文章编号：2095-2163(2022)08-0054-06 中图分类号：TP391.1 文献标志码：A

智能运维场景下的问答系统设计与应用

王 越，赵艳兴，贺 霆，赵逢波，王旭鹏，张治国
（北京宝兰德软件股份有限公司，北京 100089）

摘 要：智能运维(Artificial Intelligence for IT Operations，AIOps)场景中，问答系统的应用可以辅助运维人员完成运维任务，查询运维知识，降低企业运维成本。传统的问答系统功能单一，通常只能完成特定任务或知识查询中的一种。本文设计的多功能问答系统集成了 4 类问答功能，即：任务型、知识图谱型、问答对型、闲聊型，同时针对任务型场景深度开发 Rasa 框架，并采用基于 BERT 的改进型意图识别神经网络结构；通过真实的运维数据验证了该智能问答系统性能。结果表明，本文设计的问答系统不仅可以毫秒级返回问答结果，并且任务型场景中的意图识别准确率在原有基础上有了明显提升。

关键词：智能运维；问答系统；Rasa 框架；意图识别

Design and application of question answering system in intelligent operation and maintenance scenarios

WANG Yue, ZHAO Yanxing, HE Ting, ZHAO Fengbo, WANG Xupeng, ZHANG Zhiguo
(Beijing Baolande Software Corporation, Beijing 100089, China)

【Abstract】 In AIOps scenarios, the application of the question answering system can assist operation and maintenance personnel to complete operation and maintenance tasks, query operation and maintenance knowledge, which can reduce the operation and maintenance cost of enterprises. Usually, traditional question answering systems have a single function and can only complete one of specific tasks or knowledge queries. This paper designs a multifunctional question answering system architecture that integrates four question answering functions (Task, KBQA, FAQ, Chat). At the same time, the Rasa framework is deeply developed for task-based scenarios, and an improved intent recognition neural network structure based on BERT is adopted; The performance of the intelligent question answering system is verified by real operation and maintenance data. The results show that the question answering system designed in this paper can not only return question and answer results in milliseconds, but also the accuracy of intent recognition in task-based scenarios has been significantly improved on the original basis.

【Key words】 AIOps; question answering system; Rasa framework; intent recognition

0 引 言

在智能运维 AIOps 领域，问答系统可以帮助企业运维人员完成运维工作，如作业下发、运维知识查询等。通过与问答系统一问一答的交互，运维人员能够更加快速、便捷地完成运维工作，降低企业的运维成本。

Rasa 是当前主流的开源问答系统框架，用于构建聊天机器人和智能助手[1]。Rasa 框架的模块化和灵活设计使开发人员能够轻松构建新的扩展和功能，包含自然语言理解(NLU)模块与核心(Core)模块两部分。NLU 模块用于解析用户输入语句，识别语句中的实体、意图等信息；Core 模块负责对话管理，用于跟踪对话状态，执行对话策略，并提供可编辑的模板，方便开发人员设计多轮对话。

常见的问答系统通常功能较为单一，只能完成某些特定功能的问答场景。百度的 AnyQ 框架采用自研的 SimNet 网络结构完成研发问题对(FAQ)的问答场景。58 同城采用 qa_match 框架，基于深度学习的 2 层架构，完成 FAQ 场景问答。美团大脑基于问答对数据与知识图谱构建 FAQ 和图谱问答(KBQA)。

本文构建了一种高效的多功能问答场景融合架构，可以将任务型(Task)、FAQ、KBQA、闲聊型问答(Chat)整合为一体，通过分层的插件化设计思路，保证系统可以毫秒级快速响应用户多功能查询需求。架构中的 Task 场景与 Chat 场景采用 Rasa 框架开发，包括意图识别、词槽提取、对话状态管理、多轮对话设计等。针对运维场景的数据特点，在 Rasa 的

作者简介：王 越(1989 -)，男，硕士，高级工程师，主要研究方向：问答系统、知识图谱；赵艳兴(1977 -)，男，学士，高级工程师，主要研究方向：智能运维；贺 霆(1991-)，男，学士，高级工程师，主要研究方向：智能运维；赵逢波(1992-)，男，硕士，高级工程师，主要研究方向：问答系统、知识图谱；王旭鹏(1992-)，男，学士，高级工程师，主要研究方向：机器学习、智能运维；张治国(1978-)，男，学士，高级工程师，主要研究方向：分布式计算、智能运维。

收稿日期：2022-02-21

NLU 模块采用了本文提出的基于 BERT 的改进型意图识别神经网络结构，明显提升了意图识别模型准确率。最后，通过问答系统部署现场的真实数据，验证了本文提出的多功能问答场景融合架构与改进型意图识别模型性能的有效性。

1　系统架构设计

问答系统要求能够为不同需求的运维人员提供多种运维功能。Task 场景是问答系统的核心功能，支持运维人员通过一轮或多轮对话的方式完成具体的运维操作，如“系统健康度分析”、“创建故障群组”、“执行 OOS 系统作业”等。知识型问答场景包括 FAQ 与 KBQA，其中 FAQ 支持运维人员查询系统内的运维知识，而系统中对接的 AMDB 知识图谱数据支持运维人员对系统节点的状态、关系等信息进行查询。Chat 场景内置“功能类”与“人格属性类”等问题，“功能类”问题支持回答用户诸如“你会做什么?”、“你有什么功能”等问题，“人格属性类”问题可以回答“你是谁”、“你几岁了?”等常见闲聊问题。

多功能问答融合示意图如图 1 所示，根据功能可以分为 4 个模块：Task 插件模块、知识型问答模块、Rasa 服务模块、意图置信度赋值模块。在开发过程中，Task 场景和 Chat 场景整合在 Rasa 框架中实现，将意图训练样本同时配置在 Rasa 框架提供的配置文件中，并采用本文实现的改进型意图识别模型进行预测。对此拟展开研究分述如下。

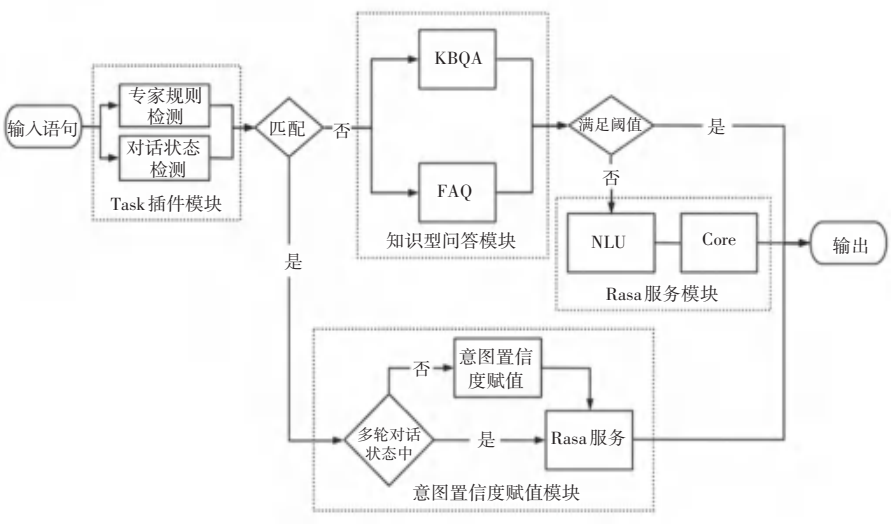


图 1　多功能问答融合示意图

Fig. 1　Schematic diagram of multi-function questions and answers fusion

1.1　Task 插件模块

用户输入语句首先经过 Task 插件模块处理，该模块包含“专家规则检测”与“对话状态检测”两种功能。其中，“专家规则检测”旨在识别实际应用中常见的 Task 场景语句，被识别的语句无需再经过知识型问答模块查询和 Rasa 意图识别模型的预测，可以大幅度提升问答系统的响应速度。系统设计了业务关键词字典与模式匹配两种方式，专家规则处理示意图如图 2 所示。

以 Task 场景中的“创建群组”类别为例，开发者可将“创建群组，拉建群组”等关键词配置在字典中，同时也可配置正则表达式“(创建|拉建)(.+?)(群组|群聊)”作为模式匹配。匹配中关键词字典或正则表达式语句则可认为被“专家规则检测”功能所识别。开发过程中，可以通过配置文件的方式或前端界面工具的方式进行专家规则的配置，也可以根据实际情况灵活地调整专家规则，或在新增意图时便捷地扩展专家规则。不同类别的专家规则之间不可有规则重叠。

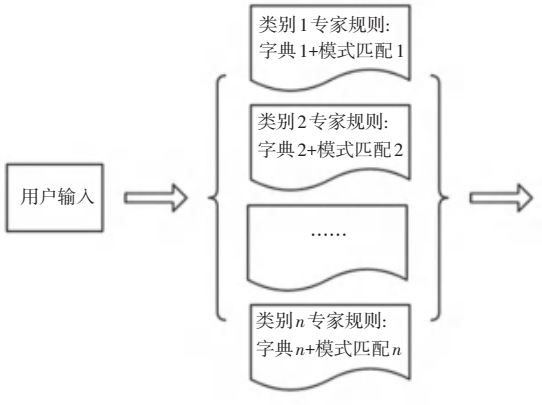


图 2　专家规则处理示意图

Fig. 2　Expert rules processing diagram

对现场用户实际输入的数据进行统计,通过“专家规则检测”功能可以识别约 40%的 Task 问答语句。分析用户使用心理发现:Task 场景中用户在使用时倾向于简短且常用的正式语句。近一半的语句可以被“专家规则检测”功能拦截是合理的,被截留的语句直接进行“意图置信度赋值”模块的处理。

Task 插件模块的“对话状态检测”功能旨在检测当前对话状态是否处于多轮对话状态中,防止意图误识别的产生。问答系统功能设计中,Task 场景支持多轮对话,用户新输入的语句不再需要其它模块的检测,将语句输入到 Rasa 服务中即可。例如:用户输入“创建群组”,系统会追问“请输入群名称”,此时待输入的群名称语义上会较为灵活;用户再输入“智能巡检方法”,该语句与 FAQ 问题重复,“对话状态检测”功能则避免了系统错误返回知识型场景的答案。Rasa 框架的 Core 模块会维护对话状态跟踪、记录对话历史与当前状态等。Rasa HTTP 服务接口可以获取所需状态信息。多轮对话示意图如图 3 所示。图 3 中,序号 1~7 代表多轮问答顺序。

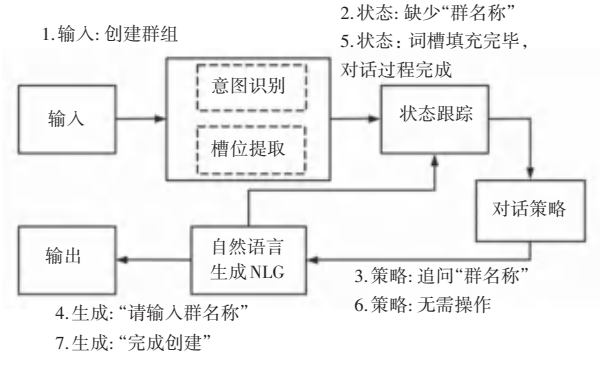


图 3 多轮对话示意图

Fig. 3 Multi-round dialogues schematic diagram

1.2 知识型问答模块

没有被 Task 插件模块识别的语句会流入知识型问答模块,当前系统开发了 KBQA 场景与 FAQ 场景两部分,都是通过对存储在知识库中的知识进行语义相似度的检索,将满足阈值要求的答案返回给用户。

KBQA 模型采用基于答案排序的方式实现[2]。图谱数据存储在图数据库 ArangoDB 中,首先对用户输入语句进行命名实体识别,将提取出的实体放在知识图谱中查找与实体相连的所有三元组,组成候选答案。其次,计算问句与候选三元组的语义相似度,选出最相似的三元组判断是否满足所设定的相似度阈值。基于答案排序的 KBQA 的工作流程图示意如图 4 所示。

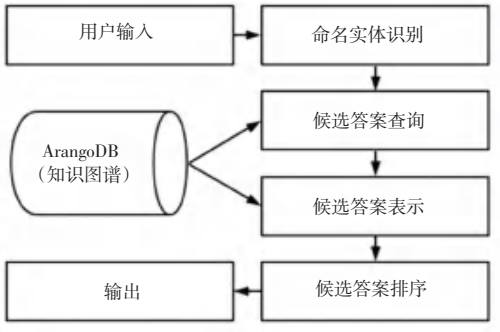


图 4 KBQA 流程示意图

Fig. 4 Schematic diagram of KBQA

FAQ 场景中将问答对数据存储在 Elasticsearch (ES)库中,采用 ES 的倒排索引快速检索用户输入的相似问句。系统在实现过程中引入了近义词表来泛化 ES 的相似度检索效果。以问句“如何 restart 电脑”为例,用户希望在 FAQ 库中匹配到“怎样重启计算机”问句,而倒排索引本身是基于共现词来创建索引,没有近义词来泛化则不能准确找到对应的候选问题。在实际生产中,可以将近义词表放入 ES 中加速查询过程,在 ES 中引入近义词词典如图 5 所示。

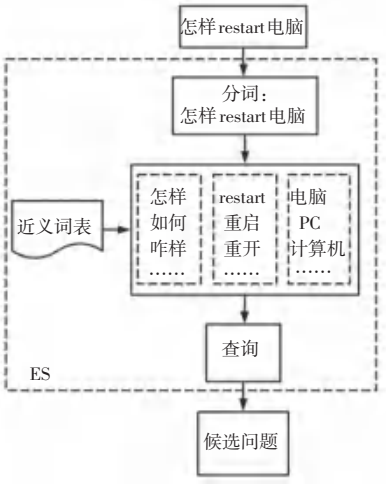


图 5 在 ES 中引入近义词词典

Fig. 5 Introducing a dictionary of synonyms in ES

预测时知识型模块同时调用 KBQA 与 FAQ 的 2 个服务接口,并行预测给出检索结果。当 2 个模型都有结果输出时,依据 FAQ 模型得到的准确率优于 KBQA 模型,因此优先选择 FAQ 的结果。

1.3 Rasa 服务模块

用户输入语句通过知识型问答模块后如果不满足相似度阈值,则进入 Rasa 服务模块。通过框架中 NLU 模块与 Core 模块可以便捷地搭建起 Rasa 问答

服务。Rasa 框架内置了多种意图识别模型与词槽提取模型，系统采用了本文提出的基于 BERT 的改进型意图识别神经网络结构，针对运维场景优化了意图识别效果。Core 模块内置多种问答策略，Form Policy 实现词槽追问功能，如创建群组时追问群名称；Fallback Policy 实现默认回答功能，当输入问答系统不能理解的语句时，系统返回默认语句“不理解您的意思，请换个说法”。

1.4　意图置信度赋值模块

符合 Task 插件模块的语句会流入意图置信度赋值模块。当系统处于对话状态之中时，该模块将问答语句输入 Rasa 服务中进行下一轮对话流程；反之，模块则主动对输入的语句进行类别置信度赋值操作。如用户输入“创建 hadoop 故障群组”，满足“创建群组”的专家规则，模块会主动将其标注为该类别，置信度赋值 1.0。赋值后的语句不再需要 Rasa 服务中的意图识别模型的判别，模块负责将语句的类别与置信度等所需信息存入 Rasa 服务状态中，从而开启接下来的对话流程。

2　意图识别

在问答系统中，意图识别属于文本分类场景，将不同的输入语句判别到对应类别、即完成意图识别功能。Rasa 框架中内置了多种算法功能，如：分词、词向量化、分类算法等，这些算法可以拼接成算法流程进行意图识别。在运维场景中，内置的算法流程在实际数据上效果欠佳，本文针对运维数据特点提出了一种基于 BERT 的改进型意图识别神经网络结构。

2.1　数据分析

Task 场景与 Chat 场景中用户输入语句均为短文本，相较于长文本，短文本对分类模型要求更高。Task 场景下用户输入语句一般由 2 部分组成：语义表示与词槽；而 Chat 场景语句也可以理解为没有词槽的 Task 语句。任务型场景数据见表 1。

表 1　任务型场景数据表

Tab. 1　Data table of task scenarios

类别	语句	词槽
创建群组	建群 / 建个群聊 /	无
	创建智能巡检群组	智能巡检
	创建执行 hadoop 定时作业群组	执行 hadoop 定时作业
执行 OOS	执行作业 / 启动作业 /	无
系统作业	执行 Linux273_nginx 作业	Linux273_nginx

如“创建群组”类意图，用户输入“建群”，“建个群聊”等带有明显语义表示的词语，较容易识别；当输入“创建执行 hadoop 定时作业群组”时，语句中包含的“执行 hadoop 定时作业”的词槽，容易引起分类错误，模型置信度一般也较低。本文尝试了 Rasa 框架内置的基于 scikit-learn 库的多种传统机器学习算法与 Rasa 自研的 DIET 深度学习算法均不能很好地解决这个问题。效果更优的 DIET 模型面对此类数据往往在意图分类时会呈现出“label：创建群组，confidence：0.54；label：执行 OOS 系统作业，confidence：0.44；label：.....”的窘境。排名第一的类别置信度与第二类别置信度相差不大且置信度低，无法超过类别阈值（一般 0.8 以上）。

2.2　模型设计

面对多变的词槽，要求意图识别模型可以学习出语句中不同词语间与不同词序的权重，对语义表示部分的词语需提高权重，对词槽部分的词语降低权重。根据这样的思路，本文设计了更优的意图识别模型。

BERT 预训练语言模型采用双向 Transformer 结构，以 Mask Language Model 和 Next Sentence Prediction 的多任务训练为目标，在自然语言处理等众多领域达到了最优效果。Transformer 结构中的自注意力机制（self-attention）是算法核心，self-attention 的数学表达式可写为：

$$Attention(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \mathrm{softmax}(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}})\boldsymbol{V} \quad (1)$$

其中，d_k 表示 $\boldsymbol{K}$ 向量的维度数，$\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}$ 分别表示查询向量、键向量、值向量，由此推导得到的数学公式为：

$$\begin{cases}\boldsymbol{W}_q\boldsymbol{x}_i = \boldsymbol{Q}\\ \boldsymbol{W}_k\boldsymbol{x}_i = \boldsymbol{K}\\ \boldsymbol{W}_v\boldsymbol{x}_i = \boldsymbol{V}\end{cases} \quad (2)$$

其中，$\boldsymbol{x}_i$ 表示输入语句的词嵌入 2 维矩阵，实际模型训练过程中使用训练样本集的 3 维句向量矩阵计算；$\boldsymbol{W}_q$ 表示查询矩阵。键向量和值向量生成方式与此相同。

$\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}$ 两矩阵相乘可解释 self-attention 的 self 概念，每个词向量会和包括自身在内的句子中所有词进行相乘。相乘后的矩阵除以向量维度 d 的平方根，在计算过程中可以使梯度更加稳定。*Softmax* 中的部分属于点积缩放的注意力机制，得到 *Softmax* 的值后再乘以值向量，最终获得经注意力机制调整的矩阵。

本文选择针对中文优化的 BERT-wwm 模型构建双通道输入的改进型意图识别网络结构,如图 6 所示。

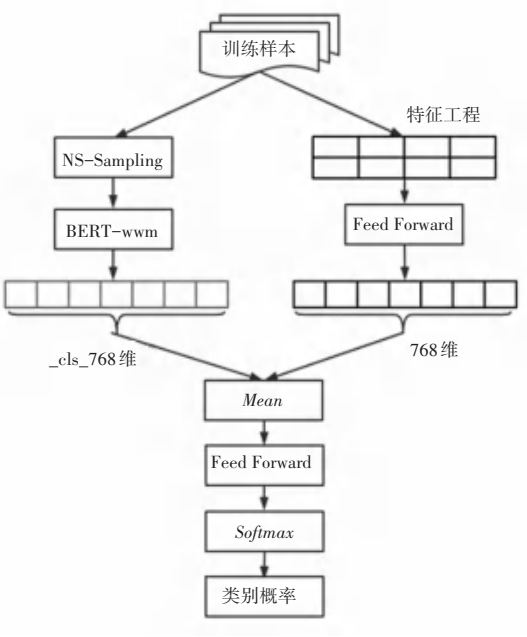


图 6 改进型意图识别网络结构

Fig. 6 Improved intention recognition network structure

左通道通过本文提出的 NS-Sampling(Sampling based on negative sample)算法流程进行训练样本设计与均衡,算法流程图如图 7 所示。

图 7 NS-Sampling 算法流程图

Fig. 7 NS-Sampling algorithm flow chart

2.2.1 数据标记

构建意图训练样本时首先需要将语句中的词槽标记出来,词槽对意图倾向没有贡献,反而会干扰模型训练。如“创建[智能巡检]的群组”,词槽“智能巡检”被标识出,这样做的目的是为了在负样本生成时程序可以自动识别,快速扩展新的 Task 场景,在工程应用中加快模型迭代速度。

2.2.2 负样本设计

本文将意图识别模型支持的意图称为正类,不支持识别的其它意图统称为负类,负类的识别能力对模型性能有着决定性影响,负样本的设计也是关键因素。本文将负样本分为 2 部分:种子负样本与程序生成负样本。对此可给出剖析论述如下。

种子负样本是在编写样本时通过人工整理特别设计的样本,这类样本容易与某些正类意图混淆,如“创建”只出现在“创建群组”的类别中,当用户只输入“创建”时,因 BERT 会将该类词语的权重学习得很高,意图识别模型会误将词语“创建”识别为“创建群组”类别。因此将其配置在负样本中,就可以主动降低这些易混淆词语的权重。

程序生成负样本来自数据标注步骤中标识出来的词槽,这些词槽不可以与正类意图重叠,在模型训练时会由程序自动补充到负样本中。

2.2.3 数据增强

数据增强采用近义词替换的方式来扩展样本较少的类别。以所有类别中样本数量最多的类别为增强数量的上限,通过分词后的样本进行近义词替换。这里不对标记的词槽进行替换。

2.2.4 过采样

经过近义词数据增强后的样本多数情况下可以达到数据均衡,个别类别可能会由于词语的近义词数量不够而增加后的样本依然较少,采用过采样的方法来提升这些类别的样本数量。

2.2.5 特征工程

意图识别模型的右通道提供特征工程入口,通过分析训练样本,本文设计了若干可以辅助提升意图识别的特征,部分特征见表 2。

表 2 特征表

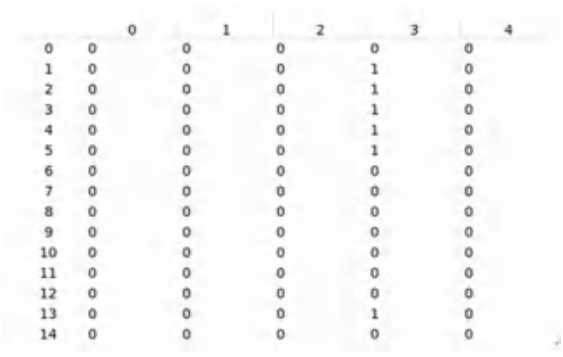
Tab. 2 Feature table

特征 1	特征 2	特征 3	特征 4
匹配 IP 正则	包含 * * 词语	以 * * 开头	以 * * 结尾

一条语句形成的特征稀疏矩阵如图 8 所示,横坐标表示特征个数,序号 0~4、共 5 个特征;纵坐标表示经过输入语句分词后的词,这里有 0~14 个、共 15 个词。矩阵中满足该特征条件的值为 1,否则为 0。

稀疏矩阵输出到全连接网络中,从而获得与左通道中 BERT-wwm 的向量相同的 768 维度向量。将左、右两端的向量进行相加后求平均(Mean)值,可以理解为右侧的 768 维度向量赋予句向量辅助特

征信息，求均值后的向量再连接全连接层与 *Softmax* 层。



	0	1	2	3	4
0	0	0	0	0	0
1	0	0	0	1	0
2	0	0	0	1	0
3	0	0	0	1	0
4	0	0	0	1	0
5	0	0	0	1	0
6	0	0	0	0	0
7	0	0	0	0	0
8	0	0	0	0	0
9	0	0	0	0	0
10	0	0	0	0	0
11	0	0	0	0	0
12	0	0	0	0	0
13	0	0	0	1	0
14	0	0	0	0	0

图 8　特征工程矩阵

Fig. 8　Feature engineering matrix

3　实验

本节设计 2 组实验用来检验多功能问答场景融合架构与改进型意图识别模型的性能。实验测试了问答系统实际响应速度与 Task 场景意图识别预测准确率。这里可做解析表述如下。

3.1　实验一

对于问答系统响应速度的测试，选用 200 条知识型语句，其中 KBQA 语句与 FAQ 语句各 100 条；选用 Task 场景包括负类在内的全部 12 种类别、250 条单轮问答语句，平均每种场景 20 条测试语句。3 种场景实验数据量之比 $KQBA:FAQ:Task = 1:1:2.5$，与实际应用时数据分布接近。实验分别测试了每种场景下单模型的耗时、融合到框架后每种场景的耗时以及综合所有场景的系统平均耗时。测试在 i7-8700 CPU @ 3.20 GHz 环境中进行。问答系统耗时结果见表 3。

表 3　问答系统耗时表

Tab. 3　Time consumption table　　ms

场景	FAQ	KBQA	Task
单模型耗时	12.2	53.6	105.7
融合耗时	14.4	55.8	97.6
平均耗时	71.3	71.3	71.3

实验表明，多功能问答场景的融合架构相比于单模型耗时没有明显增加，其中 Task 场景得益于 Task 插件模块有了小幅度提升，而综合所有场景的平均耗时较少，保证了系统可以 ms 级地快速响应。

3.2　实验二

对于 Task 场景意图识别性能的测试，实验沿用实验一中的 250 条 Task 场景测试语句，对比了基于 scikit-learn 库的传统机器学习算法 Logistic Regression 与 SVM（线性核，即 LinearSVM）、DIET 算法、微调的 $BERT_{base}$ 模型、改进型意图识别模型共 5 种。其中，传统机器学习算法的句向量化方式分别尝试了 N-gram（1，3）+ TF-IDF 与 Jieba 分词 + Word2Vec 的方式，Word2Vec 模型选择基于百度百科的语料训练，模型中包括了词向量与单个字向量。实验评价指标选择准确率见表 4。

表 4　模型预测准确率

Tab. 4　Model prediction accuracy

算法	Logistic Regression		LinearSVM		DIET	$BERT_{base}$	改进型意图识别模型
	TF-IDF	Word2Vec	TF-IDF	Word2Vec			
准确率/%	80.8	86.8	81.6	88.0	94.8	98.0	99.2

实验表明，传统机器学习算法面对运维场景的短文本分类任务性能欠佳，明显低于深度学习模型；基于 BERT 的意图识别模型效果明显高于其他方法，本文提出的基于 BERT 的改进型意图识别模型在 $BERT_{base}$ 的基础上进一步提升了预测准确率。

4　结束语

在 AIOps 场景下，本文设计了多功能问答系统架构，集成了 4 种问答功能，能够辅助运维人员完成运维工作。针对 Task 场景，通过对系统部署现场的实际运维数据进行分析，本文提出了基于 BERT 的改进型意图识别神经网络结构。实验表明，多功能问答系统架构具有较高性能，在多场景融合的情况下依然可以保持 ms 级的快速响应；而采用了基于 BERT 的改进型意图识别神经网络，Task 场景意图识别准确率效果极佳。

参考文献

[1] KONG Xiaoquan, WANG Guan, et al. Conversational AI with Rasa[M]. United Kingdom: Packt Publishing, 2021:25-27.

[2] 段楠，周明. 智能问答[M]. 北京：高等教育出版社，2018:85-92.