420	0.8533
(TR+PF+RF)原模型	0.9124	0.6562	0.8778

6 结论与展望

本文提出了一种基于深度语义的三阶段式问题检索模型,运用多任务学习的思想将模型的构建流程划分为3个阶段,进行不同任务的学习.在第1阶段,运用迁移学习的思想得到了更适用于特定领域的文本表征模型,并用该模型得到了用户问题与知识库问题的文本表征.通过计算知识库问题与用户问题的文本表征的相似度,为用户问题缩小检索范围.在第2个阶段,令用户问题和知识库问题进行多角度细粒度的深层次交互构建语义特征辅助问题检索.在最后一个阶段,基于得到的用户问题的表征和特征,采用随机森林的问题状态预测模型,通过预测候选问题的问题状态,为用户返回唯一正确答案.实验表明了该方法有效性显著高于实验比对的基线模型.目前,本文所提出的检索模型已经成功应用于国内某文旅机构智能客服系统,对该系统运营40天的数据分析表明,共产生有效问答数据1671条,为用户正确解答疑惑的问答数据有1543条,占92.34%,与上述实验数据相互印证,进一步验证了模型的有效性.本文在构造问题的外部特征上考虑了

句子的语义构成单位语义块的相似度,但在实验过程中发现语义块的相似度会直接影响到检索模型的精确度,未来的研究工作将更多着眼于提出更好的计算策略来提高词语相似度计算结果的准确度,进一步提高检索模型的准确度.

参考文献

- 赵芸,刘德喜,万常选,等.检索式自动问答研究综述.计算机学报,2021,44(6):1214-1232.[doi:10.11897/SP.J.1016.2021.01214]
- Guo JF, Fan YX, Pang L, et al. A deep look into neural ranking models for information retrieval. Information Processing & Management, 2020, 57(6): 102067. [doi: 10.1016/j.ipm.2019.102067]
- Zhou GY, Huang JX. Modeling and learning distributed word representation with metadata for question retrieval. IEEE Transactions on Knowledge and Data Engineering, 2017, 29(6): 1226-1239. [doi: 10.1109/TKDE.2017.2665625]
- Lan YS, Wang SH, Jiang J. Knowledge base question answering with a matching-aggregation model and question-specific contextual relations. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2019, 27(10): 1629-1638. [doi: 10.1109/TASLP.2019.2926125]
- Clark P, Etzioni O, Khot T, et al. Combining retrieval, statistics, and inference to answer elementary science questions. Proceeding of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2016. 2580-2586. [doi: 10.1609/aaai.v30i1.10325]
- Fan HJ, Ma ZY, Li HQ, et al. Enhanced answer selection in CQA using multi-dimensional features combination. IEEE Access, 2019, 24(3): 346-359. [doi: 10.26599/TST.2018.9010050]
- Aithal SG, Rao AB, Singh S. Automatic question-answer pairs generation and question similarity mechanism in question answering system. Applied Intelligence, 2021, 51(11): 8484-8497. [doi: 10.1007/s10489-021-02348-9]
- Gao X, Niu K, He ZQ. A convolutional neural network model for non-factoid Chinese answer selection. Proceedings of the 2017 2nd IEEE International Conference on Computational Intelligence and Applications (ICCI). Beijing: IEEE, 2017. 270-274. [doi: 10.1109/CIAPP.2017.8167221]
- Humeau S, Shuster K, Lachaux MA, et al. Poly-encoders: Transformer architectures and pre-training strategies for fast and accurate multi-sentence scoring. arXiv:1905.01969,

- 2020.
- 10 Khattab O, Zaharia M. ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2020. 39–48. [doi: [10.1145/3397271.3401075](https://doi.org/10.1145/3397271.3401075).]
- 11 Lu SQ, He D, Xiong CY, *et al.* Less is more: Pretrain a strong Siamese encoder for dense text retrieval using a weak decoder. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Punta Cana: Association for Computational Linguistics, 2021. 2780–2791. [doi: [10.18653/v1/2021.emnlp-main.220](https://doi.org/10.18653/v1/2021.emnlp-main.220)]
- 12 Gao TY, Yao XC, Chen DQ. SimCSE: Simple contrastive learning of sentence embeddings. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Punta Cana: Association for Computational Linguistics, 2021. 6894–6910. [doi: [10.18653/v1/2021.emnlp-main.552](https://doi.org/10.18653/v1/2021.emnlp-main.552)]
- 13 Qiang JP, Qian ZY, Li Y, *et al.* Short text topic modeling techniques, applications, and performance: A survey. IEEE Transactions on Knowledge and Data Engineering, 2022, 34(3): 1427–1445. [doi: [10.1109/TKDE.2020.2992485](https://doi.org/10.1109/TKDE.2020.2992485)]
- 14 Devlin J, Chang MW, Lee K, *et al.* BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis: Association for Computational Linguistics, 2019. 4171–4186. [doi: [10.18653/v1/N19-1423](https://doi.org/10.18653/v1/N19-1423)]
- 15 Reimers N, Gurevych I. Sentence-BERT: Sentence embeddings using siamese BERT-networks. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong: ACL, 2019. 3982–3992. [doi: [10.18653/v1/D19-1410](https://doi.org/10.18653/v1/D19-1410)]
- 16 Liu KH, Wang J, Zhang XJ. YNU-HPCC at SemEval-2022 task 2: Representing multilingual idiomaticity based on contrastive learning. Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022). Seattle: ACL, 2022. 211–216. [doi: [10.18653/v1/2022.semeval-1.26](https://doi.org/10.18653/v1/2022.semeval-1.26)]
- 17 Chen W, Chen JC, Zou FH, *et al.* Vector and line quantization for billion-scale similarity search on GPUs. Future Generation Computer Systems, 2019, 99: 295–307. [doi: [10.1016/j.future.2019.04.033](https://doi.org/10.1016/j.future.2019.04.033)]
- 18 Huang HY, Wei XC, Nie LQ, *et al.* From question to text: Question-oriented feature attention for answer selection. ACM Transactions on Information Systems, 2019, 37(1): 6.
- 19 黄曾阳. HNC 理论概要. 中文信息学报, 1997, 11(4): 12–21.
- 20 Qiao FF, Zhu XJ. Domain intelligent Q&A user intention recognition based on keyword separation. Proceedings of the 2020 International Conference on Culture-oriented Science & Technology (ICCST). New York: IEEE, 2020. 224–229. [doi: [10.1109/ICCST50977.2020.00049](https://doi.org/10.1109/ICCST50977.2020.00049)]
- 21 彭琦, 朱新华, 陈意山, 等. 基于信息内容的词林词语相似度计算. 计算机应用研究, 2018, 35(2): 400–404. [doi: [10.3969/j.issn.1001-3695.2018.02.018](https://doi.org/10.3969/j.issn.1001-3695.2018.02.018)]
- 22 常硕, 张彦春. 基于袋外预测和扩展空间的随机森林改进算法. 计算机工程, 2022, 48(3): 1–9. [doi: [10.19678/j.issn.1000-3428.0063249](https://doi.org/10.19678/j.issn.1000-3428.0063249)]
- 23 赵勇, 赵姜雪慧. 基于随机森林模型的滑艇艇剩余阻力预测. 华中科技大学学报(自然科学版), 2021, 49(11): 118–122. [doi: [10.13245/j.hust.211120](https://doi.org/10.13245/j.hust.211120)]
- 24 Cer D, Diab M, Agirre E, *et al.* SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). Vancouver: Association for Computational Linguistics, 2017. 1–14. [doi: [10.18653/v1/S17-2001](https://doi.org/10.18653/v1/S17-2001)]
- 25 Kim MY, Rabelo J, Okeke K, *et al.* Legal information retrieval and entailment based on BM25, transformer and semantic thesaurus methods. The Review of Socionetwork Strategies, 2022, 16(1): 157–174. [doi: [10.1007/s12626-022-00103-1](https://doi.org/10.1007/s12626-022-00103-1)]
- 26 Riyanto R, Azis A. Implementation of the Jaccard similarity algorithm on answer type description. International Journal of Informatics and Information Systems, 2022, 5(2): 76–83. [doi: [10.47738/ijiis.v5i2.130](https://doi.org/10.47738/ijiis.v5i2.130)]

(校对责编: 孙君艳)