

陈巍,熊继平,李正浩,等. 企业技术难题与专家精准对接语言模型 Expert-ChatGLM[J]. 智能计算机与应用,2024,14(10): 131-135. DOI:10.20169/j.issn.2095-2163.241018

企业技术难题与专家精准对接语言模型 Expert-ChatGLM

陈巍¹,熊继平¹,李正浩¹,陈卫东¹,李佳珂²

(1 浙江师范大学 物理与电子信息工程学院,浙江 金华 321004; 2 浙江师范大学 计算机科学与技术学院,浙江 金华 321004)

摘要:传统方式下,企业引进技术领域的专家时,通常通过简历等信息筛选。但该方式存在同质化高、精准匹配难等问题。为此,建立一个能实现专家与企业需求精准对接的模型迫在眉睫。近年来,随着人工智能的发展,出现了很多大型语言模型,如 OpenAI 的 ChatGPT、斯坦福的 Alpaca 和清华大学的 ChatGLM-6B。虽然这些模型在指令和聊天方面取得了成绩,但对于专家与企业的精准对接尚有不足。因此,提出了专家与企业精准对接语言模型 Expert-ChatGLM,该模型对研究采集构建的专家知识图谱以及专家对话数据集进行了微调,能够准确理解企业需求并提供精准的专家信息。该模型能提高寻找专家的效率和质量,具有很大的发展潜力。

关键词:专家与企业精准对接语言模型;专家知识图谱;大型语言模型;微调

中图分类号:TP399

文献标志码:A

文章编号:2095-2163(2024)10-0131-05

Expert-ChatGLM, A language model for the precise matching of technical problems and experts

CHEN Wei¹, XIONG Jiping¹, LI Zhenghao¹, CHEN Weidong¹, LI Jiako²

(1 College of Physics and Electronic Information Engineering, Zhejiang Normal University, Jinhua 321004, Zhejiang, China;

2 School of Computer Science and Technology, Zhejiang Normal University, Jinhua 321004, Zhejiang, China)

Abstract: In traditional ways, when companies hire experts in the field of technology, information screening is usually used such as resumes. However, this method faces problems such as high homogeneity and difficulty in accurate matching. Therefore, it is urgent to establish a model that can achieve precise matching between experts and corporate needs. In recent years, with the development of artificial intelligence, many large language models have emerged, such as OpenAI's ChatGPT, Stanford's Alpaca, and Tsinghua University's ChatGLM-6B. Although these models have achieved results in instructions and chatting, some deficiencies still exist in accurate matching between experts and companies. Therefore, the paper proposes the language model "Expert-ChatGLM" for the precise matching of expert and enterprise, which is fine-tuned on the collected expert knowledge graph and the expert dialogue dataset. It can accurately understand the corporate needs and provide accurate expert information. This model can improve the efficiency and quality of finding experts and has great development potential.

Key words: language model for the precise matching of expert and enterprise; expert knowledge graph; Large Language Model; fine-tuning

0 引言

近来,随着自然语言处理技术^[1]的快速发展,强大的语言聊天模型也在不断涌现。其中包括了许多大型语言模型,如 ChatGPT^[2]、Meta 的 LLama^[3]以及 Alpaca^[4]。这些大型语言模型展现出了卓越的聊

天和理解能力,因此备受大众青睐。如今,大型语言模型(Large Language Model, LLM)是设计先进的人工智能^[5]系统之一,通过使用深度学习技术进行大量文本数据的广泛训练,该类模型在各个领域都具有非凡潜力,例如在代码生成、语言翻译、文本生成以及与人类进行聊天互动方面。在众多大型语言模

基金项目:金华市公益项目(2021-4-116)。

作者简介:陈巍(1998-),男,硕士研究生,主要研究方向:深度学习;李正浩(1999-),男,硕士研究生,主要研究方向:深度学习;陈卫东(1999-),男,硕士研究生,主要研究方向:深度学习;李佳珂(1998-),男,硕士研究生,主要研究方向:深度学习。

通讯作者:熊继平(1982-),男,博士,副教授,主要研究方向:深度学习,图像处理。Email:xjping@zjnu.cn

收稿日期:2023-06-12

型中,ChatGPT取得了可观成果^[6]。

然而,ChatGPT并非旨在帮助企业精准对接各个领域的专家信息,如果将其直接用于企业与专家信息精准对接,可能会导致所提供的专家信息出现错误。这将降低企业寻找专家信息的效率,并有可能因为无法及时找到该领域的专家信息而影响企业项目的解决,进一步带来利益损失。实际上,这也是大多数LLM所存在的局限性。

为了提高专家与企业之间的精准对接语言模型的效率和准确性,研究搜集了各个领域大量的专家信息来构建数据集。然而,若想获得经过优化的专家和企业精准对接语言模型(Expert-ChatGLM)更重要的是要探索一种可用于优化的语言模型^[7]。

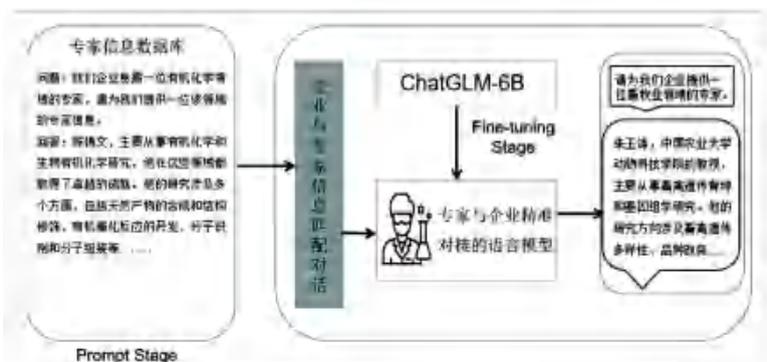


图1 Expert-ChatGLM模型训练流程图

Fig. 1 Flowchart of Expert-ChatGLM model training

1 专家与企业精准对接语言模型

1.1 专家信息数据集

为了训练专家与企业对接的语言模型,研究时首先需要收集各个技术领域内大量的专家信息。因此,通过广泛的互联网搜索收集构建了一个全面的专家信息数据库,该数据库包含了各种领域的专家姓名、研究方向、获奖荣誉等详细信息。为了确保语言模型具备高质量^[9]的对话能力,研究利用该数据库中的信息作为输入提示,调用GPT-3.5 API^[10]生成企业查询专家信息的对话,从而得到了一个专门针对专家与企业精准对接的语言模型的对话数据集。构建该数据集的主要目的是为语言模型提供精准无误的专家信息,以帮助企业在寻找自己领域的专家时高效和高精度地生成信息,供企业参考。

经过研究微调,这个模型不仅具备了ChatGLM-6B的对话性能^[11],还可以利用文中精心构造的数据集学习如何将专家信息与企业需求精准对接。最终,成功地构建了一个包含大量企业与专家对话的数据集,即Experts & Enterprises Dialogue。表1展示

Expert-ChatGLM模型训练流程如图1所示。

因此,选择使用清华大学开源的大型语言模型ChatGLM^[8],并对其进行微调,将研究构造的专家信息数据集注入其中进行训练,从而实现了企业快速且精准地匹配到所需的专家。本文的方法在以下3个方面做出了贡献:

- (1)收集了包含5万多个专家信息对话样本的数据集,用于微调LLM。
- (2)设计了一个用于微调LLM的框架。
- (3)验证了使用专家信息数据集微调的LLM模型,可以使其更好地适应企业领域的需求,具有良好的发展前景,可以实现专家与企业之间快速的精准对接。

了该数据集中的一些具体对话示例。

1.2 LoRA微调方式

LoRA微调^[12]是一种在大型语言模型上进行优化的技术,其核心思想是在指定参数(即权重矩阵)上并行添加一个额外的低秩矩阵。通过仅在训练过程中训练这些额外添加的低秩矩阵参数,可以实现微调优化。当新增低秩矩阵的“秩值”远小于原始参数维度时,新增低秩矩阵的参数数量也就很小。因此,在进行微调训练时,只需训练这些较小的参数即可,从而取得更好的表现结果,其微调流程如图2所示。通过替换图2中的矩阵A和B来冻结预训练模型进行训练,最终将可训练矩阵与冻结权重合并,从而使得存储需求和任务切换开销显著降低。

对于预训练的参数矩阵 $W_0 \in R^{m \times n}$,不去直接微调 W_0 ,而是对增量做低秩分解假设:

$$W = W_0 + BA, B \in R^{m \times r}, A \in R^{r \times n} \quad (1)$$

在训练过程中 W_0 被冻结,并且不接受梯度的更新,A与B是可训练参数。对于 $h = W_0x$,前向传播为:

$$h = W_0x + \Delta Wx = W_0x + BAx \quad (2)$$

所以 r 可以取得很小, 很多时候甚至可以直接取 1。因此, LoRA 是一种参数高效的微调方法, 至少大大降低了被优化的参数量。

表 1 专家信息对话数据集中一些企业咨询专家信息的例子

Table 1 Examples of companies consulting expert information in the expert information conversation dataset

编号	类型	对话
1	Content:	你好, 我们企业想找一位自动化领域相关的专家帮助我们解决一下问题, 你能为我们提供一位吗?
	Summary:	赵卫东, 自动化领域的杰出学者和技术专家, 主要研究方向包括智能制造、智能仪器、工业控制和机器人技术等。他在自动化领域的研究中, 主要贡献和成就包括: 在智能制造方面, 致力于智能制造技术的研究和应用, 提出了一系列先进的智能制造理论和方法, 包括基于机器学习和数据挖掘的制造过程建模与优化、智能化制造过程监控和质量控制、智能化制造系统集成等。在智能仪器方面, 主要包括基于传感器技术和数据分析的智能仪器设计和开发、智能化仪器测试和标定等。他的研究成果已经在多个领域得到了广泛应用。在工业控制方面, 主要包括基于控制理论和控制技术的工业过程控制和优化、自适应控制和智能控制等方面。他的研究成果在钢铁、石化、电力等行业得到了广泛应用。
2	Content:	你好, 我们企业遇到了一个难题, 能为我们找出一位有机化学领域相关的专家吗?
	Summary:	李建国, 主要从事有机合成化学和药物化学的研究工作。在有机合成化学方面, 李建国教授的研究兴趣主要包括以新的方法和策略合成生物活性分子和天然产物, 发展可持续化学和绿色化学的新反应及新技术等。他和他的研究团队成功地合成了多种具有生物活性的化合物, 如活性天然产物烟酰胺类, 以及一些具有抗肿瘤、抗病毒、抗炎症和抗菌活性的分子。他们也发展了多种有价值的合成方法, 如手性碳-氮偶联反应、三碳合成法、以及分子内的钯催化反应等。在药物化学方面, 他的研究团队致力于发现和研究新的药物分子, 以解决现有药物在治疗某些疾病时存在的局限性。他们利用多种手段, 包括计算机辅助药物设计、天然产物的结构优化、以及基于生物活性分子的分子修饰等, 设计和合成了多个新的具有生物活性的分子, 其中一些分子已经进入了前期的药物筛选阶段。
3	Content:	你好, 请为我们寻找一位机械领域的专家。
	Summary:	王振义, 中国著名的机械工程专家。他主要从事先进制造技术、数字化制造、高性能机床等方面的研究工作。在先进制造技术领域, 他开展了很多重要的研究工作, 包括微纳加工、超声波加工、激光加工等方面。他主持了多项国家级科研项目, 在先进制造技术方面做出了重要的贡献。也是数字化制造领域的专家之一。他主要从事数字化制造技术在机械加工、零部件制造等领域的应用和研究工作, 包括 CAD/CAM 技术、数字化加工技术等方面。
4	Content:	请为我们提供一位动物医学方向的专家。
	Summary:	陈红, 中国农业大学动物医学院的教授, 主要从事动物疫病免疫、病原微生物学、分子病理学、分子流行病学等方面的研究工作。她的研究领域主要包括动物疫病免疫机理和免疫调控、病原微生物学和病原生物学等方面。在动物疫病免疫方面, 她关注动物免疫应答的分子机制和调控网络, 研究免疫抗原的识别、处理和递呈机制、免疫应答的信号转导机制, 以及免疫细胞的分化、功能和调节等方面。在病原微生物学和病原生物学方面, 她致力于揭示病原微生物的致病机制、生物学特性和流行病学规律, 发展病原学新技术, 研究重大传染病的流行病学和防控策略等方面。
5	Content:	我们企业需要一位信息与通信方向的专家, 请提供一位此领域的专家信息。
	Summary:	吕建, 在信息与通信安全领域有着丰富经验的专家。他主要从事以下方面的研究: 在信息安全方面, 研究信息安全的基本理论、安全算法、安全协议和安全管理等问题。在网络安全方面, 研究网络安全的技术和策略, 包括入侵检测、恶意代码分析、漏洞评估等方面的问题。在量子密码学方面, 研究基于量子技术的安全通信, 开发新的量子密码算法和协议。在深度学习安全领域的应用方面, 研究如何将深度学习技术应用于信息安全和网络安全领域, 以提高安全防御的效率和准确性。他在信息与通信安全领域拥有丰富的研究经验和深厚的理论功底。

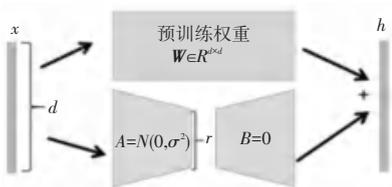


图 2 LoRA 微调流程图

Fig. 2 LoRA fine-tuning process

本文采用了一种更高效的方法, 旨在提高参数效率, 因为参数增量 $\Delta\varphi = \Delta\varphi(\theta)$ 由一个更小的参

数集 θ ($|\theta| \ll |\varphi_0|$) 表示。所以, 寻找 $\Delta\varphi$ 的任务变成了对 θ 的优化:

$$\max_{\theta} \sum_{(x,y) \in Z} \sum_{t=1}^{|y|} \log(P_{\varphi_0 + \Delta(\varphi)}(y_t | x, y < t)) \quad (3)$$

因此, 采用低秩表示来进行编码 $\Delta\varphi$, 有利于在节省计算量和内存的同时, 还能够显著减少可训练参数 $|\theta|$ 的个数。

1.3 模型的训练

研究采用了清华大学开源的 ChatGLM-6B 语言

模型,尽管 ChatGLM-6B 的参数量要小于其他大型语言模型的参数量,但其技术却与 ChatGPT 类似^[13]。该模型使用了约 1 万亿个中文和英文标记的语料库^[14]进行训练,并采用了额外的监督微调^[14]、反馈引导和使用人类反馈的强化学习。相较于其他模型,该模型在性能上也能够媲美其他大型语言模型。此外,ChatGLM 具有独特的缩放属性,使得模型可以量化,从而在单个 3060 的 GPU 上也能有效地进行推断。采用量化模型的方式可以在成本低廉的 GPU 上进行推断和计算,使得该模型更易于部署,这也是专家和企业精准对接语言模型的一个技术突破。

为了微调训练 ChatGLM-6B 模型,使用前期构建的对话演示数据集,并采用了 LoRA 微调方式。研究使用 A100 GPU 进行训练,耗时 9 个多小时。在训练过程中,将学习率设置为 $2e-5$,共进行了 1 个 epoch,最大序列长度为 512,最大目标长度为 100。

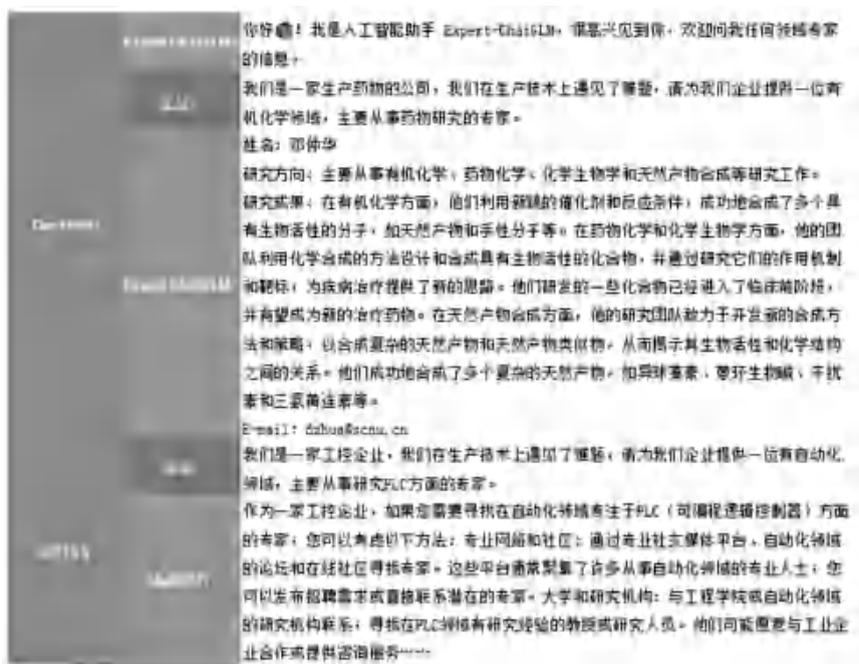


图 3 与 Expert-ChatGLM 以及 ChatGPT 进行专家信息咨询的对话例子

Fig. 3 Examples of conversations with Expert-ChatGLM and ChatGPT for expert information consultation

3 局限性

需要明确的是,目前该语言模型仅适用于学术研究。由于研发的语言模型提供的专家信息仅供参考,不能保证百分之百的准确性。因此,需要进一步收集更多的信息,构建更大的数据集来进行微调训练,从而开发出能够精准对接专家与企业的语言模

2 实验结果

经过使用本文构建的数据集,在 ChatGLM-6B 微调训练后,成功开发了一个专家与企业精准对接的语言模型。假设企业在生产过程中遇到问题,但是内部人员无法解决并亟需确保正常生产。此时,企业可以通过输入相关领域的信息,利用研究微调后的语言模型查询所需的专家并及时解决问题,从而避免企业遭受损失。运用 Expert-ChatGLM 模型,以寻找与企业所需领域专家的对话结果如图 3 所示。

研发的语言模型表现出色,与 ChatGPT 生成的专家信息相比,在准确性和全面性方面表现更加优秀。此外,本文的语言模型可以根据企业的需求精准匹配专家,从而提高企业的效率并给予有效的帮助。综上所述,本文研发的语言模型为企业提供了有力的支持,有助于解决生产过程中出现的问题并避免潜在的利益损失。

型,为企业提供高度准确的信息。只有这样,企业才能充分信任该模型并继而投入使用。

4 结束语

本论文旨在通过构建专家信息对话数据集,并使用 LoRA 微调方法对清华大学开源的 ChatGLM-6B 大语言模型进行微调训练,从而得到一个可以使

得专家与企业精准对接的语言模型 (Expert - ChatGLM), 并具有广泛应用潜力。本文研发的语言模型可以帮助企业在技术上遇到难题时, 通过与模型对话来匹配到能够解决这些问题的专家信息, 从而有效地获得帮助, 并在某些情况下避免企业利益的损失。

然而, 直接使用大型语言模型可能会提供不准确的信息, 这使得 LLM 存在许多不稳定因素。因此, 在未来的发展工作中, 让大型语言模型生成准确的结果并抑制不确定的响应, 以及通过传统的人工方法进行额外的安全检查和构建更多高质量的数据, 都是至关重要的。尽管面临各种挑战, 但研究认为通过微调训练语言模型来提高模型的性能是未来可行方法之一, 尤其是在特定领域内。

通过微调训练所得到的专家与企业精准对接的语言模型, 在此后发展中将具有显著的优势。首先, 与其他大型语言模型相比, 本文研发的语言模型生成的结果更准确、信息更全面。其次, 从企业的角度来看, 本文研发的模型能够快速为企业提供帮助, 提高企业的工作效率。分析可知, 通过不断训练和测试, 专家与企业精准对接的语言模型必将成为每个企业颇具价值的智能助手。

参考文献

- [1] 贺倩. 人工智能技术在移动互联网发展中的应用[J]. 电信网技术, 2017(2):1-4.
- [2] OUYANG Long, WU J, JIANG Xu, et al. Training language models to follow instructions with human feedback[J]. arXiv preprint arXiv:2203.02155, 2022.
- [3] TOUVRON H, LAVRIL T, IZACARD G, et al. Llama: Open and efficient foundation language models [J]. arXiv preprint

- arXiv:2302.13971, 2023.
- [4] TAORI R, GULRAJANI I, ZHANG T, et al. Stanford alpaca: An instruction - following LLaMA model [EB/OL]. [2023]. https://github.com/tatsu-lab/stanford_alpaca.
- [5] 李舟军, 范宇, 吴贤杰. 面向自然语言处理的预训练技术研究综述[J]. 计算机科学, 2020, 47(3): 162-173.
- [6] LI Yunxiang, LI Zihan, ZHANG Kai, et al. Chatdoctor: A medical chat model fine - tuned on llama model using medical domain knowledge[J]. arXiv preprint arXiv:2303.14070, 2023.
- [7] ZHANG Renrui, HAN Jiaming, ZHOU Aojun, et al. Llama - adapter: Efficient fine - tuning of language models with zero - init attention[J]. arXiv preprint arXiv:2303.16199, 2023.
- [8] DU Zhengxiao, QIAN Yujie, LIU Xiao, et al. GLM: General language model pretraining with autoregressive blank infilling[J]. arXiv preprint arXiv:2103.10360, 2022.
- [9] ZENG Aohan, LIU Xiao, DU Zhengxiao, et al. Glm-130b: An open bilingual pre-trained model[J]. arXiv preprint arXiv:2210.02414, 2022.
- [10] GILSON A, SAFRANEK C W, HUANG T, et al. How does CHATGPT perform on the United States medical licensing examination? The implications of large language models for medical education and knowledge assessment[J]. JMIR Medical Education, 2023, 9(1): e45312.
- [11] JIN Di, PAN E, OUFATTOLE N, et al. What disease does this patient have? a large-scale open domain question answering dataset from medical exams [J]. Applied Sciences, 2021, 11(14): 6421.
- [12] HU E J, SHEN Yelong, WALLIS P, et al. LoRA: Low-rank adaptation of large language models [J]. arXiv preprint arXiv: 2106.09685, 2021.
- [13] DAS A, SELEK S, WARNER A R, et al. Conversational bots for psychotherapy: A study of generative transformer models using domain-specific dialogues[C]//Proceedings of the 21st Workshop on Biomedical Language Processing. Dublin, Ireland:ACL,2022: 285-297.
- [14] XIONG Hongxin, WANG Sheng, ZHU Yitao, et al. DoctorGLM: Fine-tuning your Chinese doctor is not a Herculean task[J]. arXiv preprint arXiv:2304.01097, 2023.