基于大语言模型命名实体识别的AI智能问答优化

AI-Enhanced Q&A Optimization via LLM-Based Named Entity Recognition

施志雄，段该甲，马龙轩，吴　婕（中国联通广东分公司，广东 广州 510627）
Shi Zhixiong，Duan Gaijia，Ma Longxuan，Wu Jie（China Unicom Guangdong Branch，Guangzhou 510627，China）

摘　要：
为优化AI问答效果，提出基于大语言模型命名实体识别的优化方法。首先，通过在多种分割方式中选取最优方案，结合词语划分概率判断结果，对语料库文本进行分词。其次，在预训练的BERT模型顶部添加线性层，并通过标注数据对预测实体类别进行微调，将预测的同类标签词组合得到命名实体。最后，通过上下文构建整合用户输入与识别结果，将整合结果输入模型生成回答，并结合用户反馈优化输出。结果表明，所提方法生成结果与参考文本之间的语义相似度较高，具备较为理想的问答效果。

关键词：
大语言模型；BERT；命名实体识别；智能问答；分词
doi：10.12045/j.issn.1007-3043.2025.03.015
文章编号：1007-3043（2025）03-0080-05
中图分类号：TN915
文献标识码：A
开放科学（资源服务）标识码（OSID）：

Abstract：
To optimize the effectiveness of AI question answering，an optimization method based on large language model named entity recognition is proposed. Firstly，by selecting the optimal segmentation method from multiple options and combining it with the probability judgment results of word segmentation，the corpus text is segmented. Secondly，a linear layer is added at the top of the pre trained BERT model，and the predicted entity categories are fine tuned through annotated data. It combines the predicted same class label words to obtain named entities. Finally，it constructs and integrates user input and recognition results through context，and inputs the integrated results into the model to generate answers，and optimizes the output based on user feedback. The results indicate that the proposed method has a high semantic similarity between the generated results and the reference text，and has a relatively ideal question answering effect.

Keywords：
Large language modeling；BERT；Named entity recognition；Intelligent Q&A；Word segmentation

引用格式：施志雄，段该甲，马龙轩，等. 基于大语言模型命名实体识别的AI智能问答优化[J]. 邮电设计技术，2025（3）：80-84.

0 引言

命名实体识别技术通过从文本中抽取出具有实际含义的语义实体，从而有效理解文本的实际含义。这一技术不仅能够有效处理复杂多变的语言文本问题，还可以有效捕获文本之间的依赖关系[1]。然而，随着文本分析技术应用场景的不断衍生，传统的命名实体识别方法已难以满足其高精度、高效率的需求，基于大语言模型的命名实体识别技术应运而生，成为当前研究的热点。在命名实体识别任务中，基于大语言模型的方法能够有效捕捉文本数据之间的特征表示，可通过微调的方式在特定数据集上实现高精度识别，有效解决了传统方法依赖人工特征和规则匹配带来的局限性。

目前，智能问答与优化技术已取得了一系列重要成果。例如，文献[2]探讨了使用检索增强生成技术、大模型微调与闭环知识图谱体系来提升政企营销知识智能问答的精度，可提高至92.36%，并通过vLLM加速、数据安全、模块化架构等技术优化系统性能与安全性，促进大模型在企业中的实际应用。但是该系统

收稿日期：2025-02-16

高度依赖高质量的训练数据和知识图谱的构建，需要定期更新这些数据以保持系统的准确性和时效性。文献[3]设计了一种基于深度学习语义匹配（利用Bert模型和Faiss向量搜索）的FAQ问答系统，旨在快速搭建特定领域的问答系统，减少人工依赖，实现高效语义匹配和秒级查询响应。系统需要能够处理大规模并发查询，这对系统的扩展性和性能提出了更高的要求。文献[4]构建了中医药循证指南知识图谱，并探索了以其为知识库搭建智能问答系统，旨在增强临床决策支持，同时提供中医药领域智能化信息服务的新思路和方法。但是，构建高质量的中医药循证指南知识图谱需要专家知识和大量数据，且过程复杂。此外，随着新研究成果的出现，知识图谱需要不断更新。文献[5]提出COBERT系统，利用检索器与阅读器双算法，通过搜索冠状病毒开放研究数据集挑战赛（CORD-19）的文献，回答复杂查询，以提供COVID-19最新研究成果的精确信息，辅助决策制定。用户可能提出复杂或模糊的查询，这要求系统能够准确理解用户意图并返回相关信息。然而，这在实际应用中可能是一个挑战。

本文选择BERT作为命名实体识别的基础模型，通过对其进行微调处理，并基于识别出的命名实体，生成准确、相关的回答。

1 基于大语言模型命名实体识别的AI智能问答优化

1.1 语料库分词处理

由于命名实体通常是由多个单词或词组组成的，为了保证BERT模型能够有效捕获实体之间的关联和上下文信息，本文首先对语料中的长文本数据进行分词预处理，其具体流程如图1所示。

对于文本分词问题，假设输入的未分词字符串为$C=[c_1,c_2,\cdots,c_n]^{\mathrm{T}}$，输出的字符串为$S=[w_1,w_2,\cdots,w_m]^{\mathrm{T}}$，其中$n$和$m$分别表示处理前后的字符串数量[6]。对于给定的待处理字符串C，存在多种分割方式S，因此可以在S中找到最优的分割方案，具如式（1）所示。

$$\hat{S}=\arg\max_S P(S|C)=\arg\max_S \frac{P(C|S)P(S)}{P(C)} \quad (1)$$

其中，$P(C)$和$P(S)$分别为语料库中文本数据输入与输出字符串可能出现的频次概率，$P(C|S)$和$P(S|C)$分别表示当输入的字符串为C或输出的分词结果为S时，另一方为S或C的概率。对于语料库中给定的一

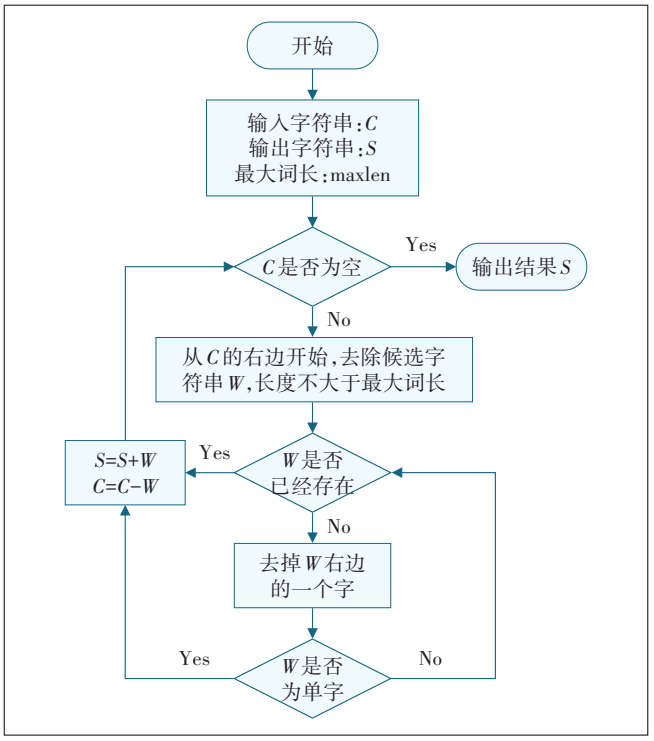

图1　语料库分词流程

段文本，可以有多种分割方案$S_1,S_2,\cdots$，因此可以计算出对应的条件概率$P(S_1|C),P(S_2|C),\cdots$[7]。为了方便讨论，本文假设分割后得到的每个字段出现的概率不受上下文相关性的影响，此时可以计算出每个分割方案对应的分词串S出现的概率，具体如式（2）所示。

$$P(X)=P\left(w_1,w_2,\cdots,w_m\right)\approx P(w_1)\times P(w_2)\times\cdots\times P(w_m) \quad (2)$$

为了衡量分词处理的结果，在完成分词后计算每个词语划分得当的概率。假设N表示语料库中文本元素的总量，则具体概率$P(w_i)$的计算如式（3）所示。

$$P(w_i)=\frac{n_{w_i}}{N} \quad (3)$$

其中，n_{w_i}表示特定词语w_i在语料库中出现的频次。通过对$P(w_i)$的值设定一个终止阈值，当遍历了输入字符串后，对所有的分词结果进行检验，若分词结果满足终止阈值要求，即可完成分词处理。

1.2 基于大语言模型命名实体识别的命名实体类别预测

智能问答系统往往需要处理来自不同领域的问题，然而不同领域的命名实体类别和模式可能存在差异[8]。用户在进行提问时通常会围绕命名实体进行展开，因此为了准确理解用户提问的意图和关注点，本文选择BERT作为命名实体识别的基础模型，通过对

其进行微调处理，预测出命名实体的类别[9]，并引入了检索增强生成技术，结合外部权威知识库，以增强模型在复杂语境下的命名实体识别能力。

本文所设计的大语言模型的起点为一个预先在广泛通用文本上训练好的BERT模型，随后收集并标注跨领域的文本数据并对这些数据进行分词预处理。这些文本数据包含了丰富的命名实体实例及其对应的类别标签[10-11]。同时，在BERT模型顶部添加一个线性层，以输出每个token的命名实体类别概率。随后，使用标注数据对模型进行微调，并通过反向传播优化损失函数，使模型能够准确预测不同领域的命名实体类别[12]。对此，本文利用交叉熵损失函数作为优化工具，以优化模型在微调过程中的性能表现，具体如式(4)所示。

$$L = -\sum_{i=1}^{N}\sum_{c=1}^{C} y_{ic} \lg(p_{ic}) \tag{4}$$

其中，y_{ic}表示真实标签的独热编码形式，p_{ic}表示模型预测得到的概率分布。本文引入检索增强生成技术，将外部权威知识库(如专业数据库、百科全书等)中的文档进行索引，以便快速检索相关信息[13]。模型利用查询语句，在索引体系中检索出与之相关的文档片段，随后依据预先设定的相似度计算方式对这些片段进行排序处理。相似度$\mathrm{sim}(q,d_i)$的计算公式如式(5)所示。

$$\mathrm{sim}(q,d_i) = \frac{q \cdot d_i}{\|q\| \|d_i\|} \tag{5}$$

其中，q表示查询向量，d_i表示第i个文档的片段的向量表示。通过式(5)可以计算出查询文档与文档片段之间的相似度，从而实现检索。

通过上述步骤即可完成对预命名实体类别预测处理。

1.3 AI智能问答生成与优化

借助微调后的大语言模型，能够依据上下文生成连贯且准确的回答。在此基础上，建立用户反馈机制，持续收集用户反馈，对大语言模型进行优化。

在完成上述命名实体类型预测后，模型会输出每个词对应的实体类型标签，根据实体类型标签可以将相同类型标签的词进行组合，从而得到命名实体[14]。为了大语言模型能够理解并生成相关回答，需要将用户输入的文本信息、命名实体识别结果以及外部的知识库信息整合为一个统一的表示形式。首先从用户输入的自然语言文本中抽取关键信息，假设f_{extract}表示信息提取函数，UserInput表示用户输入的原始文本，则信息抽取表达式如式(6)所示。

$$E = f_{\mathrm{extract}}(\mathrm{UserInput}) \tag{6}$$

其中，E表示提取的关键信息集合，而信息提取函数f_{extract}可以通过NLP技术现有的编程算法实现。为了更准确地理解问答句中的实体及其上下文关系，本文采用BiGRU网络对用户输入的问题文本与答案文本的语义信息进行编码处理，具体如式(7)所示。

$$H = \mathrm{BiGRU}(E) \tag{7}$$

其中，H表示BiGRU层的输出。假设H_q和H_a分别表示问题和答案的文本向量表示，则BiGRU网络的具体编码示意如图2所示。

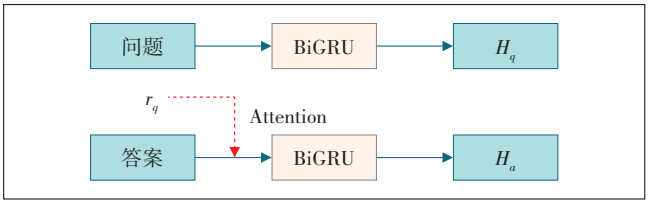

图2　BiGRU网络编码示意

其中，r_q表示原始问题的文本向量。通过上述操作，可以获得上下文信息，然后将命名实体以及上下文表示作为输入，并将其输入到大语言模型中，从而生成连贯回答[15]。回答生成的过程可以视为一个将命名实体、上下文表示等输入映射到回答文本的函数，具体如式(8)所示。

$$\mathrm{Text}_{\mathrm{answer}} = F_{\mathrm{LLM}}\left(\mathrm{Tag}_{\mathrm{entity}}, H_q, \mathrm{UserInput}\right) \tag{8}$$

其中，$\mathrm{Text}_{\mathrm{answer}}$表示生成的回答文本，$F_{\mathrm{LLM}}$表示大语言模型函数，$\mathrm{Tag}_{\mathrm{entity}}$表示识别出的实体和对应的标签序列。在生成对应的回答后，为了改进模型的输出质量，构建用户反馈机制(见图3)。

通过用户反馈机制实现AI智能问答的答案生成与优化。根据预测的实体类型标签，将相同类型标签的词进行组合，得到命名实体。通过上下文构建操作，将用户输入与实体识别结果进行整合，并输入到大数据模型中，生成对应的回答，最后结合用户反馈机制对模型的输出质量进行优化。至此，基于大语言模型命名实体识别的AI智能问答优化方法设计完成。

2 实验论证

为了验证本研究所提出的依托大语言模型进行命名实体识别(NER)增强的AI智能问答优化策略相较于传统智能问答技术，在提升实际问答性能方面的

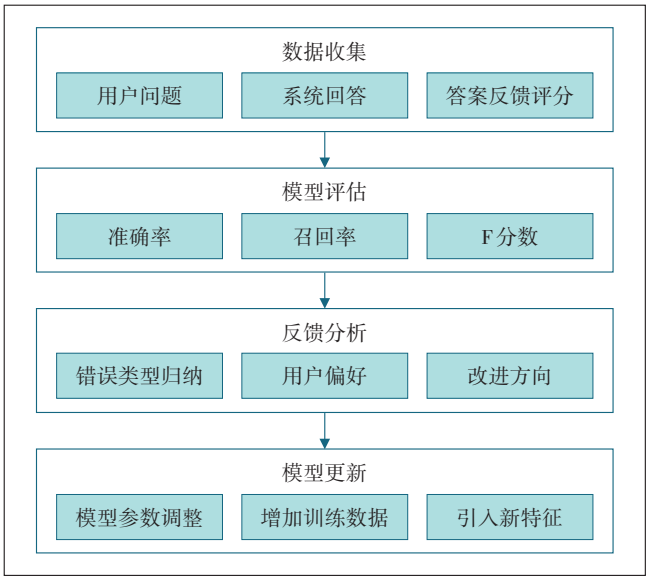

图3　用户反馈机制

优越性，本研究在理论框架构建完毕后，引入了2组基准对照实验：一组是遵循混合架构原理的智能问答技术，另一组则是依托深度学习算法的智能问答系统，进一步进行了一系列实验。该实验旨在系统性地评估和优化策略的实际应用效果，通过量化分析与质性评估相结合的方式，深入剖析该方法在问答准确性等方面的表现，从而科学论证其在实际问答场景中的优势与潜力。

2.1 实验对象

本次实验收集了不同来源的用户问题数据，涵盖科技、教育、医疗、娱乐等多个领域，形成一个多样化的统一用户问题集。该数据集包括常见问题、复杂查询、模糊表述以及潜在的歧义问题，以全面测试不同方法的应对能力。由数据集构建出的语料库分为问答文本以及回答文本，具体形式如图4所示。

文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H)

您好
打个广告可以免费的吗？
成本大概需要多少费用呢？
你们是做网络这块的吧？
我们公司网站已经做好了。
是按那个点击竞价是吧？
那无限的点击不就可以扣钱了吗
但是一直都不起效果
点击扣费我们是不考虑的。
具体情况我还不太清楚。
帮您转市场部，记一下这个号码。
谢谢！
去年就做了。
这个不是我们负责的
做推广这块业务的啊
点击费太贵小公司承受不起
信息流是什么意思啊
推广是怎么做起来的
再见

(a)问题语料库示例

文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H)

您考虑做百度推广吗
不好意思我们现在没有免费试用。
每年八千四，服务费两千四，充值六千。
是的，我们是做商业推广的。
对的就是竞价排名。
您可以设置超过三次点击就无效。
好的，后面有需要的话再联系。
可能您设置的关键词与业务不相符
不好意思打扰您了。
好的，谢谢您我都记下来了。
不客气！
建议您做网站后也可以做商业推广-
那您考虑再和我们合作吗？
您可以告诉我负责人的联系方式吗
是的我们是做商业推广的
我们现在已经是最优惠的价格了
信息流根据用户搜索习惯进行推荐
只要搜索关键词就能在网络上找到您
再见祝您生活愉快

(b)应答语料库示例

图4　语料库文件存储形式

在采用本文的算法对问题语料库中的问题进行智能回答时，需要对其进行分词处理。对此，本文设定词语划分得当的概率$P(w_i)$对应的终止阈值为0.75，当完成词语划分后，若$P(w_i)$的值高于0.75则可停止迭代，并以空格作为分隔符，将分词后的语料文件单独存储在一个语料库中，方便大语言模型直接对其进行调取。分词效果如图5所示。

基于问题文本分词结果，通过对预训练后的BERT模型进行微调，并结合命名实体识别结果以及上下文表示，生成对应的问答。待实验完成后，对每种方法所生成的回答进行记录，并与应答语料库中存

你这个什么价格啊？
你这年收费怎么收？
你做一个网站多少钱？
年费多少钱？
年费要多少？
您充值连这个总共多少钱？
您说什么价格？
您说怎么收费？

(a)原始语料文件

你 这个 什么 价格 啊
你 这 年 收费 怎么 收
你 做一个 网站 多少 钱
年费 多少 钱
年费要 多少
您 充值 连 这个 总共 多少 钱
您 说 什么 价格
您 说 怎么 收费

(b)分词处理结果

图5　语料分词示例

储的已有问答进行对比，从而进行实验对比分析。

2.2 问答效果对比

采用本文方法的用户问题智能回答结果如表1所

表1　本文方法的智能回答结果

用户问题示例	应答语料示例	本文方法智能回答结果	准确率	召回率
1. 明天天气如何？	明天晴天，温度20-25℃。	明天天气预报为晴天，温度范围在20到25摄氏度之间。	1.00	1.00
2. 明天股市开盘时间是什么时候？	明天股市开盘时间通常是上午9:30。	根据常规安排，明天股市的开盘时间预计为上午9:30 AM。	1.00	1.00
3. 我下周要去北京出差，那里的天气怎么样？	下周北京天气多变，周一到周三多云，周四到周五有小雨，周末转晴。	您下周在北京的天气预报如下：周一到周三多云，周四至周五将有小雨，周末天气转晴。	0.82	0.87
4. 明天的飞机延误了吗？	目前无法确定明天的飞机是否延误，建议您提前查询航班动态或联系航空公司。	目前暂无法直接确定明天航班的延误情况，建议您提前查询航班动态或联系航空公司以获取最新信息。	0.75	0.92

示。从表1可以看出，本文的方法能够有效分析用户的提问意图并生成对应的回答。

为提高实验结果的可靠性，本次实验以不同问答方法得到的答案与语料库中参考答案的语义相似度作为对比指标，对方法的实际问答效果进行衡量，具体对比结果如图6所示。从图6可以看出，在针对测试文本进行智能回答时，不同算法所生成的答案与参考文本之间的语义相似度均有所不同。通过数值上的对比可以看出，本文所提出的智能问答优化方法能够有效分析用户文本的意图，所生成的答案与语料库中的参考答案之间的语义相似度更高，问答效果明显优于其他2种常规的问答方法。

3 结束语

从研究意义上看，本研究不仅促进了NLP领域与大语言模型技术的深度融合，还进一步拓宽了智能问答系统的应用场景。通过精准识别文本中的命名实体，智能问答系统能够更准确地理解用户意图，从而提供更加个性化、针对性地回答，这对于提升用户体验、增强系统交互性具有不可估量的价值。

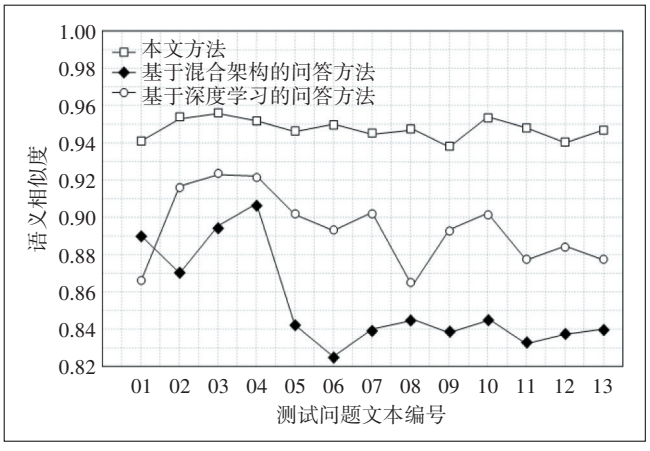

图6　不同方法的语义相似度对比结果

参考文献：

[1] 姜雨娇，黄铝文，荚子萌. 基于IMGRU-Seq2seq的自动问答方法研究[J]. 计算机应用与软件，2024，41(6)：215-222，256.
[2] 陶晓英. 基于混合架构的大语言模型智能问答系统研究[J]. 邮电设计技术，2024(5)：48-55.
[3] 武凌，黄淑芹，陈劲松，等. 基于深度学习语义匹配的通用智能问答系统的设计与实现[J]. 安庆师范大学学报(自然科学版)，2024，30(2)：84-89.
[4] 崔唐明，孙美玲，孙华君，等. 基于PICO模型的中医药循证指南知识图谱构建与智能问答系统研究[J]. 中国数字医学，2024，19(5)：20-27.
[5] ALZUBI J A，JAIN R，SINGH A，et al. COBERT：COVID-19 question answering system using BERT[J]. Arabian Journal for Science and Engineering，2021，48(8)：11003-11013.
[6] 王喆，陆俊燃，杨栋梁，等. 融合GPT和知识图谱的洪涝应急决策智能问答系统研究[J]. 中国安全生产科学技术，2024，20(4)：5-11.
[7] 丁志坤，李金泽，刘明辉. 基于大语言模型的BIM正向设计问答系统研究[J]. 土木工程与管理学报，2024，41(1)：1-7，12.
[8] 代发扬，符海东，高峰，等. 基于多角度交叉注意力机制的知识库问答方法[J]. 计算机应用与软件，2023，40(12)：33-40.
[9] 王婷，王娜，崔运鹏，等. 基于人工智能大模型技术的果蔬农技知识智能问答系统[J]. 智慧农业(中英文)，2023，5(4)：105-116.
[10] 张春红，杜龙飞，朱新宁，等. 基于大语言模型的教育问答系统研究[J]. 北京邮电大学学报(社会科学版)，2023，25(6)：79-88.
[11] 李志东，罗琪彬，乔思龙. 基于句粒度提示的大语言模型时序知识问答方法[J]. 网络安全与数据治理，2023，42(12)：7-13.
[12] 郑胜男，柳圣，鞠文慧，等. 基于自注意机制的中文医药命名实体识别算法研究[J]. 南京工程学院学报(自然科学版)，2023，21(4)：37-40.
[13] 杨波，孙晓虎，党佳怡，等. 面向医疗问答系统的大语言模型命名实体识别方法[J]. 计算机科学与探索，2023，17(10)：2389-2402.
[14] 针钰，马晓宁. 自注意力机制下复杂文本实体关系抽取方法[J]. 计算机仿真，2024，41(4)：522-526.
[15] 侯俊丞，杜浸，何之栋，等. 基于消防领域的知识图谱智能问答的研究[J]. 电信快报，2023(4)：24-33.

作者简介：
施志雄，工程师，硕士，主要从事人工智能技术研究及管理工作；段该甲，工程师，硕士，主要从事人工智能领域的前沿技术研究与创新应用工作；马龙轩，工程师，硕士，主要从事人工智能技术研究及应用工作；吴婕，工程师，学士，主要从事企业智能化应用技术研究及项目管理工作。