计算机科学
COMPUTER SCIENCE

基于大小语言模型协同增强的中文电子病历依存句法分析

许思遥, 曾健骏, 张维彦, 叶琪, 朱焱

引用本文

许思遥, 曾健骏, 张维彦, 叶琪, 朱焱. 基于大小语言模型协同增强的中文电子病历依存句法分析[J]. 计算机科学, 2025, 52(2): 253-260.

XU Siyao, ZENG Jianjun, ZHANG Weiyan, YE Qi, ZHU Yan. Dependency Parsing for Chinese Electronic Medical Record Enhanced by Dual-scale Collaboration of Large and Small Language Models [J]. Computer Science, 2025, 52(2): 253-260.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.231200054

基于大小语言模型协同增强的中文电子病历依存句法分析

许思遥[1]　曾健骏[2]　张维彦[2]　叶　琪[2]　朱　焱[1]
1 华东理工大学数学学院　上海 200237
2 华东理工大学信息科学与工程学院　上海 200237
(18817535379@163.com)

摘　要　依存句法分析是一项重要的自然语言处理任务,其目标是识别句子中词与词之间的依存关系。但在面向中文医疗电子病历的依存句法分析中,现有的研究存在以下问题:当出现缺省指示语法结构的成分和修饰成分位置多样的情况时,当前的通用解析器无法准确分析。针对该问题,提出基于大小语言模型协同增强的中文电子病历依存句法分析方法。首先,分析中文电子病历的语言特征,提出通过成分补全指示医疗文本中的特殊语法结构。然后,利用通用解析器进行依存句法分析,对于解析后的语法图,利用大语言模型的先验语法知识进行自动修正。此外,所提方法将重点放在缩小医疗文本与通用文本之间的特征分布差异上,故不受医疗领域缺少标注数据的限制。针对中文电子病历的依存句法分析,标注了444条测试样本,并对所提方法进行验证。实验表明该方法能有效地对中文电子病历进行依存分析,基于少量标注语料,LAS指标可达92.42,UAS指标可达94.60,并且在不同科室的中文电子病历上也能够达到同样显著的效果。

关键词:自然语言处理;依存句法分析;中文电子病历;大语言模型;协同增强
中图分类号　TP391

Dependency Parsing for Chinese Electronic Medical Record Enhanced by Dual-scale Collaboration of Large and Small Language Models

XU Siyao[1],ZENG Jianjun[2],ZHANG Weiyan[2],YE Qi[2] and ZHU Yan[1]
1 School of Mathematics,East China University of Science and Technology,Shanghai 200237,China
2 School of Information Science and Engineering,East China University of Science and Technology,Shanghai 200237,China

Abstract　Dependency parsing is a crucial task in natural language processing,aiming to identify the syntactic dependencies between words in a sentence. However,existing research on dependency parsing for Chinese electronic medical records faces following problems:current general-purpose parsers are unable to accurately analyze the situation when there is a lack of components indicative of grammatical structure and a variety of positions of modifiers. To address these issues,this paper proposes a method based on a dual-scale collaborative enhancement of large and small language models for dependency parsing of Chinese electronic medical records. Specifically,we first analyze the linguistic features of Chinese electronic medical records,and propose component completion to indicate special grammatical structures in medical texts. Subsequently,we utilize a generic parser for dependency parsing,for the parsed syntactic graph,we employ the prior grammatical knowledge of a large language model to modify it automatically. In addition,since our approach focuses on narrowing the feature distribution gap between medical and generic texts,it is not constrained by the lack of annotated data in the medical domain. This study annotates 444 samples for dependency parsing of Chinese electronic medical records,which validates our method. Experimental results demonstrate the effectiveness of our approach in parsing Chinese electronic medical records,achieving LAS and UAS metrics of 92.42 and 94.60 in the scenario with little data. The proposed method also shows significant performance in various departments.

Keywords　Natural language processing,Dependency parsing,Chinese electronic medical records,Large language model,Collaborative enhancement

到稿日期:2023-12-07　返修日期:2024-04-28
基金项目:上海市促进产业高质量发展专项资金(2021-GZL-RGZN-01018)
This work was supported by the Shanghai Municipal Special Fund for Promoting High-quality Development of Industries (2021-GZL-RGZN-01018).
通信作者:朱焱(zhuygraph@ecust.edu.cn)

1 引言

依存句法分析(Dependency Parsing,DEP)是一项重要的自然语言处理任务[1],它旨在识别句子中词汇之间的依存语法关系[2-3],如图1所示。

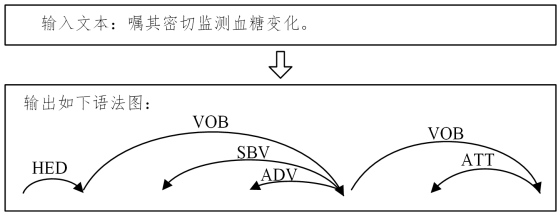

图1 依存句法分析示例

Fig. 1 Example of dependency parsing

随着医疗领域信息化建设的稳步发展,现代医学信息系统积累了大量的医疗数据。其中电子病历是重要的医疗临床数据,通过对其进行分析,可以挖掘出大量的医疗知识。电子病历的知识获取是自然语言处理(Natural Language Processing,NLP)领域研究的热点问题。DEP和信息抽取是知识获取的重要手段,而信息抽取通常依赖于句子解析出的最短路径,从而识别文本中提到的实体如何相关。换言之,DEP通过句子解析,对句子进行结构化的处理,解析复杂的上下文环境,从而为下游的信息抽取任务提供丰富信息。

当前的DEP模型大多是在大规模语料上训练的基于统计的模型,由于中文电子病历(Chinese Electronic Medical Record,CEMR)领域标注语料的匮乏,目前缺少面向CEMR领域的DEP模型。一般情况下,限定领域和通用领域的文本之间具有很大差异,其主要体现在词、词性或句子的分布上,如果直接将通用DEP模型应用在CEMR领域,模型的精度就会明显降低。本文关注CEMR的语言特征,通过分析,发现现有电子病历中存在以下两类问题,使得CEMR与通用文本之间分布差异较大,从而导致通用解析器无法准确分析CEMR。

1)缺省指示语法结构的成分:CEMR中常出现成分缺失、词性调整、并列成分过多的现象,这些现象会因其缺省指示语法结构的成分而导致通用解析器精度降低。

2)修饰成分位置多样:CEMR中常出现状语后置和定语独立的现象,这些现象会因修饰成分多样而导致通用解析器精度降低。

针对上述问题,提出基于大小语言模型协同增强的中文电子病历依存句法分析方法。具体而言,先通过成分补全指示医疗文本中的特殊语法结构,再利用通用DEP模型进行解析,且对于解析后的语法图,利用LLM的先验语法知识,进一步进行自动修正。该方法将重点放在缩小CEMR与通用文本之间的特征分布差异上,避免了医疗领域缺少标注数据的限制。

总的来说,本文的主要贡献有以下3点。

1)提出一个大小模型协同的成分补全算法。利用成分缺失识别模型,引导LLM对句子中的缺失成分进行补全,进而缩小了语料差距,改善了缺省指示语法结构的成分的问题,使得通用模型在CMER上获得了更好的解析效果,将LAS分数提升了8.25。

2)提出利用LLM的先验句法知识,对通用模型解析获得的语法图进行修正,改善了因修饰成分位置多样而导致的语法图精度低的问题,进一步提升了通用模型的性能。

3)构建了一份有444条CEMR的依存句法分析数据集,其来自儿童糖尿病和癌症两种临床文本,可用于验证CEMR的依存句法分析模型的效果。

2 相关工作

2.1 医疗领域的依存句法分析

当前的依存句法分析方法主要有以下两类。1)基于转移的方法。Chen等[4]首次将深度学习用于基于转移的方法中,为构建语法图的每一个可采取的操作分配一个概率,从而获得由一系列最优操作构成的语法图。Weiss等[5]通过束搜索和条件随机场,对Chen的方法进行增强,允许解析器撤销先前可能错误的操作。Dyer等[6]利用LSTM取代Chen的方法中的堆栈及缓冲区,通过组合已解析的短语,构建出精度更高的语法图。2)基于图的方法。Kiperwasser等[7]提出使用LSTM作为编码器提取特征,并将这些特征应用于词与词之间依存关系的评分中,从而构建高质量语法图。Dozat等[8]提出双仿方法,其改进了基于图的方法的普通评分函数。Mrini等[9]将标签信息与自注意力机制相结合,以进行双仿依存分析。

2019年,医疗领域的依存句法分析任务被作为CRAFT共享任务的子任务CRAFT-SA[10],TurkuNLP小组使用Turku解析器和UDify解析器在此任务中获得89.7分的最好表现[11]。但目前医疗领域DEP的大部分研究都是针对英语文本的,这些研究也都是针对英语本身的语言特点的。与英文DEP相比,中文DEP起步较晚,由于中文本身的语言特点,中文医疗文本的语法更加复杂。目前,中文医疗领域只有哈工大的基于模型融合的中文电子病历成分句法分析模型[12],而成分分析方法需要满足组合成分约束。因此,对于中文医疗领域的依存句法分析,仍需进行研究。

2.2 基于大语言模型的依存句法分析

LLM指通过在大型未标注语料上以无监督方式进行训练,来学习通用语言模式和特征的模型[13]。随着LLM规模的扩大,其在各种NLP任务上体现出卓越的性能[14]。Sun等[15]使用ChatGPT,将依存句法分析形式化为两步文本补全任务,该工作在仅使用ChatGPT时取得了显著效果。ChatGPT在进行简单句的DEP时效果较好,但在进行长难句特别是专业领域的长难句DEP时,分析的结果较粗略,甚至出现幻觉,并且ChatGPT自身无法确定哪两个词之间存在依存关系,分数也比传统基于双仿的模型低。

从理论上来说,LLM拥有的句法知识比传统模型多,然而在DEP任务上,其表现不如传统模型。原因在于LLM是基于自然语言语料训练的,其不擅长图结构的推理。因此,引入小模型引导LLM的形式进行研究,结合了小模型针对性强的优点和LLM解决零样本问题的能力。该方法利用通用依存分析模型提供结构化图的模板,结合LLM的海量先验

句法知识，从 CEMR 语言结构特点出发，研究医疗领域的 DEP 问题。

3 CEMR 的特点及问题分析

随着信息化在医疗领域的不断发展，几乎所有医疗机构都开始使用电子病历来记录患者就诊或住院过程中的各种医疗文本信息[16]。对 CEMR 中出现的与通用文本存在差异的各类语法结构进行了统计，如表 1 所列。电子病历缺省指示语法结构的成分和修饰成分位置多样的问题，导致通用模型对其进行 DEP 时会出现以下问题。

表 1 CEMR 中各类语言特征统计信息

Table 1 Statistical information of various language features in CEMR

	出现边数/总边数	占比/%	例句	DEP 时出现的问题
成分缺失	892/2834	31.47	尿常规：尿酮 4＋，尿糖 3＋，尿蛋白 3＋	错误地将核心词识别为“4＋”
词性调整	1034/2834	36.49	患儿 6 月余前无明显诱因出现多饮、多尿、食欲增加	错误地将核心词“无明显诱因”识别为定语
并列成分过多	214/2834	7.55	患儿 6 月余前无明显诱因出现多饮、多尿、食欲增加	错误地将“多饮”识别为“增加”的主语
状语后置	823/2834	29.04	患儿，男，3 岁；因发热 2 天，呕吐 2 次入院	错误地将状语“2 天”识别为补语
定语独立	179/2834	6.32	患儿，男，3 岁；因发热 2 天，呕吐 2 次入院	错误地将定语“男”识别为“患儿”的并列成分

1）缺省指示语法结构的成分

（1）成分缺失：省略隐含信息导致句子成分不完整，从而导致核心词及其他依存关系解析错误。

（2）词性调整：出现词性调整情况时，通用模型无法正确识别成分词性。

（3）并列成分过多：导致无法准确连接并列成分，从而导致句子层级混乱。

2）修饰成分位置多样

（1）状语后置：此时通用模型受训练数据的限制，通常错误地识别状语为补语。

（2）定语独立：此时，通用模型一方面无法识别定语的正确词性，另一方面无法识别正确的被修饰成分。

根据中心理论[17]，主语、谓语和宾语作为句子的主要成分，如果句子中缺失某一成分或成分词性识别错误，通用模型在对句子进行 DEP 时就会产生错误判断[18]，使核心词的识别与依存关系的识别出现错误。

对上述问题进行分析发现，对于问题 1），可以通过成分补全来缩小 CEMR 与通用语料之间的差距，从而改善通用模型的 DEP 结果。如图 2 所示，针对图中省略隐含信息的句子进行谓语动词补全后，通用模型可以进行正确分析；对于词性调整的情况，可以通过补全指示词性的词来引导通用模型进行正确分析，如将“患儿 6 月余前无明显诱因出现多饮、多尿、食欲增加”补全为“患儿 6 月余前无明显诱因地出现多饮、多尿、食欲增加。”后再进行通用模型解析的结果更为合理；对于并列成分过多的情况，可以通过添加连接词来避免通用模型解析混乱。而问题 2）的出现则是由于通用模型受其训练语料语法特征分布的限制，可以通过 LLM 丰富的先验语法知识对这类问题进行自动修正。

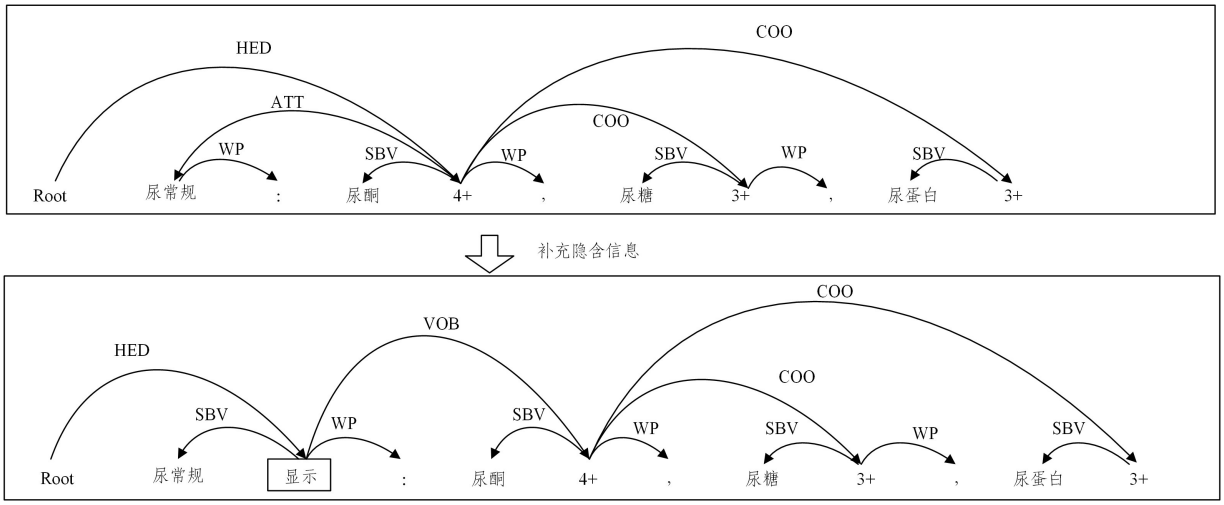

图 2 缺失成分补全示例

Fig. 2 Example of missing component completion

4 基于大小语言模型协同增强的中文电子病历依存句法分析

大小语言模型协同增强的 DEP 方法的整体框架如图 3 所示。首先，利用缺失成分识别模型（Missing Component Identification Model，MCIM）来识别输入文本的缺失位置与类型。然后，基于 LLM 迭代反思补全缺失成分，以缩小 CEMR 数据与通用领域数据间的特征差距。最后，利用 LLM 的先验语法知识对通用模型的依存句法分析结果进行自动修正。

4.1 缺失成分识别

为确定一条 CEMR 文本在什么位置存在成分缺失问题，

对句子缺失成分识别任务(Missing Component Identification, MCI)进行定义,并提出了 MCIM 模型框架。

4.1.1 缺失成分识别任务定义

对于输入的句子 $S=(x_1,x_2,\cdots,x_n)$,其中 n 是词的个数,模型输出得到一个序列 $T=(y_1,y_2,\cdots,y_n)$,其中 y_i 表示句子中第 i 个字后是否有缺失成分。若无缺失成分,则 $y_i=O$;若有缺失成分,则根据缺失成分的词性,分别标注 B-V, B-A, B-C 来代表第 3 章提到的前 3 种情况。例如对于"尿常规:尿酮 4+,尿糖 3+,尿蛋白 3+",得到的目标序列为(O, O, B-V, O, …, O),表示"规"后缺失的成分符合第一种情况。

4.1.2 缺失成分识别模型架构

MCI 任务可以看作 sequence-to-sequence 任务的变体,其整体框架如图 3 中 Step1 所示,模型包括编码和识别两部分。

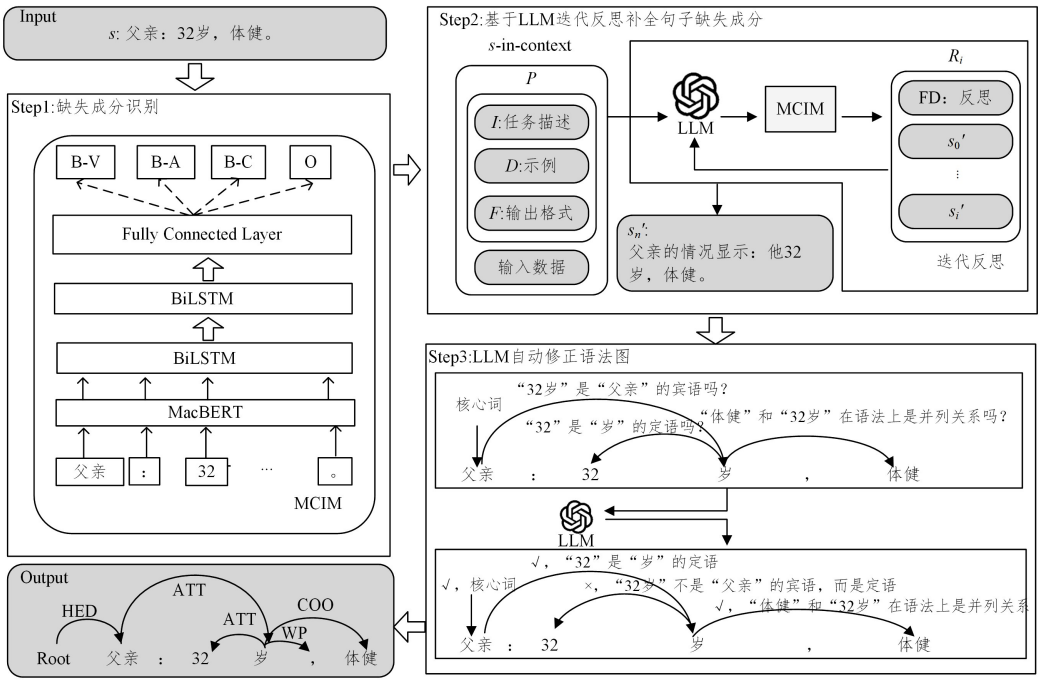

图 3 大小语言模型协同增强框架

Fig. 3 Dual-scale collaborative enhancement framework of large and small language models

1)编码层:引入 MacBERT[19] 对输入文本进行编码。输入遵循 BERT 的格式,对于句子 S,输入为:[CLS]〈S〉[SEP],输入句子经过 MacBERT 模型处理后,输出的词级特征序列为 $\boldsymbol{e}=(e_1,e_2,\cdots,e_n)$,其中 n 是输入序列的长度。$\boldsymbol{e}\in R^{n\times d_e}$,表示输入数据对应的特征向量集合,$d_e$ 是词嵌入的维度。

2)MCI 层:由双层 BiLSTM 与全连接层构成,进行缺失成分识别。

第一层 BiLSTM 将 MacBERT 的输出作为输入,输出如下:

$$\overrightarrow{\boldsymbol{h}^{(1)}}=\overrightarrow{\boldsymbol{LSTM}(\boldsymbol{e})} \tag{1}$$

$$\overleftarrow{\boldsymbol{h}^{(1)}}=\overleftarrow{\boldsymbol{LSTM}(\boldsymbol{e})} \tag{2}$$

$$\boldsymbol{h}_i^{(1)}=[\overrightarrow{\boldsymbol{h}_i^{(1)}},\overleftarrow{\boldsymbol{h}_i^{(1)}}],i=1,2,\cdots,n \tag{3}$$

$$\boldsymbol{h}^{(1)}=[\boldsymbol{h}_1^{(1)},\boldsymbol{h}_2^{(1)},\cdots,\boldsymbol{h}_n^{(1)}] \tag{4}$$

其中,$\overrightarrow{\boldsymbol{h}^{(1)}}$ 和 $\overleftarrow{\boldsymbol{h}^{(1)}}$ 是第一层 BiLSTM 的前向和后向输出;$\boldsymbol{h}^{(1)}\in R^{n\times 2d_h}$ 表示两个方向输出的合成,d_h 表示 LSTM 隐藏层的维度。与第一层相似,第二层 BiLSTM 将第一层 BiLSTM 的输出作为输入,输出如下:

$$\overrightarrow{\boldsymbol{h}^{(2)}}=\overrightarrow{\boldsymbol{LSTM}(\boldsymbol{h}^{(1)})} \tag{5}$$

$$\overleftarrow{\boldsymbol{h}^{(2)}}=\overleftarrow{\boldsymbol{LSTM}(\boldsymbol{h}^{(1)})} \tag{6}$$

$$\boldsymbol{h}_i^{(2)}=[\overrightarrow{\boldsymbol{h}_i^{(2)}},\overleftarrow{\boldsymbol{h}_i^{(2)}}],i=1,2,\cdots,n \tag{7}$$

$$\boldsymbol{h}^{(2)}=[\boldsymbol{h}_1^{(2)},\boldsymbol{h}_2^{(2)},\cdots,\boldsymbol{h}_n^{(2)}] \tag{8}$$

再由全连接层对 BiLSTM 层的输出信息进行分类,得到神经网络的标签预测概率:

$$o_j^{\mathrm{MCI}}=\sigma(W^{\mathrm{MCI}}h_j+b^{\mathrm{MCI}}),j=1,2,\cdots,n \tag{9}$$

$$\boldsymbol{o}=\{o_1^{\mathrm{MCI}},o_2^{\mathrm{MCI}},\cdots,o_n^{\mathrm{MCI}}\} \tag{10}$$

其中,W^{MCI} 和 b^{MCI} 是全连接层的参数;$o_j\in R^{\mathrm{tags}}$ 表示输入的第 j 个字符对应的标签预测。

训练过程中,采用负对数似然函数作为损失函数:

$$L_{\mathrm{MCI}}=\sum_{s\in S}\log(P(\boldsymbol{y}^{\mathrm{MCI}}s|\boldsymbol{o}_s)) \tag{11}$$

其中,S 是训练集;$\boldsymbol{y}^{\mathrm{MCI}}{}_s$ 和 $\boldsymbol{o}_s$ 分别为 s 真实的标签序列和神经网络的输出。

4.2 基于大语言模型迭代反思的缺失成分补全

为了提升通用模型在 CEMR 上的效果,提出面向 CEMR 适配的基于 LLM 反思的句子缺失成分补全以及对 DEP 结果进行修正的方法,以改进通用模型在 CEMR 上的效果。

针对第 3 节提到的 CEMR 出现的成分缺失,在 MCIM 指出缺失成分位置和缺失类型后,利用 LLM 的上下文学习能力来进行缺失成分补全,并构造反思环境来稳定 LLM 补全结果的质量。首先利用上下文学习技术,构建句子成分补全提示指令,引导 LLM 进行成分补全。然而在补全过程中,发现 LLM 的输出不稳定,存在成分丢失、输出原句等问题,

因此引入反思机制来确保 LLM 的输出质量。

构造指令引导 LLM 完成补全任务，其指令可以定义为 $P=(I,D,F)$，I,D,F 分别为任务描述、示例、输入输出格式要求。对于输入 LLM 的句子 s，首先组合 P 和 s，生成 LLM 的输入文本 s-in-context，引导 LLM 生成特定于医疗领域的补全结果 s_0'。为确保最终结果的质量，引入迭代反思。对于上一步中 LLM 输出的结果 s_i'，利用 MCIC 进行缺失成分判别后，对比 s_i' 与 s，构造反思指令 $R_i=(FD,s_0',\cdots,s_i')$，其中包含思考和历史补全结果，从而提示 LLM 进行观察，思考并输出更符合任务要求的结果，直到输出以 s 为真子序列且 MCIM 判别无缺失成分的结果 s_n' 时，结束迭代。

例如对句子“父亲：32 岁，体健。”进行 MCI 后，得到“父亲”后存在成分缺失的结果，故构造任务描述 I 为“‘父亲’后缺失了成分，请进行补全，使句子成分完整”。构造示例 D 为“输入：尿常规：尿酮 4+，尿糖 3+，尿蛋白 3+；输出：监测血糖，并予胰岛素静脉维持输注（0.1u/kg/h－0.05u/kg/h）”，输出格式要求 F 为“只在指定位置添加成分且不可删除或更改原句已有内容”。将 I,D,F 整体构成的指令 P 与 s 组合成 s-in-context 来引导 LLM 进行成分补全。

4.3 基于大语言模型先验知识的语法图自动修正

在使用通用模型对 LLM 补全后的句子 s_n' 进行 DEP 时，发现其依存分析的质量有显著提升，但仍存在部分情况通用模型无法正确分析。如第 3 章中提到的问题 2），通用模型在分析时无法识别词与词之间正确的依存关系，如图 4 所示，且这类问题无法通过补全句子成分来解决。对此，利用 LLM 在海量文本中学习到的语法先验知识，对通用模型生成的语法图结果进行修正。

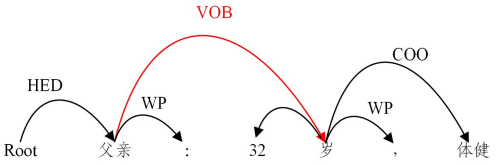

图 4 定语独立情况

Fig. 4 Case of independent determiner

由于 LLM 在预训练阶段接触了大量的自然语言文本，因此其对于高度结构化的输入的处理能力较差。故对于通用模型输出的语法图结果 D，首先进行语法单元三元组 $T=(h,t,r)$ 的采样，其中 h 为头节点，t 为尾节点，r 为两节点之间的语法关系。然后利用三元组的节点和边，构建语法问句。对于每一个语法问句构建指令 $P_d=(I,D,F)$，引导 LLM 对句子中的定语和状语成分进行修正。

如图 5 所示，对于句子“父亲：32 岁，体健。”的 DEP 结果，首先采样出语法单元三元组：{(32 岁，父亲，宾语)，(32，岁，定语)，(体健，32 岁，并列)}。针对该三元组，构建语法问句：“32 岁”是“父亲”的宾语吗；“32”是“岁”的定语吗；“体健”和“32 岁”在语法上是并列关系吗。将组合问句和任务描述输入 LLM，根据 LLM 的输出结果对语法图结果进行修正。

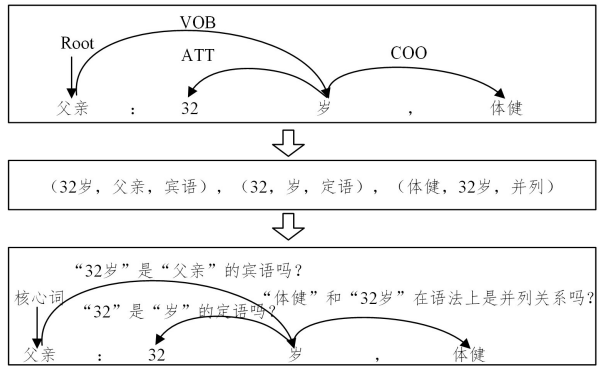

图 5 语法问句的构造

Fig. 5 Construction of grammatical questions

5 实验结果与分析

本章介绍了基线方法等实验设定，然后在 CEMR 依存句法分析数据集上验证了所提方法的性能，并对实验方法进行详细分析，通过消融实验研究了影响实验结果的因素。

5.1 主实验设定

1）CEMR 测试数据集。随机选取了某三甲医院 200 条儿科糖尿病临床电子病例文本和 Cblue 的 244 条胃癌临床电子病历文本进行标注，标注流程包括分词、词性标注和依存关系标注。为验证通用模型在电子病历语料上的依存句法分析效果，分词及词性标准都与通用模型统一，采用 PKU Multi-view Chinese Treebank[20] 标注规范。为保证数据标注的效率和一致性，先利用通用模型得出语料的分析结果，再将每条句子随机分配给两名不同的标注者进行修正标注，若不同标注者的标注结果一致，则确立为正确答案；若结果不完全一致，则将其分配给专家，由专家给出标准结果。该数据集中最小句长为 6 个字，最大句长为 79 个字，平均句长为 28 个字。

2）基线方法。使用通用依存句法分析模型 LTP[21] 作为基准。

3）测评指标。DEP 部分实验采用的评测指标为 LAS 和 UAS[22]：

$$LAS=\frac{|\{e|l_G(e)=l_P(e),e\in E_G\cap E_P\}|}{|V|} \tag{12}$$

$$UAS=\frac{|\{e|e\in E_G\cap E_P\}|}{|V|} \tag{13}$$

其中，$G=\langle V,E_G,l_G\rangle$ 为标准 DEP 结果；$P=\langle V,E_P,l_P\rangle$ 为预测的 DEP 结果；$E\in V\times V$ 为从属节点到头节点的有向边集合；$l:V\times V\to L$ 为边标记函数，L 为从属关系的标签集合。

5.2 主实验结果

为了验证本文方法的有效性，分别在儿科糖尿病临床电子病历和胃癌临床电子病历上将其与 LTP，Hanlp[23]，DDParser[24] 3 个通用模型进行了对比实验。表 2 列出了两个数据集上的 LAS 和 UAS 分数，从中可以得出以下结论。

对比表 2 中的前 5 列和后 5 列可以看出，使用本文方法对通用模型增强后，再对 CEMR 进行 DEP 时，儿科糖尿病电子病历上 DEP 的 LAS 指标最高提升了 11.22，UAS 指标

最高提升了8.93；胃癌临床电子病历上 DEP 的 LAS 指标最高提升了 9.45，UAS 指标最高提升了 8.1。本文方法能够在 CEMR 上领先基线模型，表明其可以有效处理 CEMR 上的 DEP 任务。

表 2 对比实验结果

Table 2 Experiment result comparison

	儿科糖尿病		癌症		+our method	儿科糖尿病		癌症	
	LAS	UAS	LAS	UAS		LAS	UAS	LAS	UAS
HanLP	73.69	76.91	78.52	80.91	HanLP	80.53	82.87	83.68	85.73
DDParser	77.42	78.83	80.50	84.54	DDParser	84.58	86.21	85.12	87.57
LTP	81.20	85.67	83.67	86.92	LTP	92.42	94.60	93.12	95.02

本文方法在儿科糖尿病电子病历上的效果更加显著，因为其数据集是从临床收集的原始文本且没有进行任何的加工。而胃癌临床电子病历数据集取自阿里云天池的临床发现事件抽取任务(CHIP-CDEE)，该数据集由医学专家进行审核和校验，其句子结构更为完整。此外，表 2 中的结果验证了本文方法在不同科室的电子病历的 DEP 中都有显著提升，表明该方法可作为面向 CEMR 的通用方法。

5.3 MCI 实验

1)MCI 数据集。利用 CBLUE 的中文医学命名实体识别 V2 任务提供的公开语料进行 MCI 任务数据集的构建。对于原始的完整文本，通过选择部分词进行删除来构造符合第 3 章中问题 1)提到的 3 种成分缺失情况的文本，然后对其进行标注。数据的详细情况如表 3 所列。

表 3 MCI 任务数据集的详细信息

Table 3 Details of MCI task dataset

	B-V	B-A	B-C	总量
训练集	2393	2574	2685	3633
验证集	1293	1087	963	1368
测试集	1296	1247	1023	1391

2)实验设置。本文采用 MacBERT-base 模型，该模型的参数量为 1.02×10^8，12 层，隐藏层维度为 768，12 个头。将 MacBERT 模型的学习率设置为 2×10^{-5}，其他模块学习率设置为 2×10^{-3}，使用 Adam 作为优化器，将批处理大小设置为 16 进行训练。

3)测评指标。MCI 实验采用的评测指标包括精确率 P、召回率 R 和 F_1。

$$P=\frac{TP}{TP+FP} \tag{14}$$

$$R=\frac{TP}{TP+FN} \tag{15}$$

$$F_1=\frac{2PR}{P+R} \tag{16}$$

其中，TP(True Positive)表示被正确识别的正样本；FP(False Positive)表示被错误识别的正样本；FN(False Negative)表示被错误识别的负样本；F_1 用于综合评估模型性能。MCI 的实验结果如表 4 所列。

表 4 MCI 任务的实验结果

Table 4 Results of MCI task

	P	R	$F1$
MCI	62.26	67.32	66.09

5.4 LLM 的虚幻性分析

利用 LLM 的上下文学习能力来补全缺失成分，虽然 LLM 在该任务中表现出卓越的性能，但其经常会出现幻觉。LLM 会产生偏离输入、与先前生成的上下文相矛盾或与既定事实相悖的内容[25]。为保证 LLM 补全结果的质量，对由 LLM 的虚幻性导致的错误类型进行分析与探测，并利用迭代反思机制对其错误结果进行纠正。分析结果如表 5 所列。

由于 LLM 的虚幻性，其在进行成分补全时会引入错误信息，具体会出现以下 3 种错误类型。

1)补充位置错误。如表 5 第 2 行中，文本“目前胰岛素用量：早餐前中效 4IU；晚餐前中效 2IU。”缺少词性指示词，导致解析模型无法识别“早餐前”为状语，其正确补充结果应在“早餐前”添加动词“注射”，而 LLM 的补充位置错误，无法解决该句中出现的问题。

2)引入冗余信息。如表 5 第 3 行中，LLM 为文本“尿常规：尿酮 4+，尿糖 3+，尿蛋白 3+；”正确补充缺失成分“显示”后，捏造了冗余信息“尿潜血 2+，尿白细胞 1+，尿红细胞 2+；”，这会影响下一步解析模型的 DEP 效果。

3)改动原文内容。如表 5 最后一行中，LLM 虽然为文本“监测血糖，予胰岛素静脉维持输注(0.1u/kg/h-0.05u/kg/h)”正确补充了连词“并”，但是将原文的“予”改为了“进行”。由于 LAS 和 UAS 分数的计算需保证两个语法图之间节点一致，因此改动原文内容会导致 LAS 和 UAS 分数计算错误。

表 5 LLM 幻觉分析

Table 5 Analysis of hallucination in LLM

错误类型	输入文本示例	缺失位置及类型	输出文本示例	错误比例/%	纠正比例/%
补充位置错误	目前胰岛素用量：早餐前中效 4IU；晚餐前中效 2IU。	(10,B-A)	目前胰岛素用量为：早餐前中效 4IU；晚餐前中效 2IU。	29.74	24.75
引入冗余信息	尿常规：尿酮 4+，尿糖 3+，尿蛋白 3+；	(2,B-V)	尿常规显示：尿酮 4+，尿糖 3+，尿蛋白 3+；尿潜血 2+，尿白细胞 1+，尿红细胞 2+；	24.15	21.36
改动原文内容	监测血糖，予胰岛素静脉维持输注(0.1u/kg/h-0.05u/kg/h)	(4,B-C)	监测血糖，并进行胰岛素静脉维持输注(0.1u/kg/h-0.05u/kg/h)	17.37	16.17

对于 LLM 引入的错误信息，实验中通过结合 MCI 识别的缺失位置及类型和原句内容进行探测；利用迭代反思机制纠正错误，提示 LLM 进行观察，思考并输出更符合任务要求的结果。例如对于错误信息 1)，可以进一步提示 LLM 需要补充的位置和类型；对于错误信息 2)，可以进一步控制 LLM 添加的字符数量，并提示 LLM 不要捏造信息；对于错误信息 3)，可以提示 LLM 在原文基础上进行补全，不要改动原文内容。

5.5 消融实验

本文定义缺失成分识别任务，训练缺失成分识别器，并提出通过进行缺失成分补全和 LLM 自动修正的方式提升通用模型在 CEMR 上的 DEP 效果。为了验证这 3 部分的有效性，本节以 LTP 为基础，进行了一系列消融实验。

1)缺失成分识别。定义了缺失成分识别任务，提出采用 sequence-to-sequence 的方法识别句子中成分缺失的位置及缺失类型。因此，尝试去除缺失成分识别，使 LLM 自动识别缺失成分并补全，再重新进行实验，以观察结果变化，实验结果如表 6 所列。

表 6 消融实验结果

Table 6 Results of ablation study

LTP	儿科糖尿病		癌症	
	LAS	UAS	LAS	UAS
＋ChatGPT	85.60	88.96	86.95	89.38
＋MCIM＋ChatGPT	86.92	89.07	88.91	90.29
＋ChatGPT＋feedback	87.77	90.31	89.27	91.73
＋MCIM＋ChatGPT ＋feedback	89.45	91.35	90.07	92.31
＋Auto Modify	84.56	87.72	85.92	89.28
＋ ChatGPT＋Auto Modify	88.52	90.84	90.21	91.39
＋ MCIM＋ChatGPT ＋Auto Modify	90.79	92.02	91.16	92.89
ChatGPT＋feedback＋ Auto Modify	90.28	91.79	91.37	93.74
＋MCIM＋ChatGPT＋ feedback＋ Auto Modify	92.42	94.60	93.12	95.02

由表 6 可知，与直接使用 LLM 自动识别缺失成分并补全的 DEP 结果相比，使用缺失成分识别模型对 LLM 进行引导的 DEP 结果的 LAS 指标最高提升了 2.27，UAS 指标最高提升了 2.81。这说明缺失成分识别模型可以帮助 LLM 准确识别句子中的缺失成分位置与类型，再利用 LLM 的生成能力，可以显著提升 DEP 的结果。这可能是因为 LLM 虽然能够完成缺失成分任务，但由于此任务在语料中并不常见，因此需要小模型引导。

2)基于 LLM 迭代反思补全缺失成分。为了缩小 CEMR 与通用中文文本间的差距，提出利用 LLM 补全句子中的缺失成分，以充分利用 LTP 的 DEP 能力。为验证该方法的有效性，尝试直接利用 LTP 进行 CEMR 的 DEP 分析，并直接对 LTP 的 DEP 结果进行自动修正。

利用缺失成分补全方法可以有效缩小 CEMR 与通用中文文本的差距，从而改善 LTP 在 CEMR 上的效果。表 6 中第 2 列和第 4 列的实验结果表明，该方法在零训练的前提下，将 LAS 指标提升了 5.72，将 UAS 指标提升了 3.4，这说明通过缺失成分补全可以突破 LTP 训练语料的局限性。

本节还对 LLM 迭代反思的必要性进行了验证。对比表 6 中第三列和第五列、第二列和第四列实验结果可以看出，利用 LLM 迭代反思后，实验效果均有提升，这说明迭代反思可以提升 LLM 的稳定性。

3)LLM 自动修正语法图。在对补全后的文本进行 DEP 时发现，对于 CEMR 中出现的状语后置与动词短语作状语、定语与被修饰成分独立等现象，LTP 在分析时无法识别词与词之间正确的依存关系。为了进一步修正语法图中出现的这些现象，提出利用 LLM 自动修正语法图的方法，充分利用 LLM 丰富的先验知识。本节通过对比 LLM 自动修正前后的实验结果，观察性能变化。

分别对比表 6 中前 4 栏与后 4 栏的实验结果可以看出，利用 LLM 自动修正语法图后，可以进一步提升 DEP 效果，其中 LAS 指标最高提升了 3.36，UAS 指标最高提升了 3.25，这说明 LLM 比 LTP 拥有更多、领域更广的句法知识。

结束语 针对 CEMR 的特征，通过 MCIM 模型引导 LLM 迭代反思进行文本补全，缩小其与中文通用文本间的特征差距；利用 LLM 先验知识进行语法图的自动修正，在不进行 DEP 训练的前提下提升各类 DEP 模型的效果。从实验结果来看，其仍然存在无法正确解析的语法结构。后续考虑引入医疗词典以提高分词结果的准确性，并且结合更多的医疗领域信息，进一步提升 CEMR 上的依存分析效果。

参 考 文 献

[1] EISNER. Bilexical Grammars and their Cubic-Time Parsing Algorithms [J]. Springer Netherlands, 2000, 10(7): 29-61.

[2] CHEN W L, ZHANG M, ZHANG Y. Semi-supervised Feature Transformation for Dependency Parsing[C]// Proceedings of the 2013 Conference on Empirical Methods in Natural Language. 2013: 1303-1313.

[3] TIMOTHY D, CHRISTOPHER M. Deep Biaffine Attention for Neural Dependency Parsing[C]// Proceedings of the 2017 International Conference on Learning Representations. 2017: 1-8.

[4] CHEN D Q, MANNING. A fast and accurate dependency parser using neural networks[C]// Proceedings of the 2013 Conference on Empirical Methods in Natural Language. 2014: 740-750.

[5] WEISS D, ALBERTI C, COLLINS M. Structured training for neural network transition-based parsing. [C]// Proceedings of Annual Meeting of the Association for Computational Linguistics. 2015.

[6] DYER C, BALLESTEROS M, WANG L, et al. Transition-based dependency parsing with stack long short-term memory[C]// Proceedings of the 2015 Conference on Empirical Methods in Natural Language. 2015.

[7] KIPERWASSER E, GOLDBERG Y. Simple and accurate dependency parsing using bidirectional LSTM feature representations[C]// Proceedings of Annual Meeting of the Association for Computational Linguistics. 2016: 313-327.

[8] DOZAT T, MANNING C. Deep biaffine attention for neural dependency parsing[J]. arXiv: 1611.01734, 2016.

[9] MRINI K, DERNONCOURT F. Rethinking self-attention: Towards interpretability in neural parsing [C]// Proceedings of the 2020 Conference on Empirical Methods in Natural Language.

2020:731-742.
[10] BADA M,PYYSALO S,CIOSICI M,et al. Craft Shared Tasks 2019 Overview — Integrated Structure, Semantics, and Coreference[C]// Proceedings of the 5th Workshop on BioNLP Open Shared Tasks. 2019:174-184.
[11] NGO TM,KANERVA J,GINTER F,et al. Neural Dependency Parsing of Biomedical Text: TurkuNLP entry in the CRAFT structural annotation task[C]// Proceedings of the 5th Workshop on BioNLP Open Shared Tasks. 2019:206-215.
[12] JANG Z P,GUAN Y. A Fusion Model for Chinese Electronic Medical Record Parsing [J]. ACTA Automatica Sinica, 2019, 45(2):276-288.
[13] KOPF A,KILCHER Y,RUTTE D,et al. Open Assistant Conversations-Democratizing Large Language Model Alignment [C]// Proceedings of the 2023 Conference and Workshop on Neural Information Processing Systems. 2023:1-13.
[14] WEI J, TAY Y, RISHI B, et al. Emergent Abilities of Large Language Models [J]. arXiv:2206.07682,2022.
[15] SUN X F,DONG L F. Pushing the Limits of ChatGPT on NLP Tasks[J]. arXiv:2306.09719,2023.
[16] SCHICK T,SCHÜTZE H. Exploiting Cloze-questions for Few-shot Text Classification and Natural Language Inference[C]// Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. 2021:255-269.
[17] GUNTER T D,TERRY N P. The Emergence of National Electronic Health Record Architectures in the United States and Australia:Models,Costs,and Questions [J]. Journal of Medical Internet Research,2005,7(1):e3.
[18] YEH C L,CHEN Y C. Zero Anaphora Resolution in Chinese with Shallow Parsing [J]. Journal of Chinese Language and Computing,2007,17(1):41-56.
[19] JIANG M,HUANG Y,FAN J W,et al. Parsing Clinical Text: How Good Are the State-of-the-art Parsers? [J] BMC Medical Informatics and Decision Making,2015,15(S1):1-6.
[20] SHI J L,LUO X Y. Construction of a Treebank of Learners Chinese [J]. Journal of Chinese Information Processing, 2022, 36(1):39-46.
[21] CHE W,FENG Y,QIN L,et al. N-LTP:An Open-source Neural Language Technology Platform for Chinese[C]// Proceedings of the 2021 Conference on Empirical Methods in Natural Language. 2021:42-49.
[22] PLANK B,ALONSO H M,AGIĆ Ž,et al. Do dependency parsing metrics correlate with human judgments? [C]// Proceedings of the 19th Conference on Computational Natural Language Learning. 2015:315-320.
[23] HAN H,CHOI J D. The Stem Cell Hypothesis:Dilemma behind Multi-Task Learning with Transformer Encoders[C]// Proceedings of the 2021 Conference on Empirical Methods in Natural Language. 2021:5555-5577.
[24] ZHANG S, WANG L, SUN K, et al. A practical Chinese dependency parser based on a large-scale dataset [J]. arXiv:2009.00901,2020.
[25] ZHANG Y,CUI L. Siren's Song in the AI Ocean:A Survey on Hallucination in Large Language Models [J]. arXiv: 2309.01219,2023.

XU Siyao, born in 2000, postgraduate. Her main research interests include natural language processing and dependency parsing.

ZHU Yan, born in 1984, Ph.D, associate professor. Her main research interest is graph theory and its applications.

(责任编辑:柯颖)