

# 人工智能驱动的计算机自然语言处理技术研究

王静 张雅倩

(郑州电子信息职业技术学院 郑州 451450)

**摘要** 随着 AI 技术发展,自然语言处理(NLP)在众多领域展现出了巨大的潜力,如机器翻译、语音识别、情感分析、智能客服等。文中梳理了 NLP 的基本方法,探究了深度学习在其中的作用,详细讨论了大数据与 NLP 的结合应用,并分析了预训练模型的性能及其在电子商务领域的实际应用。

**关键词:** 自然语言处理;深度学习;大数据

**中图分类号** TP399

## Research on Computer Natural Language Processing Technology Driven by Artificial Intelligence

WANG Jing and ZHANG Yaqian

(Zhengzhou Professional Technical Institute of Electronic & Information, Zhengzhou 451450, China)

**Abstract** With the development of AI technology, natural language processing (NLP) has shown great potential in many fields, such as machine translation, automatic speech recognition, sentiment analysis, intelligent customer service, etc. This paper sorts out the basic methods of NLP, explores the role of deep learning in it, discusses the combined application of big data and NLP in detail, and analyzes the performance of pre-trained models and their practical applications in the field of e-commerce.

**Key words** Natural language processing, Deep learning, Big data

## 0 引言

在自然语言处理领域应用深度学习等技术,能让计算机处理文本数据,实现语言的理解、生成和应用。语音识别技术则专注于将人类语音转换为可理解的文本形式。在自然语言处理方面,关键任务包括语义分析、情感分析、命名实体识别等<sup>[1]</sup>。

## 1 计算机自然语言处理的基本方法

### 1.1 基于规则的自然语言处理

在自然语言处理的发展早期,基于规则的方法占据着主导地位。这些方法依赖于语法和语言学规则进行任务处理,如分词、词性标注和句法解析,而这些规则大多由语言学专家制定,通过模式匹配实现简单任务。

### 1.2 统计方法在自然语言处理中的应用

通过深度分析海量的语言数据,统计方法能准确识别出语言模式的内在规律。基于概率的模型,如 N-gram 模型与隐马尔可夫模型(HMM),可通过训练数据来学习并提炼语言的深层结构,被广泛应用于语音识别、文本分类等多元化 NLP 任务中。

### 1.3 词向量模型的引入

词向量模型如 Word2Vec,以独特的视角将词汇融入向量空间。它将词语的上下文信息转换为数学向量,弥补了词袋模型在词序和语义相似度上的不足,提升了 NLP 系统的效能。

### 1.4 神经网络与深度学习方法的应用

神经网络与深度学习方法的应用,促进了自然语言处理领域的发展。RNN、LSTM 等模型能精准捕捉语言的时间依赖特性,而 Transformer 模型的出现革新了 NLP 处理方式,成为当前最热门的技术。

## 2 自然语言处理技术的应用场景

### 2.1 机器翻译

机器翻译作为自然语言处理的核心应用领域,致力于将一种语言无缝转化为另一种语言。依托于神经网络技术的翻译模型,如 Google 翻译,现已可以应对多种语言翻译任务,具备较高的准确度。LSA 可采用低维词条、文本向量代替原始的空间向量,处理大规模语料,适用于信息过滤、文本摘要、机器翻译等跨语言信息检索的生成式自然语言处理领域<sup>[2]</sup>。

**作者简介:**王静(1997—),本科,助教,研究方向为计算机技术;张雅倩(1998—),本科,助教,研究方向为计算机技术。

## 2.2 语音识别

语音识别技术能将人类的语音信号转换为文本,搭建起人机交互的桥梁。NLP则能理解和分析这些转换后的文本,被广泛应用于智能助手、电话客服等领域。近年来,融合了深度学习的语音识别系统的准确度得到提升,如Siri和Google Assistant,为用户带来了更便捷、高效的交互体验。

## 2.3 情感分析

情感分析的应用范畴极为广泛,涵盖社交媒体与电子商务中的用户反馈解析。借助NLP技术的创新,情感分析模型能洞察文本背后的情感色彩,精准判别用户的情绪倾向,如正面的赞扬、负面的批评、中立的陈述等。

## 2.4 自动文本摘要

自动文本摘要技术能从大量文本中提取关键内容,生成简短的摘要。基于神经网络的生成式摘要模型能生成自然、流畅的总结性句子,在新闻、研究报告、法律文件等需要快速获取关键信息的领域具有重要应用<sup>[3]</sup>。

## 2.5 聊天机器人

聊天机器人能借助自然语言处理技术实现高度仿真的自动对话,与人类的语言交互方式无明显差异。融入了深度学习技术的聊天机器人不仅能准确回应用户的问题,还能进行个性化的内容推荐与深入的情感互动,在客户服务、社交平台以及健康咨询等领域展现出强大的应用价值。

# 3 基于人工智能的自然语言处理技术的发展

## 3.1 语言模型的发展历程

自然语言处理是计算机AI领域的重要应用方向之一,其与计算机科学、概率统计、语言学等学科密切相关。AI作为现代计算机科学的一个重要分支,涵盖了生物学、心理学、神经科学等领域,指具有人类意识和思考能力的机器或算法,其本质是让机器模仿人类并实现与人类相仿的应答机制。自然语言处理的语言模型经历了基于统计方法到基于深度学习的变革<sup>[4]</sup>。早期的N-gram模型通过统计词语序列中词的出现概率来建模,如式(1)所示:

$$P(w_1, w_2, \dots, w_n) \approx \prod_{i=1}^n P(w_i | w_{i-n+1}, \dots, w_{i-1}) \quad (1)$$

其中, $P(w_i | w_{i-n+1}, \dots, w_{i-1})$ 是在给定前面 $n-1$ 个词的情况下词 $w_i$ 的概率。

N-gram模型可以捕捉局部依赖性,但在处理长距离依赖关系时效果有限。随后,语言模型经历了从传统统计模型向基于深度学习的模型的过渡,如BERT和GPT等预训练模型通过Transformer架构提升了语言建模能力,能有效处理更复杂的上下文和语义关系。

## 3.2 神经网络语言模型

神经网络语言模型(NNLM)能利用神经网络来学习语言的复杂模式,克服传统N-gram模型的局限性。典型的神

经网络语言模型如递归神经网络(RNN)能通过循环连接记忆先前输入的信息,如式(2)所示:

$$h_t = \sigma(W_h x_t + U_h h_{t-1}) \quad (2)$$

其中, $h_t$ 是隐藏层的状态, $x_t$ 是当前输入, $h_{t-1}$ 是前一时刻的隐藏层状态, $W_h$ 和 $U_h$ 是权重矩阵。

为处理长距离依赖问题,LSTM和GRU模型得到了应用,以避免梯度消失问题。Transformer模型能通过自注意力机制解决RNN在处理序列数据时的瓶颈,且能并行处理输入,如式(3)所示:

$$Attention(Q, K, V) = \text{soft max} \left( \frac{QK^T}{\sqrt{d_k}} \right) V \quad (3)$$

## 3.3 深度学习中的自然语言处理技术

卷积神经网络(CNN)的引入使得文本分类技术得到了快速发展。通过专门设计的卷积核,CNN能捕捉文本的局部特征,特别适用于短文本处理。在处理序列数据时,则需要应用递归神经网络(RNN)。尽管RNN在处理长序列数据时会存在梯度消失等问题,但长短期记忆网络(LSTM)和门控循环单元(GRU)的出现为此提供了有效的解决方案,使得模型能更好地捕捉长距离依赖关系。

## 3.4 大数据与自然语言处理技术的融合

在庞大的数据环境中,预训练语言模型(如BERT和GPT)能利用大数据进行自我监督学习,以有效探索语言特征。它们无需繁琐的人工标注,能通过分析海量数据,识别复杂的语言模式和上下文关系。

# 4 实验分析

## 4.1 数据集介绍

为验证不同自然语言处理模型的性能,本次实验选择了多个广受欢迎的数据集,如表1所列。其中,SQuAD 2.0主要考察模型在问答任务中的理解能力;IMDB电影评论数据集侧重于情感分析;GLUE基准则能全面衡量模型的通用性能。SQuAD 2.0是Stanford Question Answering Dataset的升级版,专为挑战机器的阅读理解能力而设计。IMDB电影评论数据集包含5万条影评,专注于情感分析,其目标是准确判断影评的正面或负面倾向。GLUE基准涵盖9项不同的自然语言处理任务,涉及分类、匹配、推理等多个维度,能为模型提供全面的评估,是评估模型的语言理解能力的重要工具。

表1 数据集的关键特性

数据集	任务类型	数据量	评估指标
SQuAD 2.0	问答	150 000对	准确率、F1值
IMDB	情感分析	50 000条	准确率
GLUE	多任务基准	不同任务	各任务不同

## 4.2 实验环境配置

实验环境配置采用主流软硬件平台,以确保实验可靠且可重复,具体包括高性能计算机集群、最新操作系统及NLP专

文库。本次实验使用的服务器配置为NVIDIA Tesla V100 GPU,显存为32 GB,能流畅处理大规模数据。主机配备了Intel Xeon Platinum 8280 CPU和128 GB内存,以提供强大的计算能力。GPU加速技术提高了模型训练与推理的速度,同时提升了数据处理效率,缩短了实验周期。编程语言选用Python 3.8,其具有广泛的社区支持和丰富的库资源。实验基于PyTorch深度学习框架,能提供良好的模型训练与调试支持。所有依赖库版本均已锁定,使用Transformers 4.9.1、TensorFlow 2.6.0和scikit-learn 0.24.2,以确保实验的复现性。操作系统选择稳定的Ubuntu 20.04 LTS,并使用CUDA 11.3版本加速GPU,结合NVIDIA的cuDNN库,实现硬件与软件的兼容。随机种子设定为42,以确保实验结果的一致性。

### 4.3 预训练模型的性能比较

本文以BERT、GPT-2和RoBERTa进行性能比较,通过在不同数据集上的训练和测试,评估了它们的准确率、F1分数和推理时间,如表2所列。可以看到,RoBERTa在两个数据集上表现最为出色,准确率和F1分数均优于其他模型,且推理速度更快。另外,BERT在IMDB情感分析任务上表现接近RoBERTa,GPT-2在文本生成等任务中也展现出独特的优势。总体而言,RoBERTa在各方面表现相对较优,但BERT和GPT-2在特定任务上也具有各自的亮点。

表2 不同预训练模型在SQuAD 2.0和IMDB数据集上的性能对比

模型	数据集	准确率/%	F1分数/%	推理时间/ 毫秒/条
BERT	SQuAD 2.0	85.2	88.6	45
GPT-2	SQuAD 2.0	82.5	85.3	52
RoBERTa	SQuAD 2.0	87.4	89.2	40
BERT	IMDB	92.1	—	30
GPT-2	IMDB	90.3	—	35
RoBERTa	IMDB	93.0	—	28

### 4.4 实验结果分析

在SQuAD 2.0问答任务中,RoBERTa模型取得了最高的准确率和F1分数。这得益于RoBERTa使用了更大规模的数据和更长的训练时间,提升了模型对复杂上下文的理解能力。RoBERTa的成功表明,更好的预训练方式可以提高模型在特定任务上的表现。在IMDB情感分析任务中,研究显示BERT和RoBERTa的表现较为接近,准确率分别为92.1%和93.0%。这表明经过预训练的Transformer模型能在情感分类中有效捕捉文本情感,而GPT-2主要用于文本生成,在分类任务中的表现稍逊。

## 5 自然语言处理技术面临的挑战

### 5.1 多语言处理的复杂性

不同语言之间存在着巨大的语法、词汇和表达方式差异,给建立通用模型带来了困难。对于资源量较少的语言而言,由于缺乏大量标注数据,模型的训练变得更加困难。

此外,多语言模型在语义翻译、理解方面常遇到歧义问题,需要进一步研究和改进。因此,为更好地应对多语言处理的挑战,需不断进行技术创新,寻求更有效的解决方案。

### 5.2 上下文理解与长距离依赖问题

在长文档或对话中,复杂的句法结构和语义关系导致模型难以完全理解上下文的潜在含义,给机器翻译、问答系统、文本生成等任务的准确性带来了挑战。因此,研究人员需要不断探索更有效的模型和算法,以提高自然语言处理任务的表现。通过引入更多语境信息、设计更复杂的网络结构或结合其他技术手段,能在一定程度上解决这些问题,进一步推动深度学习在自然语言处理领域的应用。

### 5.3 数据和计算资源的高需求

现代自然语言处理模型(如BERT和GPT)被广泛应用于各个领域,但其预训练过程会耗费大量的时间和计算资源,对中小企业和独立研究者构成了巨大的挑战。为推动技术的进步和普及,需要寻求更经济、高效的解决方案,使中小企业和独立研究者也能充分利用这些先进模型进行创新<sup>[5]</sup>。

### 5.4 隐私和伦理问题

在处理大量用户数据时,如何保护个人信息和敏感数据,同时提高模型的准确性,成为一个紧迫的问题。此外,模型在训练过程中可能会潜在地吸收带有偏见的信息,导致决策出现偏差,引发一定的社会风险。因此,必须同时考虑数据隐私保护和偏见消除,确保NLP技术的发展符合伦理规范。在加强数据安全措施的同时,还需建立对模型的解释和监督机制,减轻模型可能带来的潜在风险。

## 6 结语

本文分析了自然语言处理技术的应用,揭示了人工智能推动下的NLP发展前景和挑战。深度学习和大数据的结合提高了NLP模型的性能,尤其是在预训练语言模型的带动下,NLP在各个领域有了更广泛的应用。然而,NLP面临的挑战也随之而来,包括多语言处理的复杂性、上下文理解的难度、对大规模数据和计算资源的需求、涉及隐私和伦理问题等。因此,如何克服这些挑战将是NLP持续发展的关键。

### 参考文献

- [1] 周彦伟.人工智能在计算机科学与技术中的应用与挑战[J].信息记录材料,2024,25(6):170-172.
- [2] 陈德光,马金林,马自萍,等.自然语言处理预训练技术综述[J].计算机科学与探索,2021,15(8):1359-1389.
- [3] 陈鑫益.面向自然语言处理的对抗样本生成技术研究[D].长沙:国防科技大学,2021.
- [4] 陆正扬.基于计算机自然语言处理的机器翻译技术应用与简介[J].科技传播,2019,11(22):140-141.
- [5] 王飞,陈立,易绵竹,等.新技术驱动的自然语言处理进展[J].武汉大学学报(工学版),2018,51(8):669-678.