doi:10.6043/j.issn.0438-0479.202409024

融合词性提示学习的大语言模型依存句法分析

张振国[1,2],李 英[1,2*],余正涛[1,2],黄于欣[1,2]
(1.昆明理工大学信息工程与自动化学院,云南 昆明 650500;2.昆明理工大学云南省人工智能重点实验室,云南 昆明 650500)

摘要:[**目的**]针对大语言模型(LLMs)在依存句法分析任务上的性能尚未得到系统性探索的问题进行研究.[**方法**]首先设计多种提示策略对富资源语言中文和低资源语言越南语进行全面的 LLMs 依存句法分析性能评估,用于确定 LLMs 中蕴含句法知识的程度;然后将词性信息作为外部知识来构建提示模板,进一步激活 LLMs 对于词法和句法的理解能力,提高依存句法分析的性能.[**结果**]实验结果表明,LLMs 在富资源语言中文上的句法解析能力更强,而对于低资源语言越南语的解析能力较弱.通过对比实验,证明了外部词法知识的融入可以进一步提高其句法分析的能力,其中带标签依存分数(LAS)在中文上提升了 13.11%,在越南语上提升了 2.26%.[**结论**]设计合理的提示策略可以显著提升 LLMs 在句法分析任务上的表现,且在不同提示策略下,词性信息的加入对模型性能均产生了积极的影响.

关键词:大语言模型;依存句法分析;词性信息;提示策略
中图分类号:TP 391 **文献标志码**:A **文章编号**:0438-0479(2025)04-0606-10

Dependency parsing with large language models enhanced by part-of-speech prompting

ZHANG Zhenguo[1,2],LI Ying[1,2*],YU Zhengtao[1,2],HUANG Yuxin[1,2]
(1. Faculty of Information Engineering and Automation,Kunming University of Science and Technology,Kunming 650500,China;
2. Yunnan Key Laboratory of Artificial Intelligence,Kunming University of Science and Technology,Kunming 650500,China)

Abstract:[**Objective**] We aim to study the problem related to the performance of large language models (LLMs) on dependency parsing tasks. This problem has not been systematically explored. [**Methods**] In this study, we first design multiple prompting strategies to comprehensively evaluate the dependency parsing performance of LLMs on the rich-resource language Chinese and the low-resource language Vietnamese, in order to assess the extent of syntactic knowledge contained within LLMs. Then, part-of-speech information is used as the external knowledge to construct prompt templates, so that the LLM's ability can be further activated to understand morphology and syntax and to improve the performance of dependency parsing. [**Results**] Experimental results show that LLMs exhibit strong syntactic parsing capabilities on the rich-resource language Chinese, but weak parsing abilities on the low-resource language Vietnamese. Through comparative experiments, it is proved that the integration of external lexical knowledge can further improve its dependency parsing ability, among which the labeled dependency score (LAS) in Chinese is improved by 13.11% and in Vietnamese by 2.26%. [**Conclusions**] Designing appropriate prompt strategies can significantly improve the performance of LLMs in dependency parsing tasks. Under different prompt strategies, the addition of part-of-speech information exerts a positive impact on model performance.

收稿日期:2024-09-27 **录用日期**:2025-03-25
基金项目:国家自然科学基金(62306129,U21B2027,62366027,62266028);云南省基础研究计划项目(202401CF07121,202401BC070021,202301AS070047);云南省重大科技专项计划项目(202103AA080015,202202AD080003,202203AA080004);昆明理工大学"双一流"创建联合项目(202301BE070001-027,202201BE070001-021),云南省高新技术产业项目(201606)
通信作者:1224005374@qq.com
引文格式:张振国,李英,余正涛,等.融合词性提示学习的大语言模型依存句法分析[J].厦门大学学报(自然科学版),2025,64(4):606-615.
Citation: ZHANG Z G, LI Y, YU Z T, et al. Dependency parsing with large language models enhanced by part-of-speech prompting[J]. J Xiamen Univ Nat Sci,2025,64(4):606-615. (in Chinese)

Keywords: LLMs; dependency parsing; part-of-speech information; prompting strategy

依存句法分析是自然语言处理(NLP)的核心任务之一,旨在揭示句子中词语之间的逻辑和语义关系. 图 1 展示了通用依存树库中的一个依存树示例,其中从词"thích(喜欢)"到修饰词"cũng(也)"的有向边表示依存弧,标签"advmod"则描述了这两个词之间的依赖关系类型. 句法分析是理解语言结构的基础,对于机器翻译[1]、信息抽取[2]、语义角色标注[3]等 NLP 任务至关重要. 目前,监督依存解析模型在英语宾州树库上的带标签依存分数(LAS)和无标签依存分数(UAS)分别达到了 96.4%和 97.4%[4],在中文宾州树库上则达到了 92.5%和 93.5%[5]. 然而,这些监督解析模型在很大程度上依赖于训练数据的质量和规模[6],而对于资源匮乏的语言,如越南语,由于标注语料库的稀缺,模型的效果和性能仍不理想.

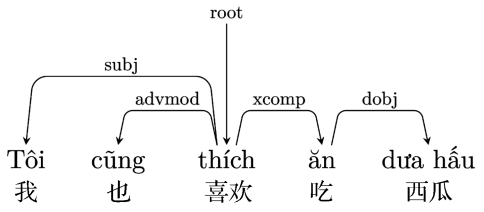

图 1　依存树示例

Fig. 1　An example of dependency tree

近年来,随着人工智能技术的飞速发展,大语言模型(large language models, LLMs)已成为 NLP 领域研究和应用的热点[7-8]. 这些模型以其卓越的语言处理能力,为自然语言理解[9](natural language understanding, NLU)和自然语言生成[10](natural language generation, NLG)领域带来了革命性的进步. 然而,这些模型在依存句法分析任务上的性能尚未得到充分评估和验证. 传统的句法分析方法通常依赖于手工设计的语法规则或统计模型[11-12],还需要大量的标注数据,这在低资源语言中难以获取. 而 LLMs 的出现为这一领域带来了新的视角和机遇[13]. 如何在有限的条件下提升 LLMs 的句法分析能力,已成为研究的热点问题.

为了解决这一问题,研究者们开始尝试通过指令和少样本数据来激活模型的句法性能[14]. 这种方法的核心在于利用少量的示例数据和明确的指令,引导模型更好地理解和应用句法规则[15]. 尽管一些初步研究显示了这一方法的有效性,但系统性的研究和验证仍然不足. 本文旨在填补这一空白,通过一系列实验探究不同 LLMs 在依存句法分析任务上的表现,并探索不同的提示策略对模型性能的影响,以期为未来的研究提供帮助.

本文使用零样本提示、单样本提示、混合提示以及少样本提示 4 种不同的提示策略对不同的 LLMs 在依存句法分析任务上的表现进行测试,以评估它们在不同训练条件下的表现. 此外,本文还进行了相关实验来验证融合词性提示学习是否能够进一步提升句法分析的性能. 实验结果表明,为 LLMs 提供适当的样本可以显著提升它们在句法解析任务上的表现. 不论是使用哪一种提示策略,词性信息都能在一定程度上提升模型的解析性.

本文贡献如下:

1) 系统性地评估了不同 LLMs 在依存句法分析任务上的表现,填补了现有研究领域的空白.

2) 通过比较零样本提示、单样本提示、混合提示和少样本提示等不同策略,深入分析不同策略在依存句法分析任务中的表现.

3) 探索将词性信息作为额外输入对 LLMs 在依存句法分析任务中性能的影响,展示了其在提升模型解析能力方面的潜力.

1　相关工作

1.1　基于预训练语言模型的方法

随着神经网络和预训练语言模型的发展,越来越多的研究者将这些技术应用于依存分析模型[16]. Kondratyuk 和 Straka[17]开发了一种在 BERT 上进行微调的多语言多任务自注意力网络,在多种语言上取得了良好的结果. Al-ghamdi 等[18]深入研究了针对阿拉伯语依存解析的 BERT 模型的微调,为针对特定 NLP 任务微调 BERT 模型提供了参考. Fernández-González 和 Gómez-Rodríguez[19]开发了一种基于指针网络的自下而上的层级依存句法分析器,并提出了两种新的基于转移的算法,包括一种从右到左解析句子的方法和一种从外向内解析的方法. Le-Hong 和 Cambria[20]提出了一种通过集成图嵌入模型来改进基于转移的依存句法分析的方法,该方法通过结合图嵌入算法训练的节点嵌入到双向循环神经网络架构中,显著提升了在多种语言上的依存句法分析性能. DO 等[21]通过欧几里得距离过滤算法选择与越南语句法相似的英语句子,并结合双仿射注意力机制提高了低资源语言解析模型的准确性. Thien 等[22]使用双向长

短期记忆网络(BiLSTM)和图神经网络(GNN)来提取上下文信息和高阶信息,并采用优化的词嵌入技术处理未知词汇,显著提升了句法分析的准确性. Nguyen 等[23]开发了首个为越南语预训练的大规模单语种语言模型 PhoBERT,该模型在词性标注、依存句法分析、命名实体识别等任务上表现出了卓越的性能. 尽管这些方法在依存句法分析中取得了一定的成功,但它们通常依赖于复杂的模型结构和大量的训练数据,尤其是在低资源语言的应用中,性能提升仍然有限.

1.2 基于 LLMs 的方法

近期研究表明,LLMs 在提升 NLP 任务性能方面表现显著,展示出了卓越的指令遵循能力,并已成功应用于多种 NLP 任务[24-25]. Lin 等[26]探索了 LLMs 在零样本情况下的依存句法分析能力,实验表明 LLMs 作为零样本依存句法分析器具有一定的潜力. Li 等[27]设计了一种利用 LLMs 增强的自训练方法,用于跨领域成分句法分析,通过迭代生成特定领域的原始语料库,并结合语法规则和置信度标准选择伪实例,显著提高了跨领域句法分析的性能. Chen 等[28]提出了一种基于条件互信息的无监督依存句法分析方法,通过整合语法约束和词性信息,提高了 LLMs 在多种语言上的依存句法分析性能. 现有的 LLMs 方法大多专注于零样本或自监督学习,虽然在资源丰富语言上取得了优异的成绩,但在低资源语言上仍然面临着性能瓶颈. 鉴于 LLMs 在生成和推理方面所展现的能力,本文尝试通过设计差异化的提示策略,并结合词性提示学习方法探索 LLMs 在中文和越南语依存句法分析中的应用,期望利用这些模型强大的的生成能力,来探索其在不同资源语言环境下的应用潜力.

2 融合词性提示学习的 LLMs 依存句法分析方法

2.1 模型

在当前人工智能领域,大型语言模型已成为主流趋势. 考虑到生成结果的准确性、处理速度、成本效益以及模型的泛化能力等多个关键因素,本文选取了以下开源及闭源模型进行评估,包括 ChatGPT-4o-mini①、ChatGPT-4o②、moonshot-v1-8k③ 和 GLM-4-0520④. 这些模型均采用最新的架构设计,具有卓越的语言理解和生成能力. 本文对这些模型在依存句法分析任务上的表现进行了综合评估.

2.2 提示策略

依存句法分析的结果必须严格遵循既定的标准格式. 生成结果的质量直接影响到后处理工作的难度,这一要求对本文的工作构成了挑战. 因此,精心设计不同的提示指令对于引导 LLMs 生成高质量的句法分析结果至关重要. 高质量的输出不仅确保了结果的准确性和可用性,同时还降低了后处理过程的复杂性. 为了进一步挖掘模型在句法分析任务中的潜力,本文采用了一系列差异化的提示策略,并结合词性提示学习方法来增强 LLMs 在依存句法分析中的表现. 词性提示学习通过将词性信息作为外部知识融入 LLMs 的输入提示,帮助模型更好地理解词法和句法结构. 本文所使用的提示策略如下,并在附录中展示了设计的提示指令:

1) 零样本提示(zero-shot prompting):不使用任何示例,直接将句子输入模型进行句法分析. 提示内容包括对越南语句法分析任务的简要说明以及期望的输出格式,模型根据其理解的任务要求生成解析结果.

2) 单样本提示(one-shot prompting):提供一个示例句子及其解析结果作为指导,以帮助模型理解任务. 示例句子的选择旨在覆盖多种句法结构.

3) 混合提示(hybrid prompting):结合单样本提示和通用模板,使用不同的提示组合来提高模型的解析能力. 通用模板提供了一套标准化的依存句法分析步骤和格式要求,用于指导模型对越南语句子进行解析.

4) 少样本提示(few-shot prompting):提供多个示例句子及其解析结果作为指导.

2.3 LLMs 依存句法分析

本文采用了多种提示策略来使用 LLMs 进行依存句法分析,如图 2 所示. 首先,将测试集中的句子输入到预先设计的提示模板中,让 LLMs 能够依据给定的格式和指导进行准确的句法分析. 本文所设计的提示模板不仅包含了清晰的指令,还特别强调了输出结果的预期格式,用来引导模型生成结构化且符合 CoNLL-U 标准的解析结果. CoNLL-U 格式是一种用于标注和表示依存句法分析结果的标准格式. 模型在输出过程中会遇到格式相关的问题,包括单词遗漏、分词错误以及输出乱码等. 为了应对这些问题,本文

① https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/
② https://openai.com/index/hello-gpt-4o/
③ https://platform.moonshot.cn/
④ https://zhipuai.cn/

实施了一系列细致的后处理措施,这些措施包括对模型输出的详尽检查、错误更正以及必要的数据清洗,确保了输出数据的准确性和可用性.

最后,将模型生成的解析结果与标准结果进行对比,得到 LAS 和 UAS.

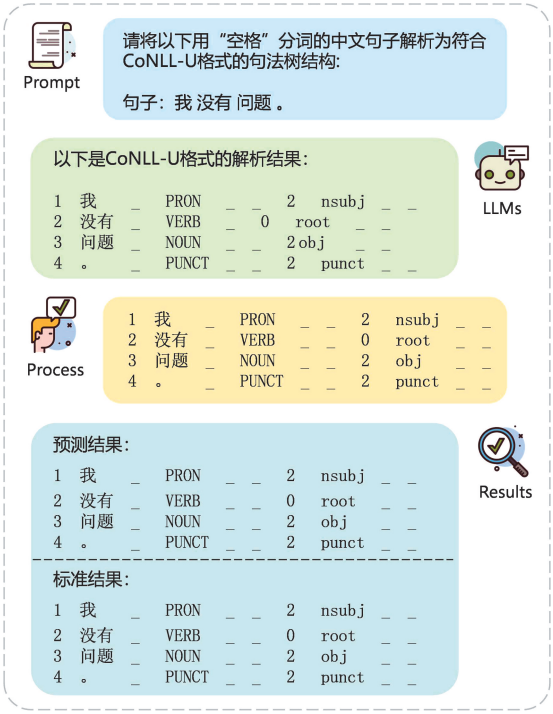

图 2　LLMs 依存句法分析方法

Fig. 2　LLMs dependency parsing method

3　实验设置

3.1　数据集

为了确保对 LLMs 在依存句法分析任务上的性能进行公正且准确的比较,本文选择了公开数据集 Universal Dependencies(UD)2.14[①] 版本中的越南语 VTB 树库和中文 Beginner 树库测试集作为实验数据. VTB 树库的测试集包含 800 条越南语句子. Beginner 树库的测试集则包含了 2 295 条汉语句子,这些句子根据其结构的复杂性和难度被细致地划分为 A1、A2、B1、B2 和 C1 5 个不同的级别. 考虑到模型输出可能存在一定的随机性,为了减少随机性对实验结果的影响,本文在所有实验中将 LLMs 的温度参数统一设置为 0.

3.2　评价指标

本文使用标准的 UAS 和 LAS 来作为依存句法分析的评价指标. UAS 仅考虑依存弧的准确性,而不考虑依存关系的类型. 相对地,LAS 则进一步考虑了依存关系类型的准确性.

4　实验分析

4.1　对比试验

本文采用了 4 种不同的提示策略来进行对比试验. 对于少样本提示,分别使用不同的样本数 k 进行测试,当 $k=10$ 时,模型性能达到最佳状态,因此最终选择 $k=10$ 作为少样本提示的设置. 考虑到 moonshot-v1-8k 和 GLM-4-0520 对越南语数据集上结果的可用性较差,因此实验重点评估了这两个模型在中文数据集上的性能. 此外,为了更全面地评估模型性能,还考虑了词性(POS)信息的影响,每种提示策略均在有无词性信息的情况下进行实验.

表 1 的实验结果显示,ChatGPT-4o 在所有提示策略下均表现出较高的性能,尤其在少样本提示下,将词性信息作为输入时,其在越南语 UD 数据集上的 UAS 达到了最高的 50.90%,在中文数据集上的 UAS 则达到了最高的 60.69%. 另外,作为中文 LLMs 的 Moonshot-v1-8k 和 GLM-4-0520,在中文数据集上同样展现出了令人印象深刻的性能,尤其是在混合提示和少样本提示下,它们的性能提升进一步验证了精心设计的提示指令在增强模型性能方面的关键作用.

在零样本提示下,意味着模型在没有接受任何特定领域样本训练的情况下进行分析,所有模型的初始性能相对较低,这反映出在没有示例的情况下,模型对任务的理解存在局限. 然而,随着提示策略的改进,尤其是引入单样本和少样本提示后,模型的性能得到了显著提升. 在没有词性信息输入的情况下,与零样本提示相比,仅提供一个示例给模型时,模型在 LAS 和 UAS 上的得分都有显著提升. 具体来说,在中文数据集上,ChatGPT-4o 使用单样本提示相比于零样本提示,LAS 得分提高了 8.77 个百分点,UAS 得分提高了 9.02 个百分点. 实验结果证明了 LLMs 自身蕴含着丰富的句法知识,通过激活模型自身的句法信息,能够有效提升其性能. 同时,适量的示例也能够有效地引导模型更好地学习和泛化句法结构,从而在不同语言的数据集上实现较高的准确率.

① https://universaldependencies.org/

表 1 LLMs 使用不同提示策略的评测结果,其中"w/o POS"表示在提示策略中未加入词性信息."w/POS"表示加入词性信息

Tab. 1 Evaluation results of LLMs using different prompting strategies,"w/o POS" means that part-of-speech(POS) information is not included in the prompt strategy."w/POS"means that part of speech information is included %

LLMs	提示策略	UD_Vietnamese-VTB				UD_Chinese-Beginner			
		LAS	UAS	LAS	UAS	LAS	UAS	LAS	UAS
		w/o POS		w/POS		w/o POS		w/POS	
ChatGPT-4o	零样本提示	26.48	46.24	28.54	47.44	34.86	49.94	42.89	55.75
	单样本提示	28.74	48.70	29.43	49.17	43.63	**58.96**	50.31	**64.15**
	混合提示	29.79	49.63	30.74	49.97	44.13	58.84	47.10	59.58
	少样本提示	**32.07**	**50.57**	**34.33**	**50.90**	**45.12**	55.87	**51.42**	60.69
ChatGPT-4o-mini	零样本提示	13.57	34.07	15.50	35.40	17.16	36.05	17.06	35.85
	单样本提示	16.97	35.46	17.17	36.19	22.13	40.05	20.89	36.46
	混合提示	21.49	42.71	22.69	43.91	21.63	39.93	24.72	43.51
	少样本提示	**27.35**	**47.06**	**28.81**	**47.97**	**33.25**	**44.62**	**37.08**	**47.34**
Moonshot-v1-8k	零样本提示	—	—	—	—	32.88	45.36	40.05	53.03
	单样本提示	—	—	—	—	38.44	51.55	41.90	53.40
	混合提示	—	—	—	—	39.06	**53.15**	40.42	51.79
	少样本提示	—	—	—	—	**40.54**	51.55	**44.62**	**54.64**
GLM-4-0520	零样本提示	—	—	—	—	28.43	**45.36**	38.20	53.03
	单样本提示	—	—	—	—	30.66	44.25	41.41	**53.13**
	混合提示	—	—	—	—	31.15	44.38	38.81	50.68
	少样本提示	—	—	—	—	**31.64**	40.54	**44.75**	52.41

混合提示是在单样本提示的基础上增加了通用模板,该模板提供了一套标准化的依存句法分析步骤和格式要求.这一设计有助于模型更好地理解提示要求并执行任务.从表 1 的实验结果可以看出,相较于单样本提示,混合提示的分数皆有不同程度的增加,说明通用模板在提升模型性能方面的有效性.

在深入分析表 1 所示的实验数据后,引入词性信息显著地影响了 LLMs 的性能表现.具体而言,当模型在处理带有词性信息的句子时,其性能表现普遍优于未引入词性信息的句子.其中越南语数据集在 ChatGPT-4o 上采用少样本提示时,LAS 提升了 2.26%;中文数据集在 GLM-4-0520 上采用少样本提示时,LAS 提升了 13.11%.词性信息的引入显著提升了模型在句法解析中的准确度,因为它提供了相关的语义和句法线索,帮助模型更好地识别和分析句子结构,有助于其在处理复杂语言现象时做出更加准确的分析.鉴于此,未来的研究可以进一步探讨如何更有效地整合词性信息,以进一步优化 LLMs 的性能.

4.2 消融实验

为了深入探讨了词性信息对模型性能的影响,本文在越南语和中文数据集上进行了消融实验.在越南语数据集上,使用 ChatGPT-4o 和 ChatGPT-4o-mini 模型进行词性信息的消融实验.在中文数据集上,则使用 moonshot-v1-8k 和 GLM-4-0520 模型进行相关的消融实验.如图 3 所示,实验首先将越南语数据集中的句子按照长度划分为 6～8、9～11、12～14、15～17、18～20 和 21～23 共 6 个区间,接着对模型在有无词性信息输入的情况下,计算了每个区间在使用不同提示策略情况下的 LAS.从实验结果来看,不论是使用哪种提示策略,在有词性信息输入的情况下模型性能普遍优于无词性信息的情况.尤其是在句子长度为 6～8 和 9～11 这两组数据上,其优势更为明显.然而,当句子长度大于 18 时,有无词性信息对于模型性能的提升却不显著.

图 4 展示了在中文数据集上的实验结果.数据集按照句子长度被划分为 2～4、5～6、7～8、9～10、11～12 及长度大于 12 的 6 个长度区间,并在不同提示策

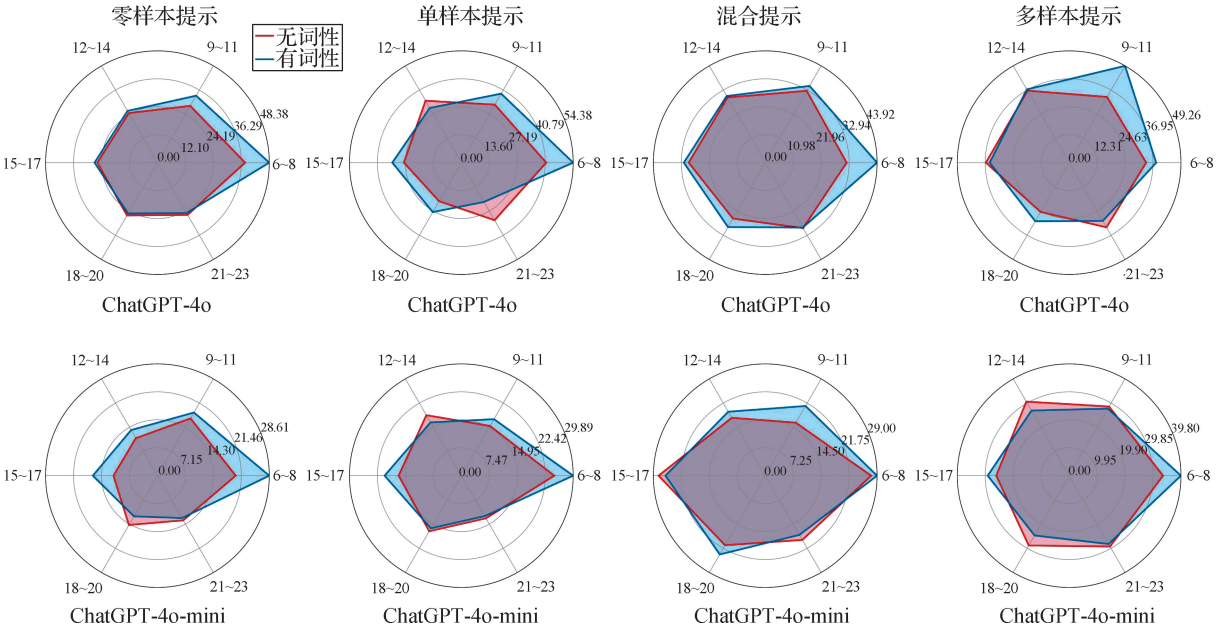

图 3　在 UD_Vietnamese-VTB 数据集上进行消融实验

Fig. 3　Conducting ablation experiments on the UD_Vietnamese-VTB dataset

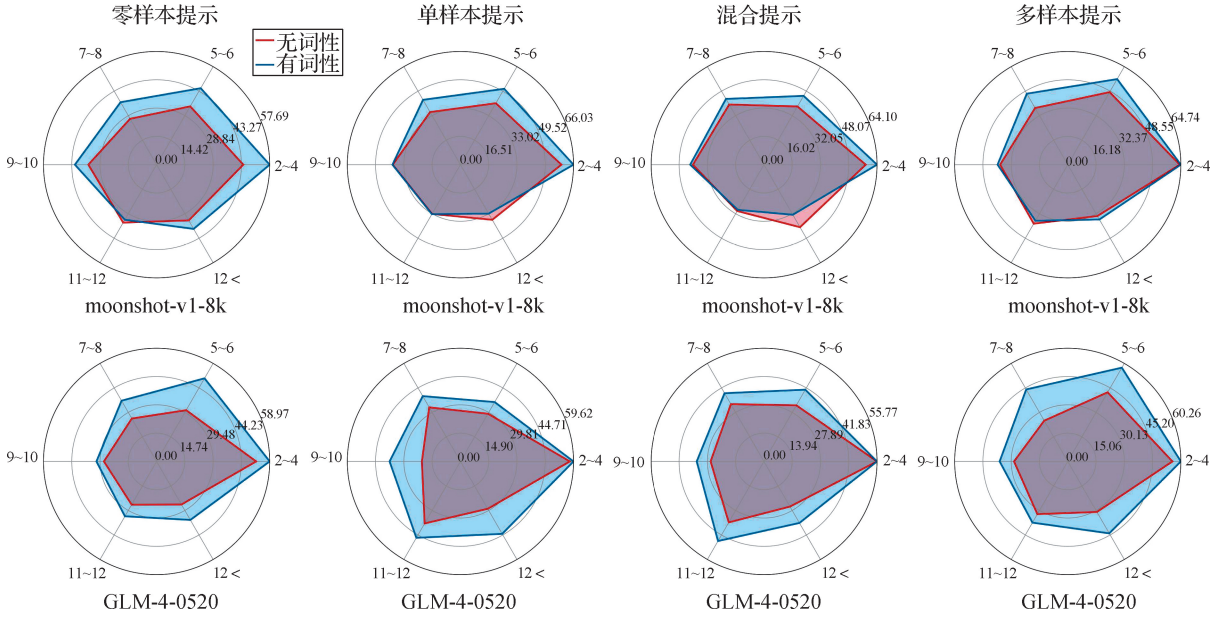

图 4　在 UD_Chinese-Beginner 数据集上进行消融实验

Fig. 4　Conducting ablation experiments on the UD_Chinese-Beginner dataset

略下计算了每个区间的 LAS. 结果显示,无论使用哪种提示策略,词性信息加入对于模型性能均有着明显的提升,尤其是在句子长度较短的区间上,LAS 都得到了很大的提升. 然而模型在处理长句时词性信息的帮助却很有限,这一情况可能是多方面因素所导致. 首先,在长句子的处理过程中,LLMs 的注意力机制需要将其注意力分配到更多的词汇和上下文中,这导致了注意力分配的稀释现象,即每个词汇的注意力权重减少,进而使得词性信息对句子解析的影响减弱. 相比之下,短句的上下文较为紧凑,模型能够将更多的注意力集中在每个词汇上,从而使得词性信息的作用更加显著. 此外,对于长句子来说,模型可能已经能够从上下文中获取足够的信息来进行解析,此时,词性信息的边际效益会减少,因为模型的主要挑战在于理解句子的长范围依赖关系,而不是依赖于单个词的词性. 在长句子处理过程中,提示中的额外信息可能占

据了模型的计算资源,使得分配给词性标注的信息量减少,影响了模型的性能. 而短句子的句法结构较简单,提供额外的词性信息对句法分析具有更显著的影响.

4.3 分析实验

本文设计了详细的分析实验来深入探究不同提示策略对模型性能的影响. 首先,实验统计了中文 Beginner 数据集中出现频率最高的 6 种标签,接着分析了 LLMs 在采用不同的提示策略时在这 6 种标签上的 LAS,对比了模型在有词性信息输入下零样本提示、单样本提示、少样本提示和混合提示策略下的表现,结果如图 5 所示.

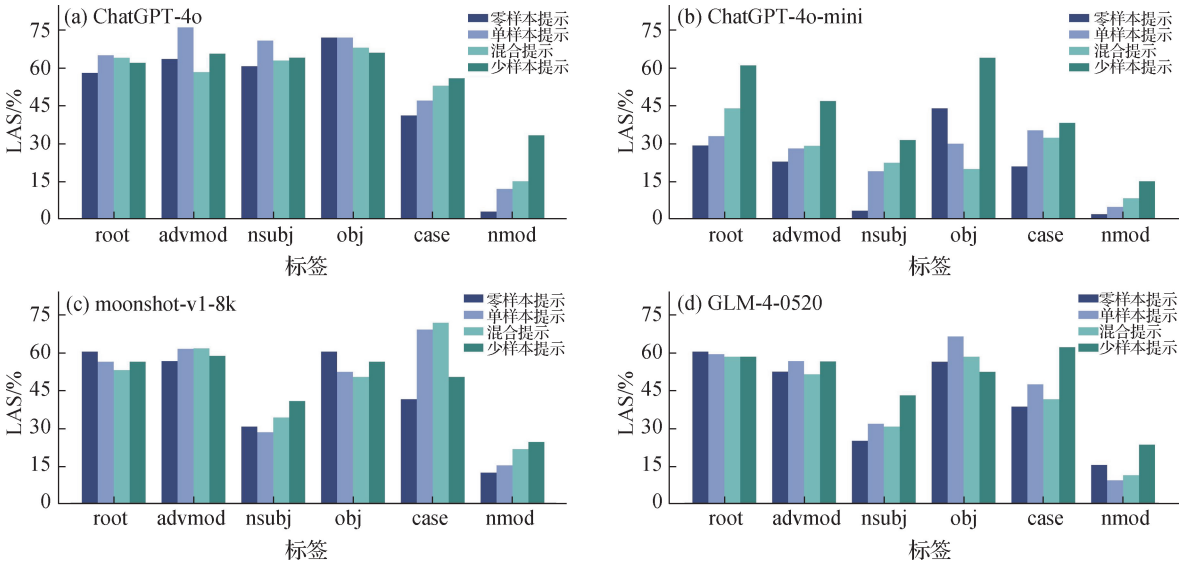

图 5 LLMs 在不同标签上的 LAS

Fig. 5 LAS of LLMs on different labels

首先,ChatGPT-4o 在所有提示策略下都展现出了卓越的性能. 这一结果表明,无论是否提供样本,ChatGPT-4o 都能够稳定地进行高质量的句法分析,特别是在各种提示策略之间,性能差异并不显著,这进一步证实了该模型在处理依存句法任务时的一致性和稳定性. 然而,在频率最低的"nmod"标签上,少样本提示的效果显著优于其他 3 种提示策略,证明少样本提示能够更好地挖掘和利用样本中的特征,提升模型的理解能力. 与此相比,对于 ChatGPT-4o-mini 模型,少样本提示在所有标签上都表现出了最佳的结果. 在较小模型中少样本提示能够有效地促进模型对样本知识的学习,从而提高其在依存句法分析中的表现.

在对两个 LLMs 的分析中,实验发现不同提示策略的效果各有高低. 这可能与模型的内部结构、训练数据和算法优化有关. 但有一个显著的共同点是,对于频率最低的"nmod"标签,4 个 LLMs 均展现了少样本提示的明显优势. 说明少样本提示在处理出现频率低的标签时具有一定的提升效果,可能是由于少样本提示能够有效地激发模型对这些标签的深入学习和理解.

此外,本文在越南语数据集和中文数据集上也进行了不同句子长度的分析实验,实验首先将数据集按照句子长度划分为 6 个区间,随后对 4 个 LLMs 在各个区间上的 LAS 进行了细致的统计分析,如表 2 和表 3 所示. 结果表明,本文设计的少样本提示在所有长度区间内均表现出色,尤其是在 ChatGPT-4o-mini 上,少样本提示的优势尤为明显,在所有区间都取得了最佳的结果. 对于短句来说,少样本提示的表现显著优于其他策略. 而零样本提示的表现最差,这与预期一致,因为未提供示例输入,模型无法有效激活其潜在的句法知识. 但给模型提供适当的示例,即便仅有一个示例,其性能亦能得到显著提升,这得益于 LLMs 强大的学习能力. 混合提示在单样本提示的基础上进一步提供了一套标准化的依存句法分析步骤和格式要求,其有效性也得到了体现.

整体来看,模型对长句的解析效果远不如短句. 无论是 LLMs 还是传统预训练语言模型,在处理长句时效果都欠佳. 尽管使用少样本提示策略可以在一定程度上提升长句的解析精度,但提升效果依然有限. 这一现象不仅是 LLMs 所面临的挑战,也是传统预训练语言模型普遍存在的难题. 长句的复杂性在于其句法结构的复杂性和依赖关系的多样性,这些因素共同增加了模型解析的难度. 此外,长句中可能存在的信

表 2　LLMs 在越南语数据集中不同句子长度上的 LAS

Tab. 2　LAS of LLMs on different sentence lengths in Vietnamese　　%

LLMs	提示策略	UD_Vietnamese-VTB/Sentence_Length					
		6～8	9～11	12～14	15～17	18～20	21～23
ChatGPT-4o	零样本提示	61.24	**52.38**	51.07	31.85	33.18	25.51
	单样本提示	62.17	45.24	**64.46**	41.11	35.61	27.43
	混合提示	62.89	39.52	49.82	33.95	30.68	29.51
	少样本提示	**63.70**	41.90	40.36	**45.31**	**49.85**	**31.83**
ChatGPT-40-mini	零样本提示	12.50	15.71	23.04	19.26	13.86	12.16
	单样本提示	45.83	26.67	35.18	17.04	19.09	15.75
	混合提示	46.50	18.10	31.79	19.63	13.71	21.89
	少样本提示	**47.83**	**35.71**	**40.71**	**28.77**	**25.23**	**26.55**

表 3　LLMs 在中文数据集中不同句子长度上的 LAS

Tab. 3　LAS of LLMs on different sentence lengths in Chinese　　%

LLMs	提示策略	UD_Chinese-Beginner/Sentence_Length					
		2～4	5～6	7～8	9～10	11～12	＞12
ChatGPT-4o	零样本提示	55.77	50.14	47.25	42.59	33.00	35.03
	单样本提示	57.69	57.22	**61.24**	38.33	46.63	**46.17**
	混合提示	58.93	51.67	55.21	35.25	**49.41**	43.48
	少样本提示	**65.38**	**57.92**	60.42	**48.09**	44.70	41.23
ChatGPT-40-mini	零样本提示	26.92	19.44	15.77	14.81	17.68	16.31
	单样本提示	29.49	20.83	21.58	19.75	21.38	20.16
	混合提示	27.56	31.25	23.59	22.59	24.33	24.37
	少样本提示	**52.56**	**46.81**	**46.73**	**29.20**	**26.18**	**30.29**
Moonshot-v1-8k	零样本提示	57.69	45.00	36.76	41.67	32.41	37.81
	单样本提示	**66.03**	51.11	43.68	39.69	33.33	32.99
	混合提示	64.10	45.00	43.01	**41.79**	29.55	32.75
	少样本提示	64.74	**56.53**	**46.88**	40.19	**37.04**	36.21
GLM-4-0520	零样本提示	58.97	50.00	36.53	31.36	33.00	35.20
	单样本提示	59.62	36.11	39.73	**37.35**	**46.46**	44.05
	混合提示	55.77	40.83	38.84	33.15	45.29	35.01
	少样本提示	**60.26**	**57.64**	**44.27**	36.30	37.63	**44.22**

息冗余和深层嵌套结构,对模型的精确度构成了额外的挑战. 因此,如何提高长句解析的准确性,依然是句法分析领域的重要问题.

4.4　案例分析

本文对不同语言模型在句法分析任务上的表现进行了深入的案例分析. 实验首先观察了 4 种模型在不同提示策略下的解析效果,每个模型都针对其测试案例进行了评估. 如图 6 所示,图中的第一行是正确的结果,红色代表模型输出的错误结果. 模型在零样本提示的情况下展现出对句法结构的基本理解,但依然存在一些偏差,普遍存在比较严重的错误,甚至在 ChatGPT-4o-mini 中存在“root”节点错误的情况. 进一步地,单样本提示为模型提供了一个具体的示例,这有助于模型更好地理解句子结构. 模型的解析效果有略微的提升,但仍然存在比较多的错误. 混合提示在单样本提示的基础上增加了一套标准化的依存句法分析步骤和格式要求,从具体的输出结果可以看出模型在理解上达到一个新的水平. 在这种策略下,模型的输出更加多样化,同时也更接近正确的句法结构. 最后,少样本提示为模型提供了更多的信息,得到的结果是最好的,句子的结构也是最准确的.

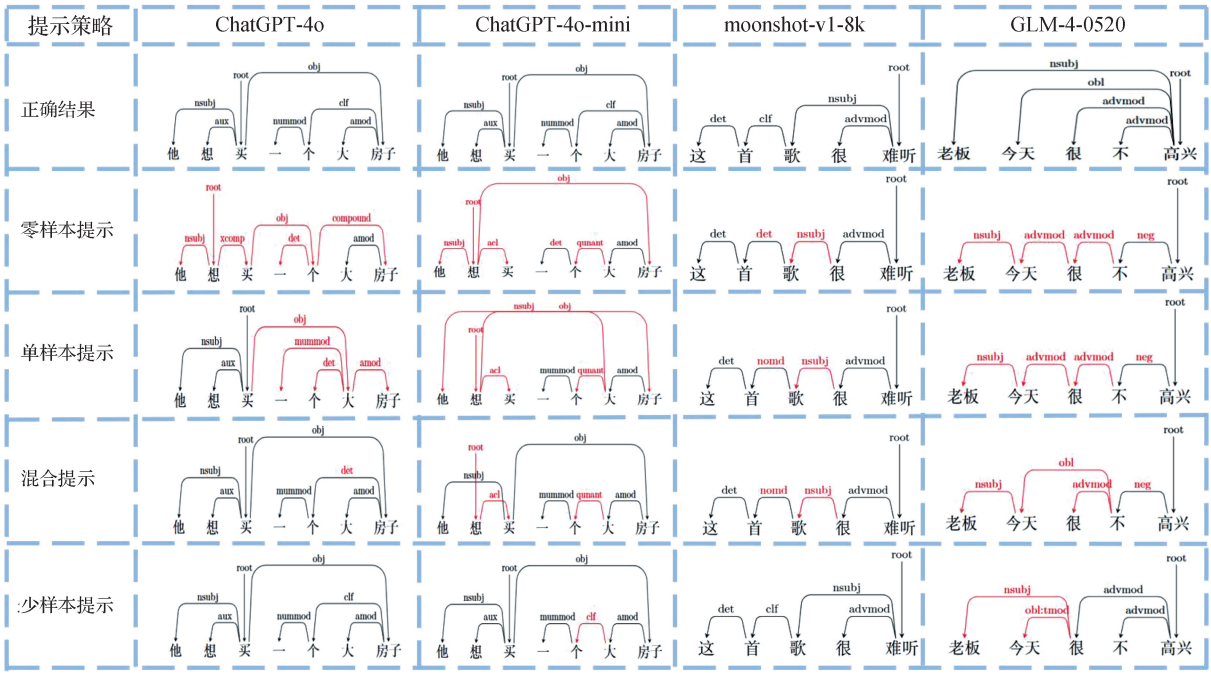

图 6 案例分析

Fig. 6 Case study

总的来说,随着提示样本数量的增加,模型的解析效果呈现出逐步提升的趋势. 这表明适当的提示策略可以显著提高模型的句法分析能力,尤其是在处理复杂语言结构时. 然而,这也说明在设计提示策略时需要细致考虑,以确保模型能够在保持准确性的同时,提高其泛化能力.

5 结 论

本文深入探究了多个 LLMs 在依存句法分析任务上的表现. 通过采用零样本提示、单样本提示、混合提示和少样本提示 4 种提示策略,结合词性提示学习方法,构建合理有效的提示模板,从而进一步增强 LLMs 在句法分析任务上的表现. 实验结果表明,LLMs 已具备一定的句法知识,合适的提示模板可以显著提升其在句法分析任务上的表现. 此外,无论采用哪种提示策略,词性信息的引入均对模型性能产生了积极的影响. 在未来工作中,我们将继续优化提示策略,并探索更多外部知识的融合方式,以提高模型在复杂句子解析和低资源语言分析中的表现.

参考文献:

[1] GONG L C, LI Y, GUO J J, et al. Enhancing low-resource neural machine translation with syntax-graph guided self-attention [J]. Knowledge-Based Systems, 2022, 246: 108615.

[2] LI Z C, PARNOW K, ZHAO H. Incorporating rich syntax information in grammatical error correction [J]. Information Processing & Management, 2022, 59(3): 102891.

[3] ZHANG Y, XIA Q R, ZHOU S L, et al. Semantic role labeling as dependency parsing: exploring latent tree structures inside arguments [EB/OL]. [2024-09-01]. https://arxiv.org/abs/2110.06865v2.

[4] AMINI A, LIU T Y, COTTERELL R. Hexatagging: projective dependency parsing as tagging [C]// Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Toronto: ACL, 2023: 1453-1464.

[5] YANG S L, TU K W. Headed-span-based projective dependency parsing [C]// Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin: ACL, 2022: 2188-2200.

[6] ZHANG S, WANG L J, SUN K, et al. A practical Chinese dependency parser based on A large-scale dataset [EB/OL]. [2024-09-01]. https://arxiv.org/abs/2009.00901v2.

[7] QIN L B, CHEN Q G, FENG X C, et al. Large language models meet NLP: a survey [EB/OL]. [2024-09-01]. https://arxiv.org/abs/2405.12819v1.

[8] HUANG H, WU S Z, LIANG X N, et al. Towards making the Most of LLM for Translation quality estimation

[C] // Natural Language Processing and Chinese Computing. Cham: Springer Nature Switzerland, 2023: 375-386.

[9] LIN W Y. Prototyping a chatbot for site managers using building information modeling (BIM) and natural language understanding (NLU) techniques[J]. Sensors, 2023, 23(6): 2942.

[10] WANG J A, LIANG Y L, MENG F D, et al. Is ChatGPT a good NLG evaluator? A preliminary study[C] // Proceedings of the 4th New Frontiers in Summarization Workshop. Hybrid: ACL, 2023: 1-11.

[11] LI Z C, CAI J X, HE S X, et al. Seq2seq dependency parsing[C] // Proceedings of the 27th International Conference on Computational Linguistics. Santa Fe: ICCL, 2018: 3203-3214.

[12] LI Z C, ZHAO H, PARNOW K. Global greedy dependency parsing[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(5): 8319-8326.

[13] HROMEI C D, CROCE D, BASILI R. U-DepPLLaMA: universal dependency parsing via auto-regressive large language models[J]. Italian Journal of Computational Linguistics, 2024, 10: 1-10.

[14] ZHANG M S, JIANG G Y, LIU S, et al. LLM-assisted data augmentation for Chinese dialogue-level dependency parsing[J]. Computational Linguistics, 2024, 50(3): 867-891.

[15] DENG H X, ZHANG X, ZHANG M S, et al. Holistic exploration on universal decompositional semantic parsing: architecture, data augmentation, and LLM paradigm[EB/OL]. [2024-09-01]. https: // arxiv. org/abs/2307. 13424v1.

[16] DOZAT T, MANNING C D. Deep biaffine attention for neural dependency parsing[EB/OL]. [2024-09-01]. https: // arxiv. org/abs/1611. 01734v3.

[17] KONDRATYUK D, STRAKA M. 75 languages, 1 model: parsing universal dependencies universally[EB/OL]. [2024-09-01]. https: // arxiv. org/abs/1904. 02099v3.

[18] AL-GHAMDI S, AL-KHALIFA H, AL-SALMAN A. Fine-tuning BERT-based pre-trained models for Arabic dependency parsing[J]. Applied Sciences, 2023, 13(7): 4225.

[19] FERNÁNDEZ-GONZÁLEZ D, GÓMEZ-RODRÍGUEZ C. Dependency parsing with bottom-up hierarchical pointer networks[J]. Information Fusion, 2023, 91: 494-503.

[20] LE-HONG P, CAMBRIA E. Integrating graph embedding and neural models for improving transition-based dependency parsing[J]. Neural Computing and Applications, 2024, 36(6): 2999-3016.

[21] DO D, DINH D, LUONG A V, et al. Adapting cross-lingual model toImprove Vietnamese dependency parsing[C] // Artificial Intelligence in Data and Big Data Processing. Cham: Springer International Publishing, 2022: 97-108.

[22] THIEN N D, TRANG N T T, QUANG T D. Applying graph neural networks for vietnamese dependency parsing[C] // Proceedings of the 7th International Workshop on Vietnamese Language and Speech Processing. [S. l.]: IWVLSP, 2020: 54-59.

[23] NGUYEN D Q, NGUYEN A T. PhoBERT: pre-trained language models for Vietnamese[EB/OL]. [2024-09-01]. https: // arxiv. org/abs/2003. 00744v3.

[24] NASUTION A H, ONAN A. ChatGPT label: comparing the quality of human-generated and LLM-generated annotations in low-resource language NLP tasks[J]. IEEE Access, 2024, 12: 71876-71900.

[25] BANG F. GPTCache: an open-source semantic cache for LLM applications enabling faster answers and cost savings[C] // Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS 2023). Singapore: Empirical Methods in Natural Language Processing, 2023: 212-218.

[26] LIN B D, ZHOU X Y, TANG B H, et al. ChatGPT is a potential zero-shot dependency parser[EB/OL]. [2024-09-01]. https: // arxiv. org/abs/2310. 16654v1.

[27] LI J L, ZHANG M S, GUO P M, et al. LLM-enhanced self-training for cross-domain constituency parsing[EB/OL]. [2024-09-01]. https: // arxiv. org/abs/2311. 02660v1.

[28] CHEN J J, HE X H, MIYAO Y. Language model based unsupervised dependency parsing with conditional mutual information and grammatical constraints[C] // Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). Stroudsburg: ACL, 2024: 6355-6366.

(责任编辑:汪　军)