基于 AI 语义分析的客户服务对话系统设计

兰一杰 [1]
LAN Yijie

摘 要 面向多轮任务型语义交互场景，文章构建了一套基于人工智能语义分析的客户服务对话系统。研究对象涵盖自然语言输入、上下文状态与企业级多源知识内容。系统由语义理解、对话状态管理、知识融合与响应决策四个模块组成。语义理解模块采用双通道分类结构，结合双向编码器表示（bidirectional encoder representations from transformers, BERT）完成意图识别与槽位抽取；状态管理模块引入门控循环单元（gated recurrent unit, GRU）实现状态表示更新；知识融合模块联合密向量检索与图神经网络（graph neural network, GNN）完成语义匹配与知识推理；响应决策模块通过动作分类与语言生成系统响应。系统在 SMP2020-ECDT、MultiWOZ 2.2 及自建语料上进行评估，采用准确率、F_1 值、BLEU 值及接口性能作为指标。结果表明各模块在识别精度、知识召回与响应质量方面性能良好，具备工程部署可行性。

关键词 人工智能；语义理解；对话系统；多轮交互；对话状态建模；知识融合

doi：10.3969/j.issn.1672-9528.2026.07.011

0 引言

面向任务的客户服务对话系统是自然语言处理与人机交互融合的重要方向，广泛应用于智能客服、信息查询等场景。随着深度学习与预训练语言模型的发展，对话系统由规则驱动逐步转向语义建模为核心的多模块交互架构，对语义理解精度、上下文保持与知识调用能力提出更高要求。近年来，多轮状态建模、知识调度与响应生成成为研究热点，系统性能与工程可部署性同步提升。

李啸天等 [1] 基于本体构建与状态结构建模，提出融合知识蒸馏的对话状态跟踪方法，提升状态预测稳定性与泛化能力。杜维等 [2] 设计模态敏感注意力机制的多模态对话模型，增强了异构信息融合下的语义一致性。蔡辰等 [3] 聚焦适老化交互，构建基于意图迁移链的语义调节模型，提升特殊人群使用体验。熊文剑等 [4] 结合云平台应用，探讨预训练模型的参数优化策略，验证轻量网络在任务型系统中的部署效率。

在此基础上，本文归纳当前对话系统在多轮状态建模、知识融合调度与响应生成一致性方面的瓶颈，构建一套融合语义理解、状态建模、知识推理与响应控制的客户服务对话

1. 南宁市勘测设计院集团有限公司 广西南宁 530000

[2] 庞顺顺 . 边缘计算场景下系统集成的轻量化优化算法设计 [J]. 产品可靠性报告 , 2025(7): 213-214.
[3] 郝舒羽 . 基于边缘计算的智能物联网环境监测系统研究 [J]. 油气田环境保护 , 2026, 36(1): 50-53.
[4] 裴志刚 , 钟天成 , 范江鹏 , 等 . 基于轻量化 AI 模型配网缺陷实时识别与边缘计算部署 [J]. 机械制造与自动化 , 2026, 55(1): 1-6.
[5] 陈智伟 . 面向边缘计算的轻量化人工智能模型研究 [J]. 互联网周刊 , 2025(22): 26-28.
[6] 董浩 , 李劭辉 , 阚诺文 , 等 . 基于深度压缩感知的联合信源信道编码方法研究 [J]. 电子学报 , 2025, 53 (7): 2178-2192.
[7] 廖四龙 , 耿卓强 . 基于边缘计算的多模态数据处理与智能决策系统研究 [J]. 电脑编程技巧与维护 , 2026(5): 85-87.
[8] 杨雄 , 卓禀航 , 吴志诚 , 等 . 边缘计算中轻量化人工智能算法优化方法研究 [J]. 贵州大学学报 (自然科学版), 2026, 43(3): 20-25.
[9] 董甲东 , 桑飞虎 , 郭庆虎 , 等 . 基于深度学习的目标检测算法轻量化研究综述 [J]. 计算机科学与探索 , 2025, 19(8): 2057-2084.

【作者简介】
赵家鹏（1984—），男，云南昆明人，本科，讲师，研究方向：计算机教学。

（收稿日期：2026-06-18 修回日期：2026-07-16）

系统，旨在提升系统对复杂语义输入的处理能力，满足高并发、多轮交互场景下的响应精度与接口性能要求。

1 系统设计核心目标

系统设计旨在构建面向任务型场景的语义驱动对话流程，实现多层级语义解析、动态状态建模与结构化知识融合。语义理解模块需支持多通道输入与非规范化表达的向量映射，适应复杂依存关系解析[4-5]。对话状态管理模块需完成上下文状态序列化与高频交互状态的实时调用，持续跟踪意图转移与语义漂移。知识融合模块需对接关系型数据库、图谱及非结构化文本，构建跨源语义索引机制。响应决策模块基于语义匹配与逻辑约束实现动作选择与生成控制。系统整体需支持 ≥1 200 次请求 /min 的并发能力，平均响应延迟 ≤150 ms。各模块应具备深度学习模型的更新能力，支持参数迭代与多轮适应，保障在复杂语境下的稳定运行与模块解耦。

2 系统总体架构

系统总体架构由语义理解模块、对话状态管理模块、知识融合模块与响应决策模块构成，采用异步消息驱动与统一语义标记机制协调各模块数据交互与逻辑联动，如图 1 所示。

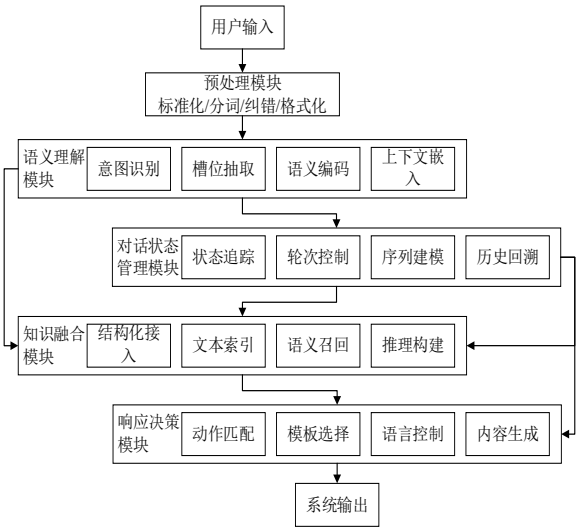

图 1 客户服务对话系统总体架构

语义理解模块完成用户输入的向量编码、意图识别与槽位抽取，输出标准化语义表示。对话状态管理模块接收语义表示，构建时序状态向量，维护轮次上下文及用户交互意图转移信息，为后续模块提供状态依赖条件。知识融合模块基于语义表示与状态向量联合检索异构知识源，完成多结构数据的语义对齐与推理调度，生成可调用的知识单元。响应决策模块综合语义、状态与知识单元执行动作选择、响应构建及策略生成。各模块通过高性能消息队列连接，采用统一数据描述规范，确保语义流、状态流与知识流在多线程环境下的数据一致性与调用闭环。架构支持模块级并行部署与动态加载，满足复杂语义任务的分布式调度需求。

3 系统模块设计

3.1 语义理解模块

语义理解模块基于 BERT-wwm-ext 模型进行深层语义建模，采用 12 层 Transformer 结构，隐藏维度 768，注意力头数 12，总参数量约 102×10^6。语义理解模型结构如图 2 所示。输入经 WordPiece 分词后映射为 token、segment 与 position 嵌入，序列长度截断至 128 token。模型运行于 NVIDIA A100 Tensor Core GPU（40 GB 显存），采用 ONNX Runtime 与 FP16 混合精度推理，平均延迟为 27 ms，输出作为全系统的语义表示基础。

意图识别任务以 BERT 池化输出接全连接层构建 18 类标签分类器，采用带权交叉熵处理类别不平衡问题。TextCNN 作为辅助分支，卷积核尺寸为 2、3、4，每组 128 个，通过拼接后引入门控融合机制与主通道集成，提升边界辨识能力。槽位抽取模块采用 BERT-BiLSTM-CRF 架构，BERT 输出输入双向 LSTM（隐藏维度 256），序列特征传入 CRF 层建模标签转移关系。得分函数用公式表示为：

$$s(x,y)=\sum_{t=1}^{T}\left(\boldsymbol{A}_{y_{t-1},y_t}+\boldsymbol{P}_{t,y_t}\right) \tag{1}$$

式中：$\boldsymbol{A}$ 为标签转移矩阵，$\boldsymbol{P}$ 为 LSTM 输出得分矩阵，y 为预测标签序列，T 为序列长度[6-7]。解码使用维特比算法获取最优标签路径。

语义对齐基于 GRU 结构建模最近 3 轮历史语义，隐状态维度 512。当前语义与历史编码残差连接后，输入多头注意力模块完成对齐计算，Softmax 归一化权重用于控制历史贡献度。输出对齐向量输入后续状态建模与知识融合模块，增强跨轮次语义一致性与歧义消解能力。

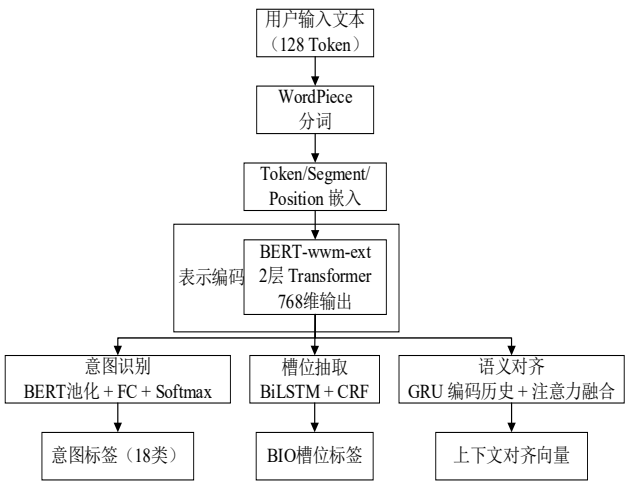

图 2 语义理解模型结构图

3.2 对话状态管理模块

对话状态表示构建采用语义上下文向量与行为控制变量联合建模。当前轮状态向量定义为：

$$\boldsymbol{s}_t = \mathrm{GRU}(\boldsymbol{W}_s[\boldsymbol{c}_t; v_t], \boldsymbol{s}_{t-1}) \tag{2}$$

式中：$\boldsymbol{s}_t$ 为第 t 轮对话状态，$\boldsymbol{c}_t$ 为语义编码向量，v_t 为行为变量，$\boldsymbol{W}_s$ 为线性变换矩阵，$t-1$ 为上轮状态，GRU 单元输出状态更新结果 [8-9]。

GRU 隐状态维度 512，历史序列窗口长度 3 轮，融合后状态表示维度为 1 024，用于建模语义依存与状态演化路径。状态转移机制引入注意力门控，对前状态进行加权调制，Query 由当前语义向量生成，Key 与 Value 为历史状态残差，缩放点积计算权重后经 Softmax 归一化，完成状态连续性控制。推理部署于 TensorRT 环境，支持 FP16 精度，平均推理延迟 12 ms。

状态缓存采用 Redis 7.0 集群，内存结构基于哈希表，每条状态记录平均 1.2 kB，生命周期 600 s，压缩方式为 LZ4 级别 3。Redis 实例配置 8 核 CPU 与 32 GB 内存，单节点支持 35 000 ops/s，保障高并发状态写入性能。状态访问接口采用 gRPC 协议，支持查询、更新与归档三类操作，最大并发 QPS 达 1 800，平均响应延迟不超过 50 ms。接口采用 Protocol Buffers 编码，索引字段包括用户 ID、轮次编号与会话 ID。状态更新逻辑由响应模块控制，确保语义、状态与响应模块之间调用链的一致性与数据完整性。

3.3 知识融合模块

知识融合模块统一建模结构化图谱与非结构化语料，构建共享语义空间。结构化知识源包括 PostgreSQL 14 与 Neo4j 5.7，图谱规模约 85 万实体、280 万关系，划分 8 类实体、18 类关系，嵌入模型采用 TransE，维度 256，训练轮数 1 000，优化器为 Adam。非结构化数据由文档、问答与操作日志构成，编码方式为（dense passage retriever, DPR）双塔结构，编码器基于 BERT-base，输出维度 768，序列长度 512，向量索引使用 FAISS（IVF+PQ），压缩比 16:1。检索调度采用结构化子图查询与语义向量检索并行执行。结构化通道通过 Gremlin 完成最多 3 跳路径遍历，非结构化通道从 FAISS 中检索 Top-5 高相似文本，平均延迟控制在 9 ms 内。多通道结果通过加权线性函数融合：

$$\mathrm{score}_i = \alpha \cdot S_i^{\mathrm{struct}} + (1-\alpha) \cdot S_i^{\mathrm{text}} \tag{3}$$

式中：S_i^{struct} 为结构化语义匹配得分，S_i^{text} 为文本相似度得分，α 为通道权重参数，经验设定为 0.6。

推理模块使用 GraphSAGE 三层图神经网络，邻居采样数 10，输出维度 128，结合路径规则与状态控制变量完成推理 [10]。模型部署于 NVIDIA RTX 6000 Ada GPU（48 GB 显存），平均时延 21 ms，输出结构化知识片段含路径与置信度。服务接口支持 gRPC 与 RESTful 协议，部署于 Kubernetes 容器，单实例配置 4 核 CPU、16 GB 内存，吞吐达 1 000 QPS，P95 响应时延不超 45 ms，支持多轮上下文注入与推理请求调度。

3.4 响应决策模块

（1）动作规划：响应决策基于意图、状态向量与知识表示联合建模，构建动作选择机制。动作空间包括 API 调用、知识回复、澄清请求与结束对话，使用三层前馈神经网络分类器，输入维度 1 536（语义 768、状态 512、知识 256），隐藏层 512，激活函数 ReLU，Softmax 输出。训练采用交叉熵损失，优化器为 Adam，学习率 1×10^{-4}，批次 64，训练 20 轮。

（2）响应生成：响应生成结合模板与语言模型双路径。模板库涵盖 46 类意图动作组合，槽位填充结合规则与上下文解析。自由文本生成使用微调 GPT-2（117×10^6 参数），最大生成 64 token，温度 0.8，Top-k=50。部署于 NVIDIA A10（24 GB 显存），平均生成时延 23 ms，输出经敏感词识别与规整处理。

（3）多轮控制：多轮控制采用状态转移图机制，节点为对话阶段，边为转移条件。控制器通过规则引擎判断动作、槽位与异常标记执行路径选择，采用堆栈结构支持子任务回溯，最大深度 6。规则判断延迟＜ 5 ms，任务流程具备显示入口与退出逻辑。

（4）接口封装：模块封装为微服务，支持 RESTful 与 gRPC 协议，接口包含动作请求、响应生成、轮次控制与终止指令。部署于 Kubernetes 集群，单 Pod 配置 2 核 CPU、8 GB 内存，支持自动伸缩，峰值并发 800 QPS，接口响应延迟＜ 50 ms。响应结构 JSON 化输出，含类型、内容、指令与置信分数。

4 实验结果与分析

4.1 实验设置

实验环境构建于高性能计算节点，CPU 为 Intel Xeon Gold 6338（32 核，2.0 GHz），GPU 为 NVIDIA A100 Tensor Core（40 GB HBM2 显存），系统内存 256 GB，操作系统为 Ubuntu 22.04，深度学习框架使用 PyTorch 2.0，推理引擎为 ONNX Runtime 1.16。实验使用 SMP2020-ECDT 意图识别数据集、MultiWOZ 2.2 对话状态追踪数据集及自建企业知识问答语料，知识库条目总数达到 120 000，覆盖 FAQ 类问句、操作文档与服务流程，文本总字数约 860 万。数据预处理包含去重、分句、分词、命名实体标注与槽位填充等步骤，所有文本统一编码为 UTF-8，最大输入序列长度为 128 token。

训练集与测试集比例设定为 8:2，模型训练过程中启用 Early Stopping 策略，最大训练轮数为 30，评估指标包括意图识别准确率、槽位抽取 F_1 值、对话状态准确率、知识召回率与响应生成 BLEU 值。接口性能测试采用 Apache JMeter 5.6 进行压力模拟，最大并发线程设置为 1 000，采样间隔 200 ms，压测时长 600 s。

4.2 结果分析

系统性能评估围绕语义理解、对话状态保持、知识检索及响应生成四个子任务进行，采用多个主流指标进行量化对比。实验环境基于 NVIDIA A100（40 GB 显存）与 Intel Xeon Gold 6338（32 核，2.0 GHz）异构计算架构，推理加速由 ONNX Runtime 与 TensorRT 共同完成。为确保公平性，各项指标与行业常用基线模型对比，基线采用未融合状态感知与多源知识检索机制的对话系统构型。测试集统一处理流程为 token 化、向量化、状态注入与多轮恢复。各项实验平均重复执行 5 次取均值，波动控制在 ±1.2% 以内。各模块核心任务及指标表现如表 1 所示。

由表 1 可见，各模块在准确性与效率维度均较基线系统获得显著提升。语义处理部分在复杂意图识别与嵌套槽位抽取任务中保持稳定输出；状态建模机制有效支撑上下文回溯与轮次控制；知识融合策略增强了任务相关性召回能力；响应模块在语言流畅度与生成一致性方面达到较优平衡。接口指标验证了系统在高并发场景下的部署可行性。整体性能表现验证了模块解耦结构与多通路协同机制的工程适应性。

表 1 系统各模块核心性能指标对比

项目	意图识别	槽位抽取	对话状态跟踪	知识召回	响应生成	系统响应效率（接口）	系统吞吐能力（接口）
指标名称	准确率（Accuracy）	F_1 值	状态准确率	Top-1 命中率	BLEU-2	P99 延迟 /ms	并发 QPS
本系统得分	0.926	0.893	0.912	0.782	0.667	47	980
对比基线得分	0.881	0.841	0.857	0.698	0.612	65	730
相对提升	5.10%	6.20%	6.40%	12.00%	9.00%	−27.7%	34.20%

5 结论

本文基于人工智能语义分析的客户服务对话系统围绕语义理解、状态建模、知识融合与响应决策四个关键环节构建完成，形成具备多轮语义保持、动态知识检索与生成式响应能力的模块化系统架构。各模块实现部署解耦，保障系统在高并发环境下的响应稳定性与处理连续性。实验结果显示，系统在意图识别准确率、槽位抽取 F_1 值、状态跟踪精度与知识命中率等核心指标上优于对比系统，接口性能在响应延迟与并发能力方面达到部署要求。系统具备良好扩展性，支持企业级知识库接入与任务流程拓展，适配复杂业务语境下的语义对话调度需求。未来将进一步聚焦跨轮语义迁移与多模态知识增强，提升系统在非结构化场景下的泛化能力与响应一致性。

参考文献：

[1] 李啸天，倪郑威．基于本体的对话状态跟踪及其知识蒸馏方法 [J]. 计算机应用与软件，2025,42(8):204-212.

[2] 杜维，朱晓瑛，许方敏，等．基于模态敏感注意力机制的多模态对话模型及应用 [J]. 计算机应用研究，2025,42(9):2590-2598.

[3] 蔡辰，郭文智，孙晓宁．人智交互情境下对话式系统的适老化设计研究进展 [J]. 现代情报，2025,45(8):146-162.

[4] 熊文剑，林炳，王娟，等．云业务 AI 对话系统模型调优方法探讨 [J]. 电信工程技术与标准化，2025,38(2):78-82.

[5] 蒋玉茹，李梦媛，陶宇阳，等．基于逻辑推理和多任务融合的认知刺激对话生成 [J]. 山西大学学报 (自然科学版), 2025, 48(3): 516-526.

[6] 张松鸣，王帅博，陈钰枫，等．一种基于级联架构与多模型融合的知识型对话系统 [J]. 中文信息学报，2024,38(7):95-105.

[7] 尹娴，冯艳红，叶仕根．基于 ChatGLM 的水生动物疾病诊断智能对话系统的优化研究 [J]. 现代电子技术，2024, 47(14): 177-181.

[8] 吕学强，张剑，穆天杨，等．嵌入知识语义的医疗领域对话系统 [J]. 计算机工程与设计，2023, 44(12): 3794-3799.

[9] 潘丽莎．基于 AI 人工智能的学前教育机器人对话系统研究 [J]. 自动化与仪器仪表，2023(5): 245-248.

[10] 邹强珍．基于改进生成对抗网络的翻译机器人智能对话系统研究 [J]. 自动化与仪器仪表，2025(9):156-160.

【作者简介】

兰一杰（1992—），男，广西壮族自治区人，本科，软件工程中级职称，研究方向：软件开发。

（收稿日期：2025-12-11 修回日期：2026-07-16）