文章类型：论文|刊号（ISSN）：2972-4236(P) / 2972-4244(O)

利用深度学习进行中文问答系统的研究与实现

金嘉诚
香港浸会大学
DOI:10.12238/acair.v2i3.8631

[摘　要] 在中文问答系统中,深度学习技术发挥了至关重要的作用,通过高效处理复杂的中文语义理解,显著提升了问答系统的精确度和工作效率。然而,实现中文问答系统面临诸多挑战。为了解决这些问题,本文提出了使用预训练中文语言模型如BERT和RoBERTa、设计高效的数据收集与标注方法、应用分布式训练与模型压缩技术,以及采用上下文理解模型如注意力机制和长短时记忆网络等策略,以此提高中文问答系统的性能和用户体验。
[关键词] 深度学习；中文问答系统；语义理解；预训练模型；上下文理解
中图分类号：O572.25　文献标识码：A

Research and implementation of Chinese question answering system using deep learning

Jiacheng Jin
Hong Kong Baptist University

[Abstract] In Chinese question answering system, deep learning technology plays a crucial role. By efficiently processing complex Chinese semantic understanding, it significantly improves the accuracy and work efficiency of question answering system. However, the implementation of Chinese question answering system faces many challenges. In order to solve these problems, this paper proposes to use pre-trained Chinese language models such as BERT and RoBERTa, design efficient data collection and annotation methods, apply distributed training and model compression techniques, and adopt context understanding models such as attention mechanisms and short term memory networks to improve the performance and user experience of Chinese question answering system.
[Key words] Deep learning; Chinese question answering system; Semantic understanding; Pre-training model; Context understanding

引言

随着大规模预训练模型的发展,如GPT-3.5和GPT-4,这些模型在处理复杂的中文语义理解和生成任务上表现出色,使得中文问答系统的精度和效率显著提高。然而,研究者在实现中文问答系统时仍然面临许多挑战,如中文语义的多义性和复杂性、海量高质量数据的获取和标注难度,以及模型的训练效率和资源消耗问题。为了解决这些问题,研究人员提出了利用预训练中文语言模型、设计高效的数据收集和标注方法、使用分布式训练与模型压缩技术,以及上下文理解模型的策略。

1 深度学习在中文问答系统中的作用

1.1卷积神经网络(CNN)和长短时记忆网络(LSTM)

卷积神经网络和长短时记忆网络在处理文本分类和序列建模任务时展示了强大的能力。CNN通过卷积操作能够高效提取文本中的局部特征,如词语或短语的模式和结构,这使得CNN在捕捉文本的局部信息上表现出色。卷积核在滑动过程中可以同时处理多个位置的特征,从而提高了模型的计算效率和特征提取能力。另一方面,LSTM通过其特殊的门控机制善于捕捉序列数据中的长距离依赖关系,能够在长文本或序列数据中保持上下文信息的连贯性,从而理解并记住长时间跨度内的重要信息。

1.2预训练语言模型：BERT和RoBERTa

预训练语言模型如BERT和RoBERTa通过迁移学习应用于特定问答任务中,进一步提高了系统的效率和精度。BERT模型由Google提出,以其双向编码器结构而著称。与传统的单向语言模型不同,BERT在训练过程中同时考虑句子中每个词语的前后文信息,从而对句子内部的语义关联有更完整的认识。这种双向的训练方式使BERT能够更准确地理解复杂的语言结构和语境。RoBERTa模型是在BERT基础上的进一步优化版本。RoBERTa通过增加训练数据的规模和训练时间,改进了BERT的训练策略,并去掉了Next Sentence Prediction任务,改为更高效的训练方式。

1.3最新的大语言模型：GPT-3.5和GPT-4

最新的大语言模型如GPT-3.5和GPT-4,在处理中文文本生成和理解任务中表现出色。它们基于Transformer架构,采用了自注意力机制,能够高效处理长距离依赖关系。多层次、多头自注意力机制使得模型能够在理解和生成文本时关注句子中不同部分的关联,提高了对复杂语义的理解能力。GPT-3.5和GPT-4通过在数百亿至数千亿参数的规模上进行训练,具备了强大的语言生成和理解能力。

这些模型在预训练阶段,通过在大规模中文语料库上进行训练,学习通用的语言知识和语义表示。随后,通过在特定任务的标注数据集上进行微调,使模型能够适应具体的应用场景,如问答系统。

2 利用深度学习进行中文问答系统的实现难题

2.1中文语义理解的复杂性

中文在语法构造上与其他语言有明显区别,缺少明确的词汇边界,这增加了语义解读的复杂性。词的多义性和歧义性也加剧了模型理解过程的复杂性。例如,“银行”在不同语境下可以指金融机构或河岸,需要模型具备较强的词汇量和深度语境理解能力。中文成语、俚语及方言等表达形式的丰富性进一步加大了语义解析难度,这类语言形式通常文化依赖性较强,语义内涵暗含,使模型的理解需要更加复杂的语义分析和文化背景知识。

中文主谓宾结构及修饰语位置较为自由,致使相同语义可由多个句式表示,对模型语法及语义解析能力要求较高。要精确了解用户提出的问题,就要求模型能灵活处理各种句式及结构变化。中文的代词指代问题及零指代现象还加大了语义理解的复杂程度,它们需要模型具备高度的推理能力及对对话语境的整体理解能力才能正确解析指代关系并给出精确答案。

2.2大量中文问答数据的获取和标注难题

中文问答数据来源零散,数据质量与覆盖范围存在显著差异。为了建立全面而优质的数据集,必须从各种来源搜集大量问答,既费时又费力,且需经过严格筛选与清理以保证数据质量。数据标注工作繁重且复杂,要求专业人员按照具体标注标准精细标注数据。由于中文语言复杂,标注时常涉及许多主观判断,例如语义是否准确、上下文是否相关等,这些都需要标注人员具备高度语言能力及领域知识。

不同标注人员对同类问题的认识与标注会有差别,从而导致标注结果不统一,直接影响训练模型质量与性能。为保证数据标注质量,必须建立详尽的标注规范与严密的审核机制,以将标注错误与不一致性降到最低。

2.3深度学习模型的训练效率和资源消耗

深度学习模型在训练过程中通常需要海量高质量数据,对数据进行预处理与准备也需要大量时间与资源。在进行模型训练时,高性能计算设备如GPU和TPU是必不可少的。GPU是最常见的选择,因其强大的并行计算能力和广泛的可用性受到青睐。其中,NVIDIA的Tesla系列、GeForce系列是深度学习训练中常用的GPU。而AMD的Radeon系列和Instinct系列也被用于深度学习训练。

深度学习模型的参数规模一般较大,具体大小取决于模型的复杂度和应用场景。例如GPT-3模型拥有1750亿个参数,BERT-Large模型有3.4亿个参数。这些模型的巨大参数规模意味着在训练时需要占用大量存储空间与内存资源,造成极大的资源消耗。

2.4上下文理解的挑战

中文语言对上下文依赖性强,某个词语或短语可能在不同上下文下有不同含义,需要模型具备很强的上下文分析能力才能正确解析用户意图。例如在会话中,用户可能略去某些信息或用代词指代前述内容,模型需要具备推理能力才能补全这些信息以保证答案正确。多轮会话的语境理解更为复杂,模型既要记住前几轮会话的主要信息,又要动态调整语境以给出持续且相关的答案。

上下文理解也涉及隐含信息识别与解析问题,尤其当用户未清晰表达自己意图时,模型需借助上下文线索进行推理与判断。这类隐含信息处理增加了模型理解难度,需要模型具备较高的语义解析与推理能力。模型在处理复杂上下文信息时也需避免信息丢失与错误理解,对其设计与训练提出了较高要求。

3 利用深度学习进行中文问答系统的实现策略

3.1采用预训练的中文语言模型,如BERT和RoBERTa

通过预训练模型可以实现问答、文本分类及命名实体识别的有效处理,大大增强系统对语义的理解与运用。BERT模型以其双向编码器结构而著称。BERT采用Transformer架构中的双向编码器表示,这意味着它在训练时同时考虑了词语的前后文信息,而不是像传统的单向语言模型那样仅考虑单一方向的上下文。通过这种双向的训练方式,BERT能够更全面地理解句子内部的语义关联,对上下文信息进行全局把握,从而在语义理解任务中表现出色。

RoBERTa模型在BERT的基础上进行了多方面的优化,进一步提升了模型的性能。首先,RoBERTa通过增加训练数据的规模和训练时间,使模型能够学习到更多的语言细节。其次,RoBERTa改进了数据处理方式,例如去掉了BERT中使用的Next Sentence Prediction任务,改为更高效的训练方式。最后,RoBERTa采用了动态掩码策略,即在每次训练时重新随机选择被掩盖的词语位置,从而提高了模型的泛化能力。

3.2设计有效的中文问答数据收集和标注方法

在中文问答数据高效采集与标注方法的设计中,可考虑在数据处理流程中引入ChatGPT与Transformer模型的自注意力机制。ChatGPT作为基于Transformer框架的自然语言生成工具,拥有出色的语言解析和生成功能,有助于更深入理解用户问题及意图,从而优化问答数据的语义标注效果。Transformer模型是由Vaswani等人在2017年提出的用于自然语言处理的模型架构,具有高度的并行计算能力和强大的表达能力。

自注意力机制特别适用于问答系统中的意图识别和情感分析任务。它能够在解析用户输入时,将关注点集中在与当前任务最相关的词语和短语上,从而更准确地理解用户的真实意图。当用户提出一个复杂的问题时,Transformer模型能够通过自注意力机制识别问题中的关键信息和上下文关系,确保对问题的全面理解,并生成准确的答案。可以先使用ChatGPT对采集到的问答数据进行初步语义理解,识别出问题中的核心概念和意图。在此基础上,借助Transformer模型的自注意力机制进一步解析数据的关键信息与上下文关系。Transformer模型中的多头注意力机制能够同时从不同角度捕获信息,有利于将多维度的信息纳入标注中,如意图识别和情感分析。

3.3利用分布式训练和模型压缩技术提高训练效率

为提升深度学习模型训练效率并减少资源消耗,采用分布式训练与模型压缩技术是行之有效的策略。分布式训练将模型训练任务分散到多个计算节点中完成,显著提升训练速度与效率。分布式训练利用多机多卡并行计算能力,同时处理大规模数据与模型参数,有效缩短训练时间。常用的分布式训练框架如TensorFlow和PyTorch为开发者执行分布式训练任务提供了大量工具与接口。

模型压缩技术显著减少资源消耗及推理时间,降低参数量及计算量。常用的模型压缩技术包括模型剪枝、量化和蒸馏。模型剪枝通过删除多余参数降低模型计算复杂度,可以在减少50%甚至更多参数的情况下,保持模型性能基本不变。量化技术将模型权重从高精度浮点数转为低精度整数表示,通常将32位浮点数转换为8位整数,这样可以减少75%的存储空间和内存带宽消耗。

3.4采用上下文理解模型,如注意力机制和长短时记忆网络

上下文理解模型,如注意力机制和长短时记忆网络(LSTM),是近年来在自然语言处理领域取得显著进展的关键技术。这些模型的使用为对话系统、问答系统及其他任务带来了更深刻的语义理解与信息处理,增强了人工智能对交互式场景的性能。注意力机制的核心思想是对模型输入序列的各要素赋予不同的权重,使模型能够重点关注重要的上下文信息。

长短时记忆网络(LSTM)作为序列建模的经典技术,凭借其特有的门控策略,高效处理长距离相关数据和序列中的时序信息。LSTM能够维持长时间序列中的重要上下文信息,并在多轮对话时保证模型的语义一致性和连贯性。

4 结束语

在中文问答系统的研究和实现过程中,深度学习技术发挥了关键作用。通过使用预训练语言模型如BERT、RoBERTa及最新的GPT-3.5和GPT-4,设计高效的数据收集与标注方法,应用分布式训练与模型压缩技术,以及采用上下文理解模型如注意力机制和长短时记忆网络,研究者们显著提高了系统的精度和效率,解决了中文语义理解、数据获取与标注、训练效率和资源消耗等挑战。

[参考文献]

[1]朱洪坤.基于深度学习的中文礼貌风格迁移方法研究[D].江西师范大学,2023.

[2]李莉,奚雪峰,盛胜利,等.深度学习中文命名实体识别研究进展[J].计算机工程与应用,2023,59(24):46-69.

[3]张岩.基于深度学习与 Markov 使用模型的测试用例生成方法研究[D].北京交通大学,2023.

[4]吕钰珂.基于深度学习的医疗智能问答系统研究与应用[D].青岛科技大学,2023.

[5]陈禧琛.基于深度学习的中文医疗社区答案选择算法研究[D].广东工业大学,2022.

[6]缪鹏飞.基于检索和知识图谱结合的开放域中文问答系统设计与实现[D].重庆大学,2022.

[7]王琪.融合多源外部知识的中文问答系统答案匹配方法研究[D].北京交通大学,2022.

[8]顾雅涵.中文文本的深度学习纠错方法研究[D].南京理工大学,2021.