文章类型：论文|刊号（ISSN）：2630-4740 /（中图刊号）：715GL012

大语言模型在生态环境领域文本编制应用的研究

钱文敏　唐芬　张映程　陈鸣*
云南省生态环境工程评估中心
DOI:10.12238/eep.v7i8.2204

[摘　要] 随着大语言模拟技术的日益发达,它在生态环境领域的运用也将越来越普遍。文章对文本的现状做出了总结,并运用私有化的大语言模式进行生态环保文本编制的具体实现途径也做出了介绍。首先,说明了当前文本处理技术中出现的几个问题并做出了说明。其次,研究了大语言模型在文本处理领域所存在的优点,并经过研究了国内一些大语言模型,对比之后,选择了私有化百度知识增强大模型,INTIE3.0大语言模型的解决途径编制了生态环境方面的研究论文。最后,还预测了私有化的大语言模式在生态环境领域的未来发展,包括技术创新和算法改进、数据库增强可持续性和可解释性等方面方向。文章还指出,私有化的大语言模型及其数据调取与咨询报告将为生态环保与可持续发展提供更强大的技术保障。
[关键词] 大语言模型；生态环境；文本处理
中图分类号：X171.1　文献标识码：A

Research on the application of large language model in ecological environment

Wenmin Qian　Fen Tang　Yingcheng Zhang　Ming Chen*
Yunnan Provincial Ecological Environment Engineering Assessment Center

[Abstract] With the increasing development of large language simulation technology, its application in the field of ecological environment will become more and more common. The article makes a summary of the current situation of the text, and introduces the specific realization ways of the ecological environmental protection text compilation by using the privatization language model. First, several problems arising in the current text-processing techniques are illustrated and explained. Secondly, it studies the advantages of large language models in the field of text processing, and studies some domestic large language models, after comparison, chooses the privatized Baidu knowledge enhancement large model, and the solution of INTIE3.0 large language model has compiled research papers on ecological environment. Finally, the future development of the big language model in the ecological environment, including technological innovation and algorithmic improvement, database enhancement sustainability and interpretability. The article also points out that the large language model of privatization and its data retrieval and consultation report will provide a stronger technical support for ecological and environmental protection and sustainable development.
[Key words] big language model; ecological environment; text processing

引言

在数字化社会,文字写作已经是各个领域不可或缺的组成部分。不管新闻、宣传文案、作品撰写或是内容营销,优质的文字撰写对传递信息、留住阅读和提高价值至关重要[1,2,3]。由于深度学习和自然语言处理技术的进展,大语言模型在编写文本方面也呈现出了前所未有的优越性。它可以认识并形成复杂的文字信息,给文字写作工作提供了革命性的改变。论文将探讨基于大语言模型的文本编写技术以及应用,并希望为生态环境领域咨询行业的蓬勃发展提供有益的借鉴。

1 文本处理的国内现状

首先,虽然文本处理技术日益发达,但是中文文本处理的困难依然很大。中文文本的语言、语义、句法等方面,和英文等西方语种有着很大不同,使得中文文本处理存在着一系列问题,如分词、词性标识、句法分析等任务的技术难度较高[3,4,5]。其次,文本处理技术在现实应用中面临着若干问题。目前国内外文本处理领域的科学研究与运用还面临相当的困难与问题,必须继续强化科学研究与实际的紧密结合,增强科技运用的科学性和稳定性,并促进各领域中的实践运用与研究发展。国内高校和学术组织中积极开展大语言模型的研制与发展,促使文字处理的有关研究进展,特别是生成文字取得了重大突破。基于此,文章

文章类型：论文|刊号（ISSN）：2630-4740 /（中图刊号）：715GL012

将提供基于私有云语言大模式的思维处理生态环境领域文本处理大模式解决咨询领域写文本获取生态环境数据复杂、撰写报告繁琐、持续时间长等问题[6,7,8]。

目前关于文本处理面临的重点问题

(1)语义理解的难度:对文本的语义理解是文本管理中的一项主要问题。因为语言的复杂性和歧义性,机器在认识和理解文字时总是受到很大的挑战。

(2)情感分析的复杂性:情感分析是文字处理中的另一项主要任务,但因为情感表现的多元化以及文化背景的不同,机械在开展情感分析时可能产生误判。

(3)语言模型的可理解度较差:由于目前的大语言模型常常没有可理解度,使得使用者无法了解模型作出决定的原因,这也影响了建模技术在某些领域的广泛应用。

(4)数据信息隐私与安全性问题:随着文本管理科技的进展,数据信息隐私与安全性问题日益凸显。怎样保障客户隐私与安全,并进行高效的文本处理是亟待解决的问题[9,10,11.12]。

2 目前在文本处理的大语言模式的优势

(1)强大的文字生成与能力:强大语言模型经过对大量文字数据的练习,掌握了大量的语言基础知识和表达方式,可以得到高质量的文字信息,同时可以掌握自然语言的语法、语义和上下文信息,增强文字处理的准确性。

(2)良好的可扩展性:大语言模型能够通过进一步扩大模型的规格和参数量,增强处理能力和有效性。同时,经过迁移学习和微调,大语言模型还可以满足对特定领域和任务的文本处理要求。

(3)有效的认知与逻辑推理功能:大语言模型通过有效的深度阅读算法和运算结构,可以在短时间内实现丰富的阅读和逻辑推理任务,大大提高了文本处理的有效性。

3 实现的技术途径

基于目前生态环境范围内的数据难、撰写报表时限长等问题；本文中提供了一个定制化私有的大语言模型体系产品的解决方案,重点面对着大国企和事业单位,该系统更强调了数据的保密性、文本的专业性,因此无法直接应用当前的大语言模型,因此一定要根据已有的大语言模型进行定制化开发,本方案优势主要包含：对私有数据的规范化处理归档、模型离线训练与推理、多模块定制化开发。

3.1大语言模型的概念

大语言模型(Large Language Model,ML)是一个采用深度学习方法的自然语言处理模式,通过对大量的文字资料集加以练习,就可以认识并形成全人类语言。大语言模型一般使用Transformer或其变种为基础结构,并利用自注意力机制和多头注意力机制进行对输入序列的全局理解。大语言模型拥有强有力的文字产生能力与理解,能够广泛应用于各类自然语言的问题,包括文字分析、问答、机器翻译等。

3.2模型分析

由于私有云大模型定制化设计解决方案所面对的主要目标对象是云南省生态环境系统的工作人员,所处理的语言都是中文,因此最终产生的主要内容就是中文报告,第二步快速完成了报告的自动形成;因为是针对生态环境这个领域,现在该项目的落脚点是研发一个定制化、可训练数据的私有大语言模型系统。该体系将针对性地服务于大规模国企和事业单位,以适应其在大数据分析保密性、模型训练效能以及定制化需求方面的特殊需求;通过调研,现在中国国内较好的大语言模型是openAI公司的GPT4、GPT-3；谷歌的Gocker、PaLM和LaMDA；百度的文心一言；阿里的通义M6；腾讯的混元；华为的XLNet中文模块；考虑到处理的语料库和文本处理的种类都是中文的,和后续数据安全,所以选择国内的大语言模型,首推BAT,而阿里的通义M六主要处理的对象是图像处理方向;华为的XLNet中文模式与腾讯的混元模式,对大型中文语料的预训练需要很大的运算资源与时间,所以首推百度的文心一言,基于百度知识增强的大模型INTIE3.0；ERIE 3.0是由百度公司开发的一个基于Transformer结构的预训练式语言模型。

3.3大语言模型的工作特点

3.3.1数据向量化

为让大语言模型识别所给出的语料,同时也为以后的计算提供基础,将给出的数据进行统计化是重要的步骤,即语料数据等向量化处理；比如能使大模型能够迅速地判定同一事件,如针对动物狗和猫、电脑；在通过把三者完成数字矩阵显示之后,大模型能够快速计算,对上述三者进行运算并得出猫、犬是同类。

3.3.2理解问题

大模型设计了许多层不同的规律,如几个算法,从各个视角去理解用户输入的问题中的词汇、句型等；并且在此基础上预言下一个出现的词语;比如如果大语言模型设计三层规矩,第一层规矩理解句子的意义,比如说“晴天”的意思肯定和天气相关;第二层是找出语句中词汇的关联性,文档中相应的词语;如果看到句子中出现晴空万里,下一个出现的词很可能是“晴天”,因为根据统计的规则,晴空万里和“晴天”的关联性非常高；第三层规则则是通过对前面所掌握的知识来掌握下一种可能产生的词汇；假设前面就已发现了“晴空万里”和“太阳伞”这些词,那么大语言模型通常认为下一句很可能出现“晴天”这种词语;第二部则通常引入基于自注意力体系的深度学习框架,如Transformer。在训练过程中,模型学生会针对输入的文本顺序,并利用自注意力机制和多头注意力机制进行对输入顺序的全局了解,从而学习语言的概率分布。GPT第三代提供了九十六层的transformer,参数为一千七百五十亿个。

3.3.3利用大量数据训练模型,增加模型的准确性

通过大量的数字对大语言模型展开练习,通常训练信息有两大类,一是来自网络中的通用信息,如大模型通过阅读网络中的网站和文字；一类数据来源于各领域的专业数据；利用上述方法对大模型加以训练,进而提升大语言模型的准确性。

3.4大语言模型的编写文本实现的基本技术途径

文章类型：论文|刊号（ISSN）：2630-4740 /（中图刊号）：715GL012

3.4.1大数据获取与处理

由于我们要建立生态环境领域的文本管理大模式,在数据的选择上必须进行除了普通数据的培训,要更加注重于专业数据的训练,可以克服生态环境领域咨询行业特有的报告编制困难、时间长以及获得数据处理不方便的问题；专业信息采集与管理,包括环评报表、竣工检验报表、集成固定污染源信息、检测数据、水质监控站监控断面信息、主要污染源在线信息、全省二次污染源检测、环评审批、竣工环境保护检验、污染许可、危废、辐射等信息数据资料核心。

3.4.2模型选取与训练

由于是根据中文文本的处理与生成,按照任务要求选取相应的深度学习结构,根据上面模型选取的理由,本文选择的大模型解决方案是基于百度知识增强大模型ANNIE3.0。因为是通过此模式能够进行定制化研发,因此还能够优化和拓展产品;然后利用预处理后的大数据分析对模型进行训练,让模型可以掌握到语言的语言、语义以及上下文信息。

3.4.3调优与微调

在建模训练流程中,按照每年更新出台的法律、规章和规范、导则、管理办法等有关文件,不断去完善大建模中相应的指标与计算规范；或者根据研究业务人员的情况和实践提出相应的研究方法和指导方法,可以采用调整超参数、采用不同的优化器等手段对模型加以调整。针对某个领域的文本处理任务,可通过领域内有关的数据对模型加以微调,以改善模型在该领域的稳定性能。

3.4.4文本生成与评价

在给定一个输入文本时,大语言模型会按照所学习到的语言概率分布,从而产生和输入有关的文本。可以通过不同的评价方法对生成的文本加以评价,如GLEU、DAOGE等。它能够对业务人员或者进行反馈,形成问题反馈体系不断的优化文本生成。

4 大语言模型的编写文本总结与发展

大语言模型在文本处理方面它利用深度学习方法对大量的信息集开展训练,可以了解语言的结构、语义和上下文信息,同时可以掌握自然语言的复杂性和歧义性。虽然大语言模型已经在文章撰写与处理等方面都获得了一定的成效,但仍存在着若干挑战,如环评报告与咨询报告、生态文明示范区规划编写以及专业报告的工作编制。所以本文通过私有云大模型解决方案进行专业报告的撰写,即通过现有送审稿的环评报告和生态文明示范区规划报告等专业数据的形成专业的私有化专业语料库,即搜集与整合数据阶段;然后,通过运用以上数据获取特征规律以及指定的特殊规则对百度知识增强大模型,INTIE3.0进行训练已超过了专业报告的技术水平；最后对大语言模型建立了一些奖励与纠正制度,以增强大语言模型编写专业报告的品质与准确性。

[参考文献]

[1]赵俊华,文福拴,黄建伟,等.基于大语言模型的电力系统通用人工智能展望:理论与应用[J].电力系统自动化,2024(006):048.

[2]张宁,豫谢辛,陈想,等.基于知识协同微调的低资源知识图谱补全方法[J].软件学报,2022,33(10):3531-3545.

[3]邓力为.基于知识图谱的文本自动生成技术的研究与实现[D].电子科技大学,2020.

[4]宋时磊,杨逸云.大语言模型的主权,安全及其治理[J].中国高校社会科学,2023(6):109-118.

[5]李敬灿,肖萃林,覃晓婷,等.基于大语言模型与语义增强的文本关系抽取算法[J].计算机工程,2024,50(4):87-94.

[6]胡志强,潘鑫瑜,文思捷,等.结合多模态知识图谱与大语言模型的风机装配工艺问答系统[J].机械设计,2023,40(S02):20-26.

[7]单斌,李豪杰,文艳伟,等.大语言模型时代的材料信息提取和数据驱动研发[J].金属功能材料,2023,30(3):1-16.

[8]杭州欧若数网科技有限公司.基于大型语言模型和图网络模型的文档检索方法和装置:CN202310693598.6[P].2023-07-14.

[9]周辉,廖浩.浅谈大语言模型在构建新型互联网出版中的实践探索[J].互联网周刊,2023(17):18-21.

[10]赵思同.一种基于对比学习和大语言模型的图数据语义搜索方法:CN202310706047.9[P].CN116450867A[2024-06-06].

[11]杨倩,林鹤.大语言模型背景下情报研究的数字化应对策略及实践场景[J].竞争情报,2023,19(3):2-13.

[12]汪骞,暴宇健.应用于角色扮演推理类游戏的大语言模型的训练方法:CN202310884926.0[P].CN116603249B[2024-06-06].

作者简介：

钱文敏(1981--),男,汉族,江苏省溧阳市人,博士,单位：云南省生态环境工程评估中心,正高级工程师,研究方向：大气污染防治。

通信作者：

陈鸣(1988--),男,汉族,湖南邵阳人,硕士,单位：云南省生态环境工程评估中心,工程师,研究方向：自然语言处理。