

DOI: 10.11830/ISSN.1000-5013.202411013



面向多义词例句语料生成的大模型 微调指令自动化生成框架

张子龙¹, 胡渲郎¹, 牛林峰¹, 郝瑜鑫², 王华珍¹

(1. 华侨大学 计算机科学与技术学院, 福建 厦门 361021;

2. 华侨大学 华文教育研究院, 福建 厦门 361021)

摘要: 首先,构建包含主体描述集和指令示例列表的人工指令集,作为指令池的初始化输入;然后,将指令池中的指令输入大模型,生成多条机器指令与其对应的语料,并对生成的语料进行文本修正,以获取符合要求的多义词语料;最后,采用编辑距离算法进行机器指令去重,并使用谱聚类算法对候选机器指令进行聚类,从而实现机器指令的自动化生成。通过更新的指令池,实现多义词例句语料的迭代生成。结果表明:构建的多义词例句数据集及其对应的大模型机器指令集具有较好的语言多样性、内容多样性;文本构建的多义词例句数据集在例句长度、情感、词汇标准等级难度、主题等方面能满足第二语言学习者的需求。

关键词: 大型语言模型; 指令生成; 多义词; 例句生成; ChatGPT

中图分类号: TP 3

文献标志码: A

文章编号: 1000-5013(2025)03-0328-09

Framework for Automated Generation of Fine-Tuning Instructions for Large Model in Ploysemy Example Sentence Corpora Creation

ZHANG Zilong¹, HU Xuanlang¹, NIU Linfeng¹,
HAO Yuxin², WANG Huazhen¹

(1. School of Computer Science and Technology, Huaqiao University, Xiamen 361021, China;

2. Chinese Education Research Institute, Huaqiao University, Xiamen 361021, China)

Abstract: First, a manual instruction set containing a body description set and a list of instruction examples is constructed as the initial input for the instruction pool. Then, input the instructions from the instruction pool into the large model to generate a number of machine-generated instructions corresponding to their corpora, the generated corpora are refined with text correction to obtain the desired polysemy example sentence corpus. Finally, the edit distance algorithm is used to remove the weight of machine instructions, and the spectral clustering algorithm is used to cluster the candidate machine instructions, thereby achieving automated generation of machine instructions. By updating the instruction pool, iterative generation of the polysemy example sentence corpus is realized. The results show that the constructed polysemy example sentence dataset and its corresponding large model machine instruction set exhibit good linguistic diversity and content diversity. The constructed polysemy example sentence dataset meets the needs of second language learners in terms of sentence

收稿日期: 2024-11-29

通信作者: 王华珍(1978-),女,副教授,博士,主要从事人工神经网络深度学习、自然语言处理、知识图谱和人工智能教育的研究。E-mail: wanghuazhen@hqu.edu.cn。

基金项目: 教育部中外语言交流合作中心 2021 年国际中文教育研究课题(21YH30B)

length, sentiment, vocabulary difficulty standard level , and topics.

Keywords: large language model; instruction generation; polysemy; example sentence generation; ChatGPT

中文作为一种复杂的语言,具有丰富的多义词现象,即一个字或一个词有多个不同的意义。对于汉语二语学习者而言,理解和运用多义词是一个难点。在词汇学习和阅读理解中,学习者需准确理解多义词在具体语境中的意义,逐步培养推测词义和理解句意的能力,以避免多义词可能带来的歧义和语言交际上的误解。与此同时,随着语料库语言学的兴起,语料库在汉语作为第二语言教学领域中的作用也日益显现。语料库可以提供大量真实语言使用的例句和语境,帮助学习者更好地理解多义词在不同语境中的用法和含义。因此,高质量的多义词资源建设日益受到关注。然而,目前对于带有多义词义项标注的语料库研究还相对较少,特别是多义词例句语料库需要进一步研究,以提供更多高质量的多义词资源供学习者和教师使用。

近年来,大语言模型(LLM)领域实现了突破性的进展,如 GPT-3^[1]、LLaMa^[2]等模型在自然语言任务中表现出卓越的性能。通过适当的微调指令,可以有效地引导这些模型产出预期的响应,进而在低资源领域的零样本生成任务中实现质量的显著提升。这一方法为解决多义词语料不足的问题提供了新的思路。然而,目前许多 LLM 都严重依赖人工指令,并需要经过大量手动调试才能得到一组好的指令数据集。这种人工构建指令数据集的过程既耗时又耗力,并且可能受到人类主观偏见和误差的影响。为了克服这些限制,自动化生成指令的框架成为当前研究的热点。针对这一问题,学者们已提出一些自动化生成指令的方法,但这些方法仍有不足之处。一是自动化生成指令的方法通常是基于模型的反馈进行迭代更新的,但它们往往没有将领域知识融入迭代逻辑设计中,导致生成的指令缺乏可解释性;二是这些方法通常使用模型生成结果的质量作为评价标准,而没有直接对生成的指令进行评估,忽略了指令语义对自动化生成的促进作用。基于此,本文提出一种面向多义词例句语料生成的大模型微调指令自动化生成框架。

1 相关工作

1.1 多义词语料研究

一词多义是世界不同语言在各个历史时期都普遍存在的现象。多义问题一直都是语言学家关注的问题^[3]。其中,较为典型的是基于词典的多义词研究,多集中于对词典义项设置的研究或比较不同时期汉语词典中多义词的义项异同。胡长虹^[4]比较了《国语辞典》和《现代汉语词典》中 1 450 个常用多义动词,发现与《国语辞典》相比,《现代汉语词典》义项的增加是主流,词义有复杂化趋势。周娟^[5]比较了《现代汉语词典》2002 年的增补本和 2005 年的第 5 版,发现多义词义项发生了义项增加、义项减少、义项分立和义项合并 4 个方面的变化。陈国华等^[6]分析了《汉语大词典》义项失序的问题。

此外,将计算机技术、自然语言处理技术、大数据与人工智能引入多义词语料研究也成为当前研究的热点。李安^[7]以《现代汉语分类词典》义类体系为基础,通过计算语义相似度,测量多义词义项的语义距离,并把多义词义项之间的关系分为跨义类、同义类和近义类 3 种关系类型。Lopez-Arevalo 等^[8]采用 WordNet 获取歧义词汇真实语义的方法,实现在特定领域中的词义消歧。Al-Saiagh 等^[9]提出一种模拟退火和粒子群优化混合的启发式算法,将改进的 Lesk 方法作为混合粒子群优化算法的目标函数,度量歧义词汇在不同语义类下的概率。Rahman 等^[10]提出一种基于语义扩展知识进行词义消歧的方法,并将其应用于文本查询中。通过对输入文本进行语义扩展来选择歧义词汇的正确含义,从而获得与输入文本相关的文本信息。

综上所述,前沿技术在研究词义消歧方面取得了显著进展。然而,对于多义词语料库研究,特别是关于多义词例句语料库的研究仍然较少。

1.2 LLM 微调指令自动化生成

微调指令是一种明确且规范的指导语句,用于引导模型的行为,以实现特定任务或目标。微调指令提供了一种自然且直观的方式,使人类可以与大型语言模型进行交互和使用。自动化生成 LLM 微调指令的研究可以分为以下 3 个领域。

1) 基于模板和规则的微调指令生成。根据任务类型和数据格式设计固定的指令模板,并将任务和数据的具体信息填入模板中,以生成相应的微调指令。例如,Wang 等^[11]提出 Super-NaturalInstructions,其中,包含多个自然语言处理任务和数据集的指令,它们使用简单的指令模板,如“给出一个句子,判断情感倾向”或“给出两个单词,判断它们是否同义”,以生成不同任务的指令。这种方法直观而简单,但可能缺乏灵活性和创造性,无法涵盖更复杂和多样的任务场景。Xu 等^[12]提出 Evol-Instruct 方法,旨在增强大型语言模型遵循复杂指令的能力。

2) 基于思考链的微调指令生成。利用 LLM 自身的知识和推理能力,生成一系列相关的问题和答案,形成思考链,再将思考链作为微调指令来引导语言模型完成目标任务。如 Liu 等^[13]引入逻辑链思维的微调指令数据集 LogiCoT,有效提高了 GPT-4 在复杂推理任务上的性能。此外,Zelikman 等^[14]提出 STAR 技术,该技术在一个循环中生成一步一步的解释,以提高 LLM 在复杂推理任务上的性能。这种方法具有较强的创造性,但难以控制思考链的长度和复杂性,并且可能需要大量的计算资源和时间来生成思考链。

3) 基于迭代学习的微调指令生成。利用 LLM 自身的反馈信息来不断优化指令,根据历史的输入输出数据和误差信息修正和优化控制指令。例如,Wang 等^[15]提出了 Self-Instruct,它通过从 LLM 自身生成大量的指令、输入和输出样本,并对其进行筛选和修正,再使用这些样本来微调原始的语言模型。此外,Zhou 等^[16]提出一种自动生成和选择指令的自动提示框架,展示了 LLM 在生成指令方面强大的能力。然而,基于迭代学习的方法依赖于模型的自我生成能力和反馈信息,可能在生成过程中面临指令精确性不高的问题。

1.3 ChatGPT 语料生成

近年来,使用 ChatGPT 生成高质量且多样化的语料已成为一种新颖而有效的方法。这种方法能够扩展语料库的规模,提供更多样的训练数据,并涵盖更广泛的领域和话题,从而提升自然语言处理模型的性能和适用性。这种技术对于改进文本生成任务、对话系统和语言理解等领域具有重要意义。利用 ChatGPT 生成语料,研究人员和开发者可以更好地训练和优化模型,使其在不同应用场景下表现出更强的语言生成能力和适应性。这种方法的发展将为自然语言处理领域带来更广阔的可能性,并推动其在实际应用中的进一步发展。Xu 等^[17]提出一种自聊天方法,通过引导 ChatGPT 从对话数据集中随机抽取问题或关键句子作为核心话题,生成大量数据。

然而,鉴于 ChatGPT 的通用领域特性,现有的语料生成研究和应用主要集中在通用常识领域。因此,如何使 ChatGPT 适应特定语料领域的垂直性成为中文语料生成的主要挑战。迄今为止,尚未见利用 ChatGPT 生成多义词例句语料生成的研究。

2 面向多义词例句的大模型微调指令自动化生成框架

面向多义词例句的大模型(大语言模型)微调指令自动化生成框架包括人工指令集构建、指令生成与语料修正、基于编辑距离相关的机器指令去重、基于谱聚类的机器指令示例采样 4 个步骤。面向多义词例句的大模型微调指令自动化生成框架,如图 1 所示。

2.1 人工指令集构建

人工指令集为引导大语言模型生成创新且多样化的指令提供上下文示例,构建生成多义词例句语料的人工指令集 $I_b = \{I_f, I_u\}$ 。其中, I_f 为人工指令主体描述集, $I_f = \{x_1, \dots, x_k\}$,每个组份 $x_z(z=1, 2, \dots, k)$ 对应着不同的指令生成限制描述,限制描述来源于领域垂直性约束知识,如多义词例句语料的情感、词性、语法结构、释义、长度、数量等多维度约束; I_u 为人工指令示例列表,其组合限制描述中的领域垂直性约束知识

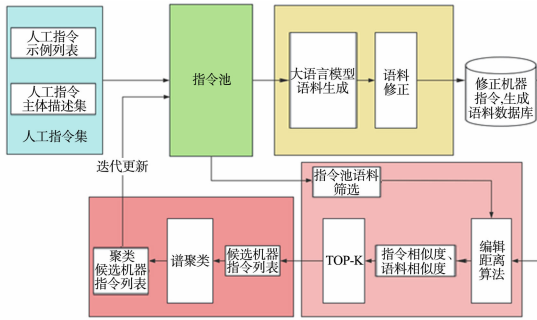


图 1 面向多义词例句的大模型微调指令自动化生成框架图
Fig. 1 Framework Diagram for automated generation of fine-tuning instruction for large model in polysemy example sentences corpora creation

的指令生成限制描述,限制描述来源于领域垂直性约束知识,如多义词例句语料的情感、词性、语法结构、释义、长度、数量等多维度约束; I_u 为人工指令示例列表,其组合限制描述中的领域垂直性约束知识

构建出指令,作为指令范例。以“阿姨”这个词的指令示例为例,“生成包含‘阿姨’这个词的 7 个例句。其中,这个词在句子中的词性为名词,且这个词的释义为‘对跟自己母亲同辈、年纪也差不多的女性的称呼。认识的或不认识的都可以用’。生成的例句长度不要超过 15 个字,带有负面的情感色彩且定中结构。不要回答除答案以外的其他内容”。

人工指令集用于指令池(Istbase)的初始化。指令池是大模型的输入端,由 I_t 和动态可变的指令示例列表(I_s)两部分组成。初始化时,指令池中的指令示例列表 I_s 为人工指令示例列表 I_u 。在后续的迭代轮次中,指令池内的指令示例列表将由模型生成的机器指令不断更新。

2.2 指令生成和语料修正

大语言模型获取指令池进行生成任务,其生成结果不仅包括多义词例句语料,还包括机器指令,即输出是多份的机器指令-生成语料(i_m, d)= $\text{LLM}(\text{Istbase})$ 。其中, i_m 为 LLM 生成的机器指令; d 为 LLM 生成的多义词例句语料。

为了确保生成的多义词例句语料符合中文教学要求,使用例句长度控制、语法修正、句子词汇难度控制 3 个指标进行修正,以保留有效的语料。

1) 例句长度控制。例句长度的控制是为了确保生成的例句语料适用于中文教学场景而进行的重要步骤。设定一个最大长度阈值 g_{\max} ,以确保例句在所需范围内。如果例句超过了最大阈值 g_{\max} ,会将该机器指令-多义词例句样本对舍弃。这是为了确保例句的紧凑性和易读性,避免过长的例句导致学习者难以理解或吸收。这有助于提高例句的可读性和可理解性,为学习者提供更好的学习体验和教学效果。例句长度(l_d)控制的计算公式为

$$\text{if } l_d > g_{\max}, \quad \text{drop.} \tag{1}$$

2) 语法修正。语法修正是确保生成的机器指令与例句语料在语法上正确的关键步骤。采用 HanLP 的语法分析工具分析和纠正例句语料中存在的语法错误,如不完整的句子结构和拼写错误等。通过该工具的应用,能有效识别并修正这些语法问题,确保例句语料在语法上的准确性和合理性。首先,将例句语料输入 HanLP 的语法分析器,该工具能够对句子进行细粒度的分析,包括例句语料的语法错误,如不完整的句子、错别字等。基于这些分析结果,能够检测到不符合语法规则的句子,并进行相应的修正。语法修正过程能够有效地提高例句语料的语法正确性,使生成的机器指令更加准确和可理解。

3) 句子词汇难度控制。句子难度控制是避免生成的例句语料中使用过于复杂或晦涩的词汇,以减少读者的认知负担,提高句子的可读性和流畅性。首先,将生成的例句语料进行分词。然后,将分词后形成的词汇进行词汇等级检测。最后,统计句子中域外词数量的占比。如果占比超过了最大阈值 p_{\max} ,会将该机器指令-多义词例句样本对舍弃。句子词汇难度控制过程能够有效地控制例句语料的难度,使生成的机器语料更加符合学习者的阅读水平。句子词汇难度的计算公式为

$$\text{if } N_{\text{ew}}/l_d > p_{\max}, \quad \text{drop.} \tag{2}$$

式(2)中: N_{ew} 为句子中域外词的数量。

通过上述方法对生成的多义词例句语料进行修正,最终可得修正的多义词例句语料 d_{tec} ,将其进一步送入修正机器指令-生成语料数据库。

2.3 基于编辑距离算法的机器指令去重

为了增强指令池的示例指令,减少机器指令-生成语料数据库中修正机器指令的差异性和冗余性,采用基于编辑距离算法进行机器指令的采样。该算法通过衡量修正机器指令与指令池示例指令之间的编辑距离,将编辑距离最小的修正机器指令作为采样结果。首先,将修正机器指令-生成语料数据库和指令池中的指令示例转化为字符串。然后,利用编辑距离算法计算修正机器指令与指令池指令示例之间的编辑距离,以及修正生成语料与修正机器指令-生成语料数据库中语料之间的编辑距离,即两个字符串之间相互转化所需的最小编辑操作次数。最后,通过加权求和计算,依此筛选合适的机器指令。

2.4 基于谱聚类算法的指令示例采样

为了系统化地降低候选指令集中的冗余性,提高数据处理的效率,采用谱聚类算法对候选机器指令列表中的机器指令进行聚类 and 采样。

首先,通过向量化器(Vec)将机器指令示例转化为特征向量,再进行候选机器指令示例特征向量之

间的相似度计算,使用余弦相似度进行计算,从而构建相似度矩阵,即

$$\mathbf{V}_{i_m} = \text{Vec}(i_m), \tag{3}$$

$$\mathbf{S} = \frac{\mathbf{V}_{i_m} \cdot \mathbf{V}_{i_m+1}}{\|\mathbf{V}_{i_m}\| \|\mathbf{V}_{i_m+1}\|}. \tag{4}$$

式(3)、(4)中: $\mathbf{V}_{i_m}, \mathbf{V}_{i_m+1}$ 均为通过平均池化获取的修正机器指令的向量; \mathbf{S} 为候选机器指令示例的相似度矩阵。

然后,针对 \mathbf{S} ,基于无向图来计算候选机器指令示例的度矩阵(\mathbf{D}),即

$$\mathbf{D} = \text{diag}(\mathbf{S}). \tag{5}$$

将 \mathbf{S} 减去度矩阵,可得拉普拉斯矩阵(\mathbf{L}),即

$$\mathbf{L} = \mathbf{S} - \mathbf{D}. \tag{6}$$

对拉普拉斯矩阵使用指数函数 `eigen` 进行特征分解,得到特征向量,并将其作为新的特征表示。

最后,将新的特征向量输入 K -means 聚类算法中进行聚类操作。为了使每个样本到其所属簇中心点的距离最小,定义目标函数 J 为

$$J = \sum_{i=1}^K \sum_{x_j \in C_i} \|x_j - \mu_i\|^2. \tag{7}$$

式(7)中: C_i 为第 i 个簇; x_j 为 C_i 的某一点; μ_i 为 C_i 的中心点; K 等于指令池中指令示例列表的大小。

通过最小化目标函数,得到每个簇的中心点。计算 C_i 簇内每个样本与中心点 μ_i 的距离,找到离中心点 μ_i 最近样本 x_{rep}^i ,将其作为 C_i 簇的代表性样本加入聚类候选机器指令列表 T_{m_K} ,计算过程为

$$x_{\text{rep}}^i = \arg \min_{x_j \in C_i} \|x_j - \mu_i\|^2. \tag{8}$$

对 K 个聚类簇分别进行计算,得到聚类候选机器指令列表 T_{m_K} 。此外,为了逐渐降低人工介入对指令自动生成过程的影响,采用逐步减少人工示例指令的权重的方法,逐渐增加机器指令的影响。具体而言,引入一个衰减率参数(取值范围为 $0 \sim 1$),用于调整指令池中示例指令的减弱幅度。通过衰减率参数的计算,确定指令池中需要减弱的指令数量,并随机移除相应数量的指令。然后,从候选机器指令的聚类列表中随机选择与减弱数量相当的指令,并将其添加到指令池中进行更新,更新公式为

$$N_n = (1 - \alpha^n) \times N_{n-1}. \tag{9}$$

式(9)中: N_n 为第 n 轮指令池指令示例列表规模; N_{n-1} 为第 $n-1$ 轮的衰减个数; α^n 为第 n 轮的衰减率。

3 多义词例句语料库构建

为了验证提出的面向多义词例句的大模型微调指令自动化生成框架的有效性,将 ChatGPT(gpt-3.5-turbo)作为大语言模型。

3.1 实验设置

3.1.1 领域垂直性知识约束的设置 在使用面向多义词例句语料生成的大模型微调指令自动化生成框架生成多义词例句语料过程中,设置等级标准多义词词表、语法结构、例句长度等领域垂直性约束。

针对多功能的外国人学汉语词典《学汉语词典》,采用版面分析与正则匹配方法抽取每个多义词的词条信息,包括词、拼音、词性、义项编号、义项、例句集等词汇要素,形成结构化的学汉语多义词词表,共包含 11 864 个词条。针对《国际中文教育中文水平等级标准》(以下简称《等级标准》)1~4 级中的每个词汇,抽取在学汉语词典数据集的词性、释义等信息,形成 1~4 级标准多义词词表 GS_poly,即该词表的每个词为多义词,且每个词都属于《等级标准》中的范畴,具有 1~4 级标准等级词汇要素信息。1~4 等级标准多义词词表 GS_poly 共包含 728 多义词,2 475 条词条信息。例如,GS_poly 中多义词[安定]有 2 个义项,因此,包含 2 个词条信息,具体为[‘安定’,‘形容词’,‘生活、情绪等平静,没有不安’], [‘安定’,‘动词’,‘使人的情绪平静’] }。

3.1.2 人工指令集构建 在使用基于大型模型微调的自动指令生成框架生成多义词例句语料时,构建一个包含 10 个组份的人工指令集合 $I_h = \{I_t, I_u\}$ 。人工指令主体描述集,如表 1 所示。人工指令示例列表,如表 2 所示。

表 1 人工指令主体描述集
Tab. 1 Manual instruction body description set

组份	人工指令主体描述集	说明
1	你被要求提供多样化的例句生成任务指令, 这些任务指令将被提供给 GPT 模型, 以生成包含“阿姨”这个字的例句。 以下是你提供指令需要满足的要求:	模型目标相关描述, 针对 GPT 模型在文中研究的面向多义词例句的大模型微调指令自动化生成框架中所承担的任务进行说明。设计时应说明大语言模型的输入和输出, 特别地, 需要具体到某个特定的多义词, 如“阿姨”, 以展示模型生成特定多义词“阿姨”的例句生成指令目标
2	不同指令中不要有相同动词, 要最大化指令的多样性	指令内容多样性, 是指在各个指令中尽量避免重复使用某些动词和类似的表达方式。设计时应考虑指令内容的多样性
3	指令应包含生成例句的数量	指令生成例句的数量约束。同时生成例句数量过少会降低语料生成效率, 同时生成例句数量过多则可能导致生成例句的高度相似性, 降低语料生成质量
4	指令中应该包含生成例句的情感, 如例句中的情感可以为正面、负面、中性	应考虑例句中所代表的情感色彩, 如例句所表达的情感可能包含正面、负面、中性
5	指令用中文书写	指令语言类型, 指令应该用中文书写, 以确保指令的语言中文教学场景需求。设计时应包含语言的选定
6	指令中应该限定生成例句的语法结构, 如定中结构、状中结构、主谓结构、主谓宾结构、主谓双宾结构、主系表结构、主谓状结构等句子类型	指令生成例句的语法约束, 能够丰富生成例句的多样性
7	指令中应该限制生成例句的长度	指令生成例句长度约束, 过短或者过长的例句会给实际应用中学生的理解带来困难。文中研究对句长进行约束, 句长不超过 20 字
8	指令应包含生成指令的长度	指令长度约束, 过长的指令输入可能会导致指令无法完全被处理, 进而影响结果的准确性
9	指令应该为 2~3 个句子	指令长度约束, 过长的指令会给模型的学习和理解带来困难, 过短的指令则不足以提供丰富的信息。设计时应注意指令的长度, 以便模型更好地理解指令
10	指令中应该限制生成例句所对应的多义词的释义	用于生成例句的特定多义词的义项约束, 防止生成该多义词的其他义项的例句

表 2 人工指令示例列表
Tab. 2 List of manual instructions example

示例	人工指令示例列表	说明
1	生成包含“阿姨”这个词的 7 个例句。其中, 这个词在句子中的词性为名词, 且这个词的释义为“对跟自己母亲同辈、年纪也差不多的女性的称呼。认识的或不认识的都可以用”。生成的例句长度不要超过 15 个字, 带有负面的情感色彩且定中结构。不要回答除答案以外的其他内容	由组份 1、3、4、6、7、8、10 的组合表达
2	以中文教学的场景, 生成包含“阿姨”这个词的 6 个例句让学生学习。这个词在句子中的词性为名词, 且这个词的释义为“对跟自己母亲同辈、年纪也差不多的女性的称呼。认识的或不认识的都可以用”。生成的例句带有正面的情感色彩, 每句例句的长度不要超过 12 个字, 不要回答除答案以外的其他内容	由组份 1、3、4、6、7、8、10 的组合表达
3	用“阿姨”这个词造 5 个句子。并且这个词在句子中的词性为名词, 且这个词的释义为“对跟自己母亲同辈、年纪也差不多的女性的称呼。认识的或不认识的都可以用”。生成的例句带有中性面的情感色彩, 且生成的例句为状中结构, 每句例句的长度不要超过 16 个字, 不要回答除答案以外的其他内容	由组份 1、3、4、6、7、8、10 的组合表达
4	设定一个场景, 你需要生成 6 个包含“阿姨”这个词的例句来供来华留学生学习。并且这个词在句子中的词性为名词, 且这个词的释义为“对跟自己母亲同辈、年纪也差不多的女性的称呼。认识的或不认识的都可以用”。生成的例句带有正面的情感色彩, 且生成的例句为主谓双宾结构, 每句例句的长度不要超过 20 个字。不要回答除答案以外的其他内容	由组份 1、3、4、6、7、8、10 的组合表达
5	您是一名教育来华学生的汉语老师, 你需要生成包含“阿姨”这个词的 4 个例句来供学生学习。并且这个词在句子中的词性为名词, 且这个词的释义为“对跟自己母亲同辈、年纪也差不多的女性的称呼。认识的或不认识的都可以用”。生成的例句带有中性的情感色彩, 且生成的例句为主谓状结构, 每句例句的长度不要超过 18 个字, 不要回答除答案以外的其他内容	由组份 1、3、4、6、7、8、10 的组合表达

该人工指令集合包括任务目标定义、面向指令的设计规范及生成多义词例句的相关参数。任务目标定义使 GPT 模型能够生成多义词例句。设计规范方面包括多义词的词性、释义、长度、数量等, 以确

保生成的指令具有多样性和适应性。此外,设计规范还有助于模型生成符合预期的指令和多义词例句语料。人工指令示例列表 I_0 由 5 个不同的指令示例组成,这些示例涵盖了情感、词性、语法结构、释义、长度、数量等多个方面。

3.1.3 多义词例句语料的后处理 为了生成符合中文教育场景需求的多义词例句语料,对生成的多义词例句语料进行精细数据后处理。在分析生成的原始例句数据集后,观察到以下 3 个问题:1) 语料格式多样性,由于 ChatGPT 的不可控性,生成的语料除文本格式外,还包括了 JSON 格式的数据;2) 回复内容冗余性,由于 ChatGPT 的交互模式特点,生成的语料可能包含与例句无关的回复;3) 例句的重复性,在多次迭代生成多义词例句语料时,可能出现生成相同例句的情况。为了应对这些问题,首先,删除非文本格式的噪声数据;然后,移除与例句无关的回复;最后,筛选出生成语料中的重复例句。通过上述数据后处理工作,最终获得约 24 万条高质量且符合国际中文教育标准的多义词例句语料。

3.2 多义词例句语料的评估指标

为了评估生成的多义词例句质量,设置的客观指标为平均字数、情感指数、《等级标准》词汇难度匹配度、《等级标准》主题匹配度。

平均字数是多义词例句语料的总字数除以例句语料的句子数量,平均字数($N_{ave,w}$)的计算公式为

$$N_{ave,w} = \frac{N_{t,w}}{N_s}。$$
 (10)

式(10)中: $N_{t,w}$ 为例句语料的总字数; N_s 为例句语料的句子数量。

情感指数是指带有正面和负面情感例句语料的数量总和在例句语料的句子数量中的占比。采用百度 AI 开放平台的情感倾向分析 API 对生成的例句语料进行情感检测,情感指数(E)的计算公式为

$$E = \frac{N_{s,pos} + N_{s,neg}}{N_s}。$$
 (11)

式(11)中: $N_{s,pos}$ 为正面情感的例句语料的数量; $N_{s,neg}$ 为负面情感的例句语料的数量。

《等级标准》词汇难度匹配度式是指多义词等级与该多义词例句语料中词汇最高等级之间一致的程度。首先,对该例句语料进行分词,并统计该例句中词汇的最高等级;然后,判断例句中词汇的最高等级是否与该多义词等级一致;最后,将符合该条件的例句语料数量除以例句语料的数量,可得《等级标准》词汇难度匹配度(M_d),其计算公式为

$$M_d = \frac{N_{s,d}}{N_s}。$$
 (12)

式(12)中: $N_{s,d}$ 为满足条件的例句数量(即例句中最高等级与词汇等级标准一致的例句数量)。

《等级标准》主题匹配度是指多义词等级与该多义词例句语料所对应主题等级之间一致的程度。《等级标准》主题匹配度(M_t)的计算公式为

$$M_t = \frac{N_t}{N_s}。$$
 (13)

式(13)中: N_t 为例句语料中多义词等级与主题等级一致的例句数量。

主题等级表,如表 3 所示。多义词语料主题是从百度 AI 开放平台中的文章分类 API 中获取的,文本通过映射方法将百度主题集(26 种)与《等级标准》等级主题集进行对应,从而获取多义词例句语料的主题等级。《等级标准》主题匹配度指标用于考察多义词等级与多义词语料的主题等级的一致性。

表 3 主题等级表

Tab. 3 Theme level table

百度主题集 《等级标准》的主题 《等级标准》的主题等级			百度主题集 《等级标准》的主题 《等级标准》的主题等级		
娱乐	兴趣爱好	1	时尚	兴趣爱好	1
游戏	兴趣爱好	1	搞笑	兴趣爱好	1
家居	日常起居	1	动漫	兴趣爱好	1
音乐	兴趣爱好	1	综合	兴趣爱好	1
美食	用餐	2	母婴育儿	家庭生活	2
汽车	交通	1	教育	教育	3

续表

Continue table

百度主题集 《等级标准》的主题 《等级标准》的主题等级			百度主题集 《等级标准》的主题 《等级标准》的主题等级		
旅游	出行经历	3	文化	课程情况	3
健康养生	健康状况	4	宠物	动物	4
社会	社会现象	5	时事	社会新闻	6
军事	矛盾纠纷	6	科技	科学技术	7
情感	心理情感	7	体育	体育	7
星座运势	心理情感	7	历史	历史	8
国际	国际事务	9	财经	商业贸易	9

3.3 多义词例句语料结果与分析

针对最终获得的约 24 万条多义词数据集进行分析。将该多义词例句数据集与《学汉语词典》等级多义词例句数据集的差异进行展示。《学汉语词典》等级多义词例句数据集是由学汉语多义词词表中匹配标准等级 1~4 级得到,包含 728 个 1~4 级多义词,2 475 条词条信息,形成 6 299 个多义词例句。

3.3.1 多义词例句的主观指标评价 引入人工评估,其评估结果具有主观性,设计的主观指标包括表达流畅度和倾向性。表达流畅度指例句的流畅性、易理解性和语言表达的地道程度;倾向性指例句是否遵循通常的中文常识和实际教学场景的规范。

从该多义词例句数据集中随机选择 200 个例句样本,并请 3 位中文教育领域专家进行评估。每个例句样本由 3 位专家独立评估,评分范围为 1~5(1 表示较差,5 表示优秀)。最后,将 3 位专家的评分取平均值作为最终的评估结果。

例句语料流畅度和倾向性的评估结果分别为 4.9、4.7。通过面向多义词例句的大模型微调指令自动化生成框架生成的多义词例句语料在各个评估指标上都表现良好,这说明生成的例句语料能够符合中文教育需求,同时也能满足不同教育自然语言处理任务的数据需求。

3.3.2 多义词例句的客观指标评价 通过客观指标比较不同数据集之间的差异,结果如表 4 所示。

表 4 多义词例句数据集的相关指标

Tab. 4 Related indicators of polysemy example sentence dataset

数据集	语言	样本数	$N_{ave,w}$	E	M_d	M_t
大模型生成等级多义词例句数据集	中文	240 564	18.6	0.970	0.044	0.331
《学汉语词典》等级多义词例句数据集	中文	6 299	12.0	0.956	0.022	0.277

由表 4 可知:大模型生成等级多义词例句数据集的《等级标准》词汇覆盖度、例句长度、《等级标准》词汇难度匹配度符合国际中文教学要求;与其他例句数据集相比,大模型生成等级多义词例句语料具有更高的《等级标准》主题匹配度,说明利用大模型能实现低资源领域语料的构建。

4 结论

构建人工指令集作为指令池的初始化输入,并利用大语言模型生成多条机器指令及其对应的语料。通过文本修正和长度修正以及句子词汇难度控制,获取更符合要求的多义词语料。采用编辑距离算法和谱聚类算法进行机器指令采样和聚类,实现机器指令的自动化生成。通过使用 ChatGPT(gpt-3.5-turbo)模型,成功地生成了约 12 200 条机器指令和 24 万条多义词例句文本。指令集涵盖了涉及多义词例句的不同任务。生成的多义词例句数据集具有较好的语言多样性和内容多样性。通过客观指标和专家主观评价,验证了生成的多义词语料的质量和契合度,表明其能满足中文学习者的学习需求。因此,利用大模型进行低资源领域语料构建具有可行性。

参考文献:

[1] BROWN T, MANN B, RYDER N, *et al.* Language models are few-shot learners[J]. Advances in Neural Information Processing Systems, 2020, 33: 1877-1901.

[2] TOUVRON H,LAVRIL T,IZACARD G,*et al.* Llama: Open and efficient foundation language models[EB/OL]. (2023-02-27)[2024-12-24]. <https://arxiv.org/abs/2302.13971>.

[3] 赵颜利,董博,雷燕.我国语义标注领域研究现状分析[J].福建师范大学学报(自然科学版),2020,36(4):17-24,36. DOI:10.12046/j.issn.1000-5277.2020.04.003.

[4] 胡长虹.《国语辞典》和《现代汉语词典》常用多义词义项处理对比研究[D].烟台:鲁东大学,2013.

[5] 周娟.《现代汉语词典》新旧版本多义词义项变化计量研究[D].南宁:广西大学,2011. DOI:10.7666/d.y1952844.

[6] 陈国华,李申.《汉语大词典》义项失序问题研究[J].辞书研究,2015(1):10-18. DOI:10.3969/j.issn.1000-6125.2015.01.002.

[7] 李安.多义词义项的语义关系及其对词义消歧的影响[J].语言文字应用,2014(1):29-37.

[8] LOPEZ-AREVALO I,SOSA-SOSA V J,ROJAS-LOPEZ F,*et al.* Improving selection of synsets from WordNet for domain-specific word sense disambiguation[J]. Computer Speech & Language,2017,41:128-145. DOI:10.1016/j.csl.2016.06.003.

[9] AL-SAIAGH W, TIUN S, AL-SAFFAR A, *et al.* Word sense disambiguation using hybrid swarm intelligence approach[J]. PloS One,2018,13(12):e0208695. DOI:10.1371/journal.pone.0208695.

[10] RAHMAN N,BHOGESWAR B. Improvement of query-based text summarization using word sense disambiguation [J]. Complex & Intelligent Systems,2020,6:75-85. DOI:10.1007/s40747-019-0115-2.

[11] WANG Yizhong,MISHRA S,ALIPOORMOLABASHI P,*et al.* Super-NaturalInstructions: Generalization via declarative instructions on 1600+ NLP tasks[EB/OL]. (2022-04-16)[2024-12-24]. <https://arxiv.org/abs/2204.07705>.

[12] XU Can,SUN Qingfeng,ZHENG Kai,*et al.* Wizardlm: Empowering large language models to follow complex instructions[EB/OL]. (2023-04-24)[2024-12-24]. <https://arxiv.org/abs/2304.12244>.

[13] LIU Hanmeng,TENG Zhiyang,CUI Leyang,*et al.* Logicot: Logical chain-of-thought instruction-tuning data collection with GPT-4[EB/OL]. (2023-10-28)[2024-12-24]. <https://arxiv.org/abs/2305.12147>.

[14] ZELIKMAN E,WU Yuhuai,MU J,*et al.* Star: Bootstrapping reasoning with reasoning[J]. Advances in Neural Information Processing Systems,2022,35:15476-15488.

[15] WANG Yizhong,KORDI Y,MISHRA S,*et al.* Self-instruct: Aligning language models with self-generated instructions[EB/OL]. (2022-12-21)[2024-12-24]. <https://arxiv.org/abs/2212.10560>.

[16] ZHOU Yongchao,MURESANU A I,HAN Ziwen,*et al.* Large language models are human-level prompt engineers [EB/OL]. (2022-11-03)[2024-12-24]. <https://arxiv.org/abs/2211.01910>.

[17] XU Canwen,GUO Daya,DUAN Nan,*et al.* Baize: An open-source chat model with parameter-efficient tuning on self-chat data[EB/OL]. (2023-04-03)[2024-12-24]. <https://arxiv.org/abs/2304.01196>.

(责任编辑: 钱筠 英文审校: 陈婧)