人工智能

中文检索增强生成在政务热线领域的创新与应用

张黎，陈国润，郑荣，代晓菊，李铮，刘长东
（上海理想信息产业（集团）有限公司，上海 201315）

摘　要：探讨了自然语言处理（NLP）技术在政务热线领域的应用，特别是基于检索的生成（RAG）模型的创新使用。传统的政务热线面临响应速度慢、服务质量不稳定等问题，而NLP技术的应用能够实现服务的自动化和智能化。详细介绍了RAG技术在智能客服、智能问答和智能处理系统中的应用，并通过上海市民服务热线12345的案例，展示了如何利用检索增强生成技术和文本转语义向量提升知识库管理的效率和服务质量。实验评估表明，新方法在准确性上较原RAG结构提升了48%，证明了其在政务热线服务中的实际效用。未来，将继续优化技术，推动政务热线服务的智能化发展。

关键词：大模型；智能客服；检索增强生成
中图分类号：TP393
文献标志码：A
doi: 10.11959/j.issn.1000−0801.2025063

0　引言

政务热线是政府与公众之间沟通的重要桥梁，它能够帮助公众解决各种问题，提高政府的服务质量和效率。然而，传统的政务热线存在着一些问题，如响应速度慢、服务质量不稳定、处理问题效率低等。随着人工智能技术的不断发展，自然语言处理（NLP）技术在政务热线领域的应用逐渐受到关注。

NLP技术是一种能够理解和解释人类语言的技术，它通过对语言的分析和处理，可以帮助政务热线实现自动化、智能化的服务。其中，基于RAG的NLP技术在政务热线领域的应用具有重要的创新意义。

RAG是一种基于检索的生成模型，它结合了检索和生成两种技术的优点，能够在保证生成质量的同时，提高生成速度和效率。在政务热线领域，RAG技术可以应用于智能客服、智能问答、智能处理等多个方面，从而实现政务热线的创新与发展。

首先，RAG技术可以应用于智能客服系统。传统的政务热线客服系统需要大量的人工客服，而且服务质量受到人员素质、经验等因素的影响，存在很大的不稳定性和不可预测性。而基于RAG技术的智能客服系统可以通过对历史对话数据的分析和学习，实现对用户问题的自动理解和回答，从而提高服务质量和效率。

其次，RAG技术可以应用于智能问答系统。传统的政务热线问答系统通常是基于规则匹配的，对于一些复杂的问题，系统往往无法给出准确的答案。而基于RAG技术的智能问答系统可以通过对大量历史问答数据的分析和学习，实现对用户问题的自动理解和回答，从而提高问答的准确性和效率。

收稿日期：2024−11−09；**修回日期：**2025−02−06

最后，RAG技术可以应用于智能处理系统。传统的政务热线处理系统通常需要人工进行分类、转接、处理等操作，效率低下且容易出错。而基于RAG技术的智能处理系统可以通过对历史处理数据的分析和学习，实现对用户问题的自动分类和处理，从而提高处理效率和准确性。

综上所述，基于RAG的NLP技术在政务热线领域的应用具有重要的创新意义。通过实现智能客服、智能问答和智能处理等功能，RAG技术能够提高政务热线服务的质量和效率，提升公众的满意度和信任度，从而推动政务热线向智能化、高效化、人性化的方向发展。

1 技术架构

1.1 中文检索增强生成技术

检索增强生成技术（retrieval-augmented generation，RAG）是一种结合了检索和生成技术的自然语言处理方法，专门针对中文语境下的文本生成任务。这种技术的核心思想是在生成文本的过程中利用检索到的相关信息来增强生成模型的能力，以提高生成文本的质量、相关性和多样性。

在检索增强生成技术中，通常包括以下几个关键步骤。

（1）检索：根据输入的查询或者上下文信息，从大规模的中文语料库中检索出与当前生成任务相关的信息。这些信息可以是文档、段落、句子或者词组，它们作为生成的参考材料。

（2）排序：将搜索出来的片段根据与问题的相关性排序。

（3）生成：利用检索并排序的相关信息，结合语言生成模型（如ChatGLM、通义千问等）来生成新的文本。生成模型在训练时会学习如何利用检索到的信息来提高生成文本的质量。

（4）整合：将检索到的信息与生成模型生成的文本进行整合，调用大模型能力进行总结回答。

中文检索增强生成技术的优势如下。

（1）提高准确性：通过检索相关信息，生成模型可以获得更多的上下文信息，从而提高生成文本的准确性，避免大模型幻觉问题。

（2）增强灵活性：随着知识的更新，即时更新检索片段，避免重复训练模型，有助于灵活应对知识库更新。

（3）提升相关性：检索到的信息与生成任务直接相关，可以溯源回查，给予充分的解释，确保生成的文本有事实依据。

1.2 文本转语义向量

在语义学领域，文本转语义向量是指将文本转换为语义向量的过程，语义向量是一种表示文本特征的向量，可以用来表示文本的语义信息。将文本转换为语义向量可以帮助我们更好地理解和分析文本数据，这也是所有处理文本的重要且必要的第一步工作。

现在主流的文本转语义向量方法有如下几种。

（1）基于规则的方法。这种方法使用专家知识和规则来将文本转换为语义向量。这些规则可以基于文本的语法、词汇和句子结构等方面来确定。其中一些方法包括TF-IDF、TextRank、BM25等。

（2）基于机器学习的方法。这种方法使用机器学习算法来训练语义向量。其中一些算法包括朴素贝叶斯、支持向量机等。

（3）基于深度学习的方法。这种方法使用深度学习模型来训练语义向量。其中一些模型包括卷积神经网络（CNN）、循环神经网络（RNN）、Transformer、BERT和GPT等。

本文针对深度学习的方法，采用BGE（BAAI general embedding），这是由智源研究院发布的一款开源中英文语义向量模型。该模型在语义向量模型的领域取得了显著进展，特别是在中文语义

表征方面表现出色，在中文语义向量综合表征能力评测C-MTEB中展现了卓越的性能，其检索精度大约是OpenAI Text Embedding 002的1.4倍。

2　政务热线领域创新与应用

2.1　政务热线知识库特点

（1）业务多样性和知识管理挑战。政务热线的服务内容非常广泛，涉及经济调节、市场监管、社会管理、公共服务、生态环境保护等多个领域。因此，如何有效地汇集这些领域的知识，并将其整理为规范化、体系化的形式，以便话务员能够快速准确地向民众提供信息和服务，是政务热线知识管理面临的一大挑战。

（2）跨部门协同和知识管理及时性。政务热线的知识库管理需要内部各部门之间的紧密协作，同时也涉及与其他政府机构、公共组织以及私营部门的合作。这种跨部门协同对于确保知识的全面性和准确性至关重要，但也带来了协调沟通的复杂性。

（3）知识库与座席交互体验的易用性。政务热线的知识库必须设计得既全面又易于使用。这意味着知识库的内容应当是结构化的、更新及时的，并且能够以使用者友好的方式呈现，以便话务员能够迅速检索到相关信息，为民众提供准确的指导和帮助。

（4）知识管理的即时性。政务热线知识库是热线服务的核心支柱，其政策文件、法律法规可能会有变化，需要知识库管理及时更新，避免市民获取过期的指引。

综上所述，政务热线知识库在业务多样性、跨部门协同、易用性、底层逻辑方面具有显著特点，这些特点决定了建立政务热线领域知识库具有困难性。

2.2　政务检索增强生成知识库

以上海市民服务热线12345为例，探索检索增强技术的知识库功能和特点。

上海市民热线委办局有125个，如市交通委、市水利局、XX区人民政府等。每个委办局都有自己的知识文件，上传给市民热线，由统一的知识库管理员录入，话务员通过关键字进行搜索查找。引入检索增强生成技术后，每个委办局的知识文件都被细致地分类和标签化，确保了信息的有序和可查找性。例如，市交通委提供的知识可能包括最新的交通法规、公交线路调整信息、交通拥堵的应对策略等。市水利局的知识文件则可能包含防汛措施、水域管理法规、水利工程建设进展等。而区人民政府的知识文件则可能聚焦于地区性的政策、社区服务、文化活动等信息。话务员在接到市民的电话时，只需要简单输入问题，系统便会迅速匹配出最相关的知识条目，并提供话术建议，极大地提升了工作效率和准确性。

此外，这种技术的应用还极大地提升了市民服务热线的服务质量。话务员能够提供更加准确和及时的信息，减少了市民的等待时间，提高了解决问题的速度。同时，通过数据分析，市民服务热线还能够对服务流程进行优化，提升整体的运营效率。

总之，以上海市民服务热线12345为例，检索增强技术的应用，不仅提高了话务员的工作效率，还优化了市民的服务体验，展现了技术应用为现代服务业带来的创新和发展。

2.3　问题检索拓展

由于政策多、范围广，市面上的RAG技术虽然可以给出答案，但准确性无法保证，经测试，使用原生RAG技术，导入全部知识库后，话务员随机找问题测试，答案的准确率仅为63%，不满足上线应用标准。分析上海12345热线服务全流程，梳理数据留存情况，创新性地采用了以下方法进行检索拓展。

（1）从话务员接电话历史操作记录中提取知识库点击记录，并与其对话文本进行关联，从对

话文本中提取市民的咨询问题，从而建立真实通话中的问题与知识库的对应关系。对历史数据进行批量处理，搜集到问答库中的全部问答关系共有20 000对。

（2）重构RAG流程，首先将搜集的问答库中的问题进行Embedding，单独建立索引，放在全部知识库片段索引前面，当接收到市民问题query来的时候，先匹配问答库，如果匹配成功（相似度大于0.9），则直接将该问题对应的知识点给到下一步流程，如果没有成功，再用全部知识库片段索引进行匹配。

（3）将匹配好的知识文件或者知识库片段提供给大模型，设置prompt为“你现在是一个话务员，请根据输入的诉求内容，生成简单明确的工单描述，请务必参考诉求内容，不要生成无关的信息”，大模型会给出一个综合性的回复。

3 实验与评估

针对本文所述方法进行实验，话务员整理出了有200条问答的测试集，每条为市民咨询的问题和对应的正确答案。分别对原RAG结构和本方法给出的答案进行测试，每个问题根据回答的准确性，给出0~10分的评分，非常准确且无任何多余回答的得10分，完全不对的得0分，其他根据情况区间打分。对比测试记录见表1。

综上所述，在有200个问答的测试集上，本论文结构回答得到的分数相较于原RAG结构，提升了48%，其中大部分问题从原RAG结构基本无法解答，变为了基本正确无误。

4 结束语

本文深入探讨了基于RAG技术的自然语言处理（NLP）在政务热线领域的应用，以及如何通过技术架构的创新来解决传统政务热线面临的挑战。引入中文检索增强生成技术和文本转语义向量的方法，不仅提高了政务热线服务的准确性和效率，而且增强了服务的灵活性和相关性。特别是上海市民服务热线12345的应用案例，展示了如何通过检索增强生成知识库和问题检索拓展方法，显著提升话务员的工作效率和市民的服务体验。

实验与评估部分进一步证实了本文方法的有效性。通过对200条测试集的评分对比，本文方法在准确性上相较于原始RAG结构提升了48%，这表明了本文方法在处理政务热线问题时的显著优势。随着技术的不断进步和优化，基于RAG的NLP技术将为政务热线带来更多创新的可能性，为公众提供更高效、更智能、更人性化的服务。

在未来的工作中，本团队将继续探索并优化

表1 对比测试记录

问题编号	问题内容	原RAG结构回答得分	本论文结构回答得分
1	驾驶证丢失怎么办?	8	9
2	我要出国，哪些国家可以免签?	7	10
3	徐汇区小学转学怎么办理手续?	0	8
4	公司不发年终奖，可以仲裁吗?	0	7
5	公司原来在徐汇区，想迁到临港，怎么办理?	3	10
…	…	…	…
200	医保卡丢了，去哪里补办?	4	9
平均分		5.8	8.6

RAG技术在政务热线领域的应用，同时，也期待与更多政府部门和机构合作，共同推动政务热线服务走向智能化和现代化。通过不断进行技术创新和服务改进，致力于构建一个更加开放、透明、高效的政府服务平台，以满足公众日益增长的服务需求，并为建设智慧型政府贡献力量。

参考文献：

[1] 孙维纬. 知识检索增强的对话系统研究[D]. 济南: 山东大学, 2023.
[2] 文森, 钱力, 胡懋地, 等. 基于大语言模型的问答技术研究进展综述[J]. 数据分析与知识发现, 2024, 8(6): 16-29.
[3] 周扬, 蔡霈涵, 董振江. 大模型知识管理系统[J]. 中兴通讯技术, 2024, 30(2): 63-71.
[4] 袁飞. 基于预训练模型的问答系统关键技术研究[D]. 电子科技大学, 2021.
[5] 朱茜. 我省首批数字政府大模型场景应用清单发布[N]. 安徽日报, 2023-12-06(001).
[6] 蒋钰玲. 面向视觉问答的多模态信息增强方法研究[D]. 南京邮电大学, 2023.

[作者简介]

张黎（1987– ），男，上海理想信息产业（集团）有限公司工程师，主要从事NLP自然语言处理领域的研究和技术开发工作。

陈国润（1978– ），男，上海理想信息产业（集团）有限公司高级工程师，主要从事大数据与人工智能领域的研究和管理工作。

郑荣（1981– ），男，上海理想信息产业（集团）有限公司高级工程师，主要从事大数据与人工智能领域的研究和管理工作。

代晓菊（1990– ），女，上海理想信息产业（集团）有限公司工程师，主要从事NLP自然语言处理领域的研究和技术开发工作。

李铮（1986– ），男，上海理想信息产业（集团）有限公司工程师，主要从事NLP自然语言处理领域的研究和技术开发工作。

刘长东（1993– ），男，博士，现就职于上海理想信息产业（集团）有限公司，主要从事自然语言处理与大语言模型领域的研究与技术开发工作。