基于本体信息增强的人类表型概念识别

祁杰蔚, 罗凌*, 杨志豪, 王健, 林鸿飞
大连理工大学，计算机科学与技术学院，大连，116024
jwqi@mail.dlut.edu.cn,{lingluo,yangzh,wangjian,hflin}@dlut.edu.cn

摘要

从文本中自动识别人类表型概念对疾病分析具有重大意义。现存本体驱动的表型概念识别方法主要利用本体中概念名和同义词信息，并未充分考虑本体丰富信息。针对此问题，本文提出一种基于本体信息增强的人类表型概念识别方法，利用先进大语言模型进行数据增强，并设计本体向量增强的深度学习模型来提升概念识别性能。在GSC+和ID-68两个数据集上进行实验，结果表明本文提出方法能够利用本体丰富信息有效提升基线模型性能，取得了先进结果。

关键词： 生物医学概念识别 ； 人类表型本体 ； 本体信息增强

Ontology Information-augmented Human Phenotype Concept Recognition

Jiewei Qi,Ling Luo*, Zhihao Yang, Jian Wang, Hongfei Lin
School of Computer Science and Technology,Dalian University of Technology,Dalian,116024
jwqi@mail.dlut.edu.cn,{lingluo,yangzh,wangjian,hflin}@dlut.edu.cn

Abstract

Automatic phenotype concept recognition from text is of great significance for disease analysis. Existing ontology-driven phenotype concept recognition methods mainly utilize concepts and synonyms information within the ontology, without fully considering the rich information of the ontology. To address this issue, this paper proposes an ontology-enhanced method for human phenotype concept recognition, which utilizes advanced large language models for data augmentation and designs a deep learning model enhanced with ontology vectors to improve concept recognition performance. Experimental results on the GSC+ and ID-68 datasets demonstrate that the proposed method effectively leverages the rich ontology information to enhance the performance of baseline models, achieving state-of-the-art results.

Keywords: Biomedical Text Mining , Human Phenotype Ontology , Ontology Information Enhancement

1 引言

* 通讯作者

基金项目：国家自然科学基金（62302076）；中央高校基本科研业务费（DUT23RC (3)014）；SMP-智谱AI大模型交叉学科基金（ZPCG2024010204)

表型是基因型和环境共同作用的生物体的特定物理外观，是疾病背后生理病理过程的表现，了解表型和疾病之间的关联有助于重大疾病的诊断和治疗。人类表型本体(Human Phenotype Ontology, HPO)全面收集和定义了人类疾病的表型特征，并提供了一套标准化词汇用于描述人类表型异常，在世界上被广泛使用 (Robinson et al., 2008; Köhler et al., 2019; Köhler et al., 2021)。从海量生物医学文本中自动识别出HPO中的表型概念，对于理解人类发病复杂原因，推动生物医学研究的发展，以及提高罕见疾病的临床诊断准确率等方面都具有重要的意义 (Groza et al., 2015)。

人类表型概念识别任务旨在从非结构化的生物医学文本中自动抽取表型概念，如表1所示，该任务包括命名实体识别(Named Entity Recognition, NER)和实体标准化(Named Entity Normalization, NEN)两个子任务，即从文本中识别出人类表型实体并映射到HPO中的标准术语上(HPO ID号)。但目前没有充足的表型概念识别人工语料库资源用来训练基于有监督学习的NER和NEN模型进行流水线识别。因此，大多现存表型概念识别工具主要采用基于词典的方法，例如OBO(Open Biological and Biomedical Ontologies) (Taboada et al., 2014)、NCBO(National Center for Biomedical Ontology) (Jonquet et al., 2009)、Doc2hpo (Liu et al., 2019)等。这些方法通过利用HPO中的概念标准名和提供的同义词构造词典，然后设计基于文本匹配的方法进行HPO概念识别。这类方法简单易实现，并获得了较高的识别准确度，但无法识别词典中不存在的表型概念同义词，存在召回率低的问题。

输入文本	A syndrome of brachydactyly (absence of some middle or distal phalanges), aplastic or hypoplastic nails, symphalangism...			
概念识别	14	27	brachydactyly	HP:0001156
	29	71	absence of some middle or distal phalanges	HP:0009881
	73	103	aplastic or hypoplastic nails	HP:0001798
	...	...	...	...

Table 1: 人类表型概念识别样例

近年来，针对基于词典方法召唤率低的问题，研究者们提出了本体驱动的深度学习方法来识别表型概念，例如NeuralCR (Arbabi et al., 2019)、PhenoBERT (Feng et al., 2022)和PhenoTagger (Luo et al., 2021)等。这类方法将NER和NEN两个子任务转换成一个文本分类任务，通过使用HPO中每个概念的标准名称和同义词构建远程监督数据集，用来训练深度学习分类模型。识别阶段将输入文本转化为n元词作为概念候选，并将其输入训练好的分类器进行HPO标识符分类，高于一定阈值的候选被识别为表型概念。这类方法不依赖于人工标注的数据集，同时利用深度学习模型有效提升了概念识别的召回率，在多个表型概念识别测试集上获得了目前先进结果。

上述本体驱动的深度学习方法虽然取得了先进结果，但多数方法主要利用了HPO中的概念名和提供的同义词，其他本体信息并未得到充分利用。而在HPO中，不仅包含了表型概念的同义词信息，还提供了概念定义、概念的上下位关系等丰富信息，这些信息都有助于人类表型概念识别任务。如图1所示，HPO本体结构为树状层次结构，存在上下位关系的术语通常具有十分密切的关联信息。例如“Neoplasm by histology”是其父节点“Neoplasm”更具体的一种表型，这类层次结构信息对于模型识别和区分相似表型概念有重要作用。此外，HPO中多数概念只提供了一个同义词，甚至没有提供同义词。在基于远程监督的深度学习方法中，同义词能够提供更丰富的语义特征，有助于提升分类模型的性能，因此如何对本体中没有同义词信息的概念进行数据增强也值得进一步研究。

最近，大语言模型(Large Language Models, LLMs)在自然语言处理领域的表现引起了广泛的关注，它具有良好的自然语言生成能力、上下文理解能力和知识融合的能力，在处理复杂的自然语言任务方面如命名实体识别、关系抽取、机器翻译等任务上取得优异的性能 (Wang et al., 2023b; Zhou et al., 2023; Wang et al., 2023a)。在一些例如医学、法律、金融等特定领域 (Luo et al., 2024; Cui et al., 2023; Chen et al., 2023)，大模型也可以生成高质量、专业化的内容。目前LLMs在人类表型概念识别任务上处于初步探索阶段，Groza (2024)等人使用GPT-3.5-turbo和GPT-4模型通过构建提示模板对人类表型概念进行抽取，但存在资源耗费大的问题，且大模型生成概念位置和HPO标签ID并不准确，识别结果低于传统方法。

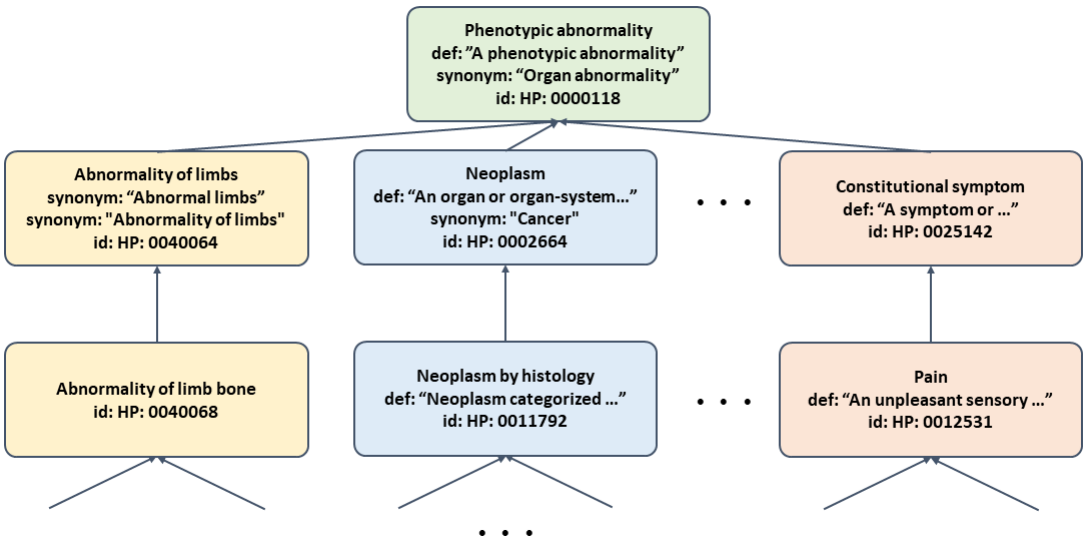

Figure 1: HPO中的信息结构

针对这些问题，本文提出一种基于本体信息增强的人类表型概念识别方法。该方法在先进的本体驱动方法的基础上，借助大语言模型的文本生成能力，利用丰富的HPO本体信息探索了数据增强和模型增强两方面来提升人类表型概念识别的性能。在数据增强方面，利用当前先进的大语言模型GPT-4产生概念术语同义词来增强远程监督训练集；在模型增强方面，通过知识表示方法将HPO的本体结构信息转换为本体向量，用其作为先验知识融入深度学习模型提升模型性能。本文的主要贡献总结如下：

(1) 利用HPO中概念名、概念定义等信息设计大语言模型的提示模板(Prompt)，借助目前先进的大语言模型GPT-4为HPO中缺乏同义词的概念生成其同义词，进行远程监督训练集的数据增强。

(2) 提出了一种融合本体结构向量的深度学习模型对表型概念进行识别。通过使用知识表示学习方法将HPO层次结构信息转换为本体向量，将其与预训练语言模型学习到的文本特征进行融合，提升模型性能。

(3) 在GSC+和ID-68两个表型概念识别数据集上进行实验验证，结果表明本文提出的数据增强和本体向量增强均能提升基础模型性能，两者融合后概念识别性能得到了进一步提升，在两个数据集上取得了最高的平均F1值0.805。这说明了本文方法能够更有效地利用本体丰富信息，验证其有效性和鲁棒性。

2 相关工作

最近，人类表型概念识别任务受到了广泛的关注，但由于人工标注语料库稀缺以及表型概念表达多样等问题，人类表型概念识别仍然是生物医学文本处理中的一项具有挑战的任务。现存人类表型概念识别工具主要基于词典匹配方法，包括OBO (Taboada et al., 2014)、NCBO (Jonquet et al., 2009)、Doc2hpo (Liu et al., 2019)、MI(Monarch Initiative) (Shefchek et al., 2020)等。这些基于词典的方法主要是利用HPO构造词典，并通过文本匹配技术对表型概念进行识别。尽管这些方法在精确度上表现出色，但是无法识别出词典中未包含的的概念同义词，存在召回率低的问题。近年来，随着深度学习技术的快速发展，研究者们提供了本体驱动的深度学习方法来进行表型概念识别。由于缺乏人工标记数据集，这些方法大多使用HPO构建远程监督数据集训练深度学习模型进行识别。其中NeuralCR (Arbabi et al., 2019)方法使用稀疏矩阵表示两个概念节点之间的上下位关系，然后融入卷积神经网络(Convolutional Neural Network, CNN)进行概念识别，有效提升了识别召回率，但该方法无法抽取重叠实体概念。PhenoBERT (Feng et al., 2022) 进一步根据本体的层次结构，设计训练了分层CNN结构，能够更准确地识别表型概念。PhenoTagger (Luo et al., 2021)是一种词典和深度学习的混合方法，在词典基础上结合了先进的预训练语言模型BioBERT，同时考虑了重叠概念的抽取，在人

类表型概念识别任务上取得了先进结果，但该方法仅利用了HPO中的概念名和同义词信息。本文方法在先进PhenoTagger基础上，重点研究利用丰富本体信息提示模型性能。

为了更充分利用本体结构信息，本文研究了知识表型学习方法来融合本体结构信息。知识表示学习方法旨在将知识库中的实体和关系嵌入到低维稠密的实值向量空间中，如TransE (Bordes et al., 2013)、TransR (Lin et al., 2015)、ConvE (Dettmers et al., 2018)和SemNE (Wang et al., 2021)等。TransE是一种用于知识表示学习的模型，将实体和关系使用向量分布式表示，通过调整向量以近似表示头实体加上关系等于尾实体这一过程。虽然这种方法简单高效，但无法处理复杂关系。为了解决这个问题，TransR被提出，它将实体和关系嵌入到不同的语义空间中，以更好地捕捉复杂相关性。Trans系列模型是浅层模型，简单且参数较少，但在抓取复杂信息方面有所限制。ConvE是一深度学习模型，融合了卷积神经网络和图嵌入技术将实体和关系嵌入到低维向量中，利用卷积操作捕获实体和关系之间的信息。SemNE是一种语义网络编码器，通过学习特征映射函数，并以预训练方式应用于多路语义网络，联合建模语义信息和网络拓扑结构，丰富了网络表示，提高了语义网络的表达能力。

3 本体信息增强的表型概念识别方法

本文方法构建了本体信息增强的深度学习标注器。在深度学习标注器中，首先利用HPO构建远程监督训练集，并探索了使用GPT-4产生同义词进行数据增强，然后用其训练本体向量增强的表型概念分类模型，最后对输入文本产生n元候选通过训练好的分类器得到深度学习的识别结果。此外，构建了词典标注器进行结果集成，得到最终的表型概念识别结果。

3.1 深度学习标注器

3.1.1 大语言模型增强的远程监督训练集构建

由于表型概念识别人工语料库缺乏，本体驱动方法主要利用HPO中每个概念名、同义词及对应的概念ID，构建远程监督训练集。对于每条概念，将其概念名和其同义词分别产生<**X**:术语文本, **Y**:HPO ID>实例对作为训练集的正例，同时，从生物医学文献中构造一些不属于任何已有概念的负例。即从海量无标注生物医学文本中随机抽取n个连续单词，若不在HPO中，则为其分配新的标签“HP:None”作为负例。在构建远程监督训练集的过程中，本文对HPO本体中的概念进行统计分析，结果如图2所示。其中超过40%的概念未提供同义词信息，其余8424条概念包含了同义词，但这之中约50%的概念只包含一条同义词信息。同义词信息的缺乏致使这些概念在训练集中仅有一条样本，从而限制了模型对它们语义信息的学习能力。

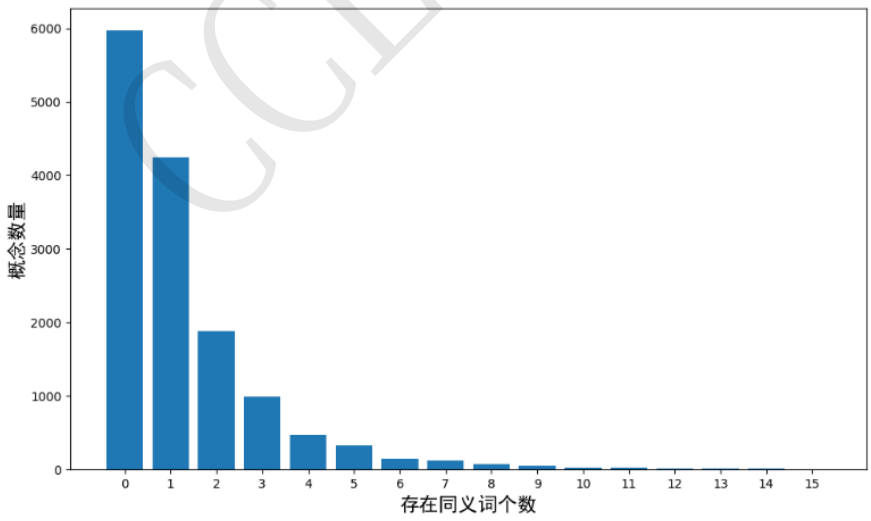

Figure 2: 人类表型概念同义词数量分析

为了缓解同义词信息不足的问题，本文采用了先进的大语言模型GPT-4进行数据增强。先前研究已表明 (Shanahan et al., 2023)，通过设计详细的提示工程(例如，专业领域专家的角色扮演、语义完整的提示模板、提示样例、思维链等)可以让大语言模型显著提升其在该领域

的表现。因此，本文在Prompt设定中将大语言模型设定为医学专家的角色，并通过类似思维链 (Wei et al., 2022)的方式，逐步分析概念定义中涉及的症状、病因等要素，并结合人类表型本体的层次结构信息，以及同义词与原始术语之间的语义一致性，通过具体示例来激发大语言模型的理解能力。具体Prompt请参考**附录A.同义词生成Prompt构建**。

3.1.2 本体向量增强的表型概念分类模型

基于Transformer (Vaswani et al., 2017)的预训练语言模型，如BERT (Devlin et al., 2018)等，在自然语言处理领域中的各项任务上，相比较于词向量模型、循环神经网络和长短期记忆网络等传统方法有了性能上的提升。本文选择使用生物医学版本的预训练语言模型，如BioBERT (Lee et al., 2020)、Bioformer (Fang et al., 2023)和PubMedBERT (Gu et al., 2021)等作为文本编码器。

本文在生物医学预训练模型基础上，利用知识表示方法获取本体向量，提出一种本体向量增强的表型概念分类模型，模型结构如图3所示。与原始的BERT类似，BioBERT等三种预训练模型也使用WordPiece (Wu et al., 2016)作为嵌入方式，将远程监督训练集中的$\mathbf{X}$作为输入，通过预训练模型编码后得到其隐层向量。具体而言，以输入文本$x_i = (w_1, w_2, \ldots, w_n)$，在每个输入前插入一个特殊标签[**CLS**]得到输入文本序列x_i^{in}:

$$x_i^{in} = \{[\mathbf{CLS}], w_1, w_2, ..., w_n\} \tag{1}$$

然后使用预训练模型对x_i^{in}进行文本特征提取，采用[**CLS**]对应的向量作为输入文本表示v_i。其计算公式如下:

$$v_i = \mathbf{PreTrainModel}(x_i^{in}; \mathbf{W}_0) \tag{2}$$

其中，**PreTrainModel**表示使用预训练语言模型进行编码的过程如BioBERT等,$\mathbf{W}_0$为预训练语言模型的可训练参数。

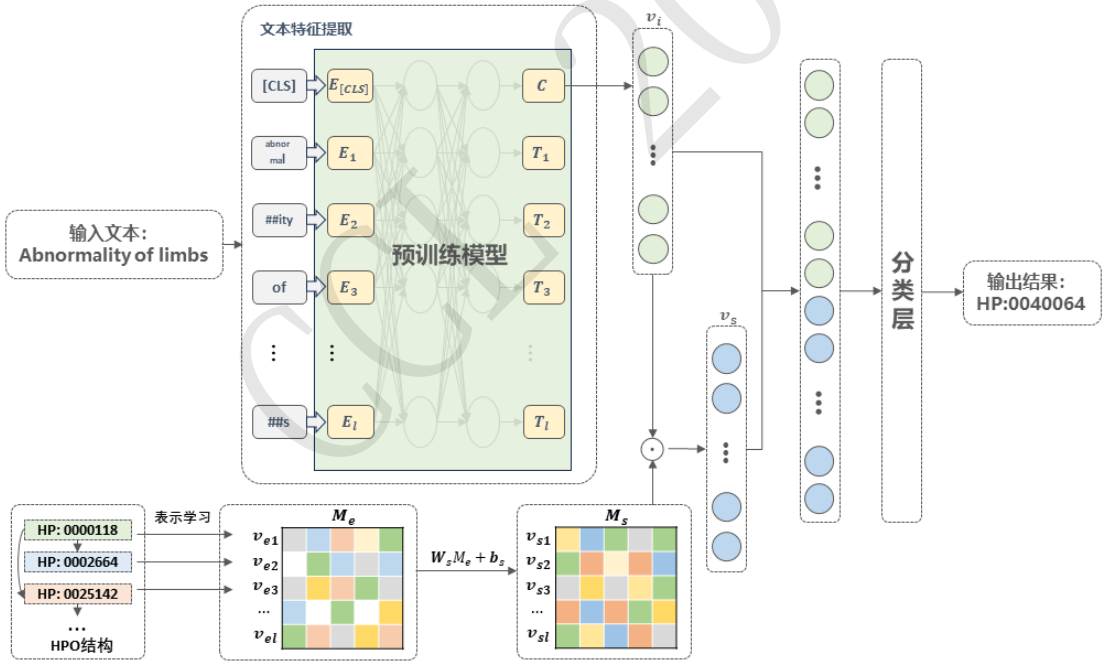

Figure 3: 本体向量增强模型结构图

在本体向量构建部分，本文采用了四种不同的知识表示学习方法对HPO结构信息进行表示。用T={V,E}表示HPO树状结构，其中V表示HPO中的概念节点，E表示概念节点之间的边。通过知识表示学习方法对HPO树T提取结构嵌入矩阵M_e。在M_e中，每一行表示一个概念实体的嵌入向量。具体过程如下:

$$M_e = \mathbf{RL}(T; \mathbf{W}_1) \tag{3}$$

其中，**RL**代表使用知识表示学习方法提取结构矩阵的过程，如TransE等，W_1是训练过程中的可训练参数。然后通过一个自定义的结构信息学习层得到结构嵌入矩阵的信息，得到HPO结构信息矩阵M_s，其计算公式如下：

$$M_s = \boldsymbol{W}_s M_e + \boldsymbol{b}_s \tag{4}$$

其中，$\boldsymbol{W}_s$和$\boldsymbol{b}_s$为可学习的参数和偏置。之后将结构信息矩阵M_s和输入编码向量v_i进行点积运算，得到融合结构信息的向量$v_s = M_s \odot v_i$，其中$\odot$表示点积运算。最后将编码向量和融合向量进行拼接，通过一个分类层进行分类得到最后的分类结果y_i'，即：

$$y_i' = Softmax\big(Dense([v_i; v_s]; \mathbf{W}_2)\big) \tag{5}$$

其中$\mathbf{W}_2$为分类层的可学习参数。通过以上操作可以得到深度学习方法的预测结果y_i'。

最后，本文采用稀疏交叉熵作为损失函数。其损失计算如下：

$$L(y_i, y_i') = -\frac{1}{N}\sum_{i=1}^{N} y_i log(y_i') \tag{6}$$

其中y_i是远程监督数据集中$\mathbf{Y}$的真实标签。

3.1.3 概念识别过程

在识别阶段，首先使用NLTK (Bird et al., 2009)将输入文本分句、分词和词性标注，然后生成句子中所有的n元词作为候选概念。接下来，对候选概念使用规则方法进行过滤，即根据词性删除以标点符号、介词和连词开始或结束的候选。接下来使用训练好的预训练模型对每个候选对象进行分类。最后，被分类HPO ID标签且预测分数(即预训练模型预测的标签概率分数P)高于阈值t的候选项被识别为HPO概念，其中t是一个超参数，根据本文方法在开发集上的性能进行选择。

3.2 词典标注器与结果合并

在PhenoTagger研究中表明融合基于词典方法和深度学习方法可进一步为了提高识别结果的准确率，本文采用同样策略进行词典方法识别结果融合。首先使用HPO中的概念名和同义词构造词典，采用Trie树数据结构 (Fredkin, 1960)来存储HPO词典，进行精确的字符串匹配识别概念。得到词典匹配和深度学习识别的结果后，根据下面四条规则和候选结果得分(词典结果赋予1分，深度学习方法识别结果分值为模型分类概率分值)来获得最终识别结果：1)保留所有不重叠的概念；2)如果重叠的概念具有相同的概念ID，则保留得分最高的概念；3)如果重叠概念在文本中起始和结束位置相同，但被映射到不同的概念ID中，则保留得分最高的概念；4)如果重叠的概念在文本中有不同的开始或结束位置，不同的概念ID，则所有重叠的概念都被保留。

4 实验结果与分析

4.1 数据集和实验设置

本文主要使用HPO Gold Standardized Corpus plus (GSC+) (Lobo et al., 2017)数据集和ID-68 (Feng et al., 2022)数据集进行实验。随机选择GSC+数据集的10%作为开发集用于超参数选择，其余的数据用作测试集。ID-68数据集包括68份真实的临床记录 (Anazi et al., 2017)，数据集统计结果如表2。其中GSC+数据集规模相对较大，用作主数据集测试方法各模块有效性，将GSC+开发集上设置好的模型直接在ID-68数据集上进行测试。

	医学文本数	概念数量	唯一概念数量
GSC+	228	2122	1433
ID-68	68	866	800

Table 2: 数据集统计

本文采用在开发集上进行随机搜索的方式选择超参数，其中学习率设置为5×10^{-6}，batch size大小设置为64，n元词的最大长度设置为10，深度学习标注器的分类阈值t设置为0.95，并根

据HPO中概念最大长度将n元词的最大长度设置为10。使用概念识别任务上广泛采用的评价指标即文档级别的宏平均的Precision，Recall和F1值作为衡量指标 (Luo et al., 2021)。文档级别只需考虑每个文档中的标签集合而忽略文本中的确切位置。

4.2 大语言模型数据增强的实验结果

为了验证使用大语言模型增强远程监督训练集的有效性，本节考察了在仅使用GPT-4增强训练集而不改变模型结构的情况下对人类表型概念识别性能的影响。针对此目的，本节中对比三种基座模型BioBERT (Lee et al., 2020)、Bioformer (Fang et al., 2023)和PubMedBERT (Gu et al., 2021)的性能。本文使用GPT-4对不包含同义词的术语概念产生3个同义词，经过实验对比证明在使用2个同义词时性能最优，详细对比结果请参考**附录B.同义词数量在GSC+数据集上对模型的性能影响**。下述实验均使用2个同义词进行模型性能对比。使用GPT-4对HPO同义词增强后在GSC+测试集上的性能对比结果如表3所示，其中Dict表示使用词典匹配方法，DL表示深度学习(括号中为不同基座模型下的结果)，Hybrid为词典匹配方法和深度学习方法结合的结果(括号中为不同基座模型下的结果)。

方法	未用GPT-4增强			使用GPT-4增强		
	Precision	Recall	F1	Precision	Recall	F1
Dict	0.756	0.565	0.647	0.765	0.587	0.664
DL(BioBERT)	0.754	0.700	0.726	0.762	0.736	0.749
DL(Bioformer)	0.767	0.621	0.686	0.743	0.700	0.721
DL(PubMedBERT)	0.755	0.706	0.730	0.749	0.720	0.734
Hybrid(BioBERT)	0.766	0.712	0.738	**0.768**	**0.751**	**0.760**
Hybrid(Bioformer)	**0.773**	0.674	0.720	0.760	0.719	0.739
Hybrid(PubMedBERT)	0.762	**0.725**	**0.743**	0.766	0.736	0.751

Table 3: GPT数据增强在GSC+测试集上的结果。其中Dict表示使用词典匹配方法，DL表示深度学习方法，Hybrid表示词典和深度学习混合方法，加粗表示最高值。

实验结果表明：(1) 使用GPT-4生成同义词后，对词典匹配方法的准确率和召回率分别提升了0.9%和2.2%，表明GPT-4生成的同义词中包含了原有词典中不存在的正确概念；(2) 使用GPT-4扩充远程监督训练集也能提高深度学习方法的性能，其中在Bioformer模型上提升最为明显，达到了3.5%，在BioBERT模型上的提升也达到了2.3%；(3)混合词典和深度学习方法后，GPT-4增强的BioBERT模型获得了最高F1值。这说明使用GPT-4生成同义词后，能够改善HPO中缺少同义词概念的识别性能，并且对于原始识别性能较弱的基座模型提升效果更加明显。

4.3 本体向量增强的实验结果

为了验证使用本体结构信息增强概念识别的有效性，本节考察了在不使用大语言模型增强训练集的情况下，仅使用本体结构信息增强模型的识别性能。主要探究不同本体向量维度和不同本体表示方法对模型性能的影响。

4.3.1 本体向量维度对模型性能的影响

本实验主要探索不同嵌入矩阵中向量维度对人类表型概念识别的影响，使用SemNE方法作为结构信息嵌入向量生成方法，BioBERT作为基座模型，在GSC+测试集上进行对比，实验结果如图4所示。

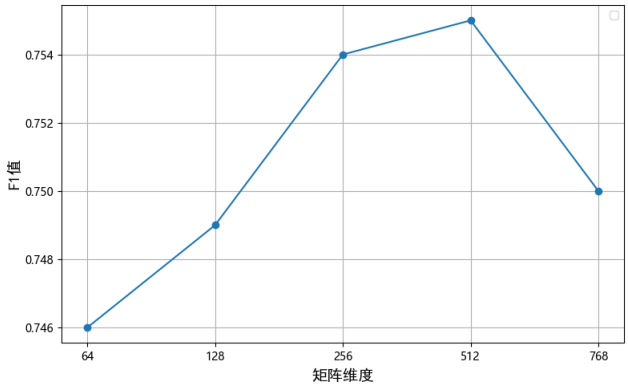

Figure 4: 不同维度的本体向量在GSC+数据集上对模型的性能影响

从实验结果可以看出，随着维度从64增加至512，模型性能持续提升，但在768时性能出现下降。这可能是由于随着向量维度的增加，嵌入向量中包含的结构信息变得更加丰富，但进一步增加维度可能导致一些过拟合情况，从而影响了性能。因此，后续实验选择512作为本体向量的维度。

4.3.2 本体向量表示方法对模型性能的影响

本实验主要探索不同知识表示方法对人类表型概念识别的影响，分别使用TransE、TransR、ConvE和SemNE四种方法作为结构信息嵌入向量的提取方法，维度设置为512，同时，测试了BioBERT、Bioformer和PubMedBERT不同的基座模型，以评估不同提取方法的通用性。在GSC+数据集上的实验结果如图5所示，其中“Original”表示不使用嵌入向量，“Random”表示使用随机函数随机生成嵌入向量。

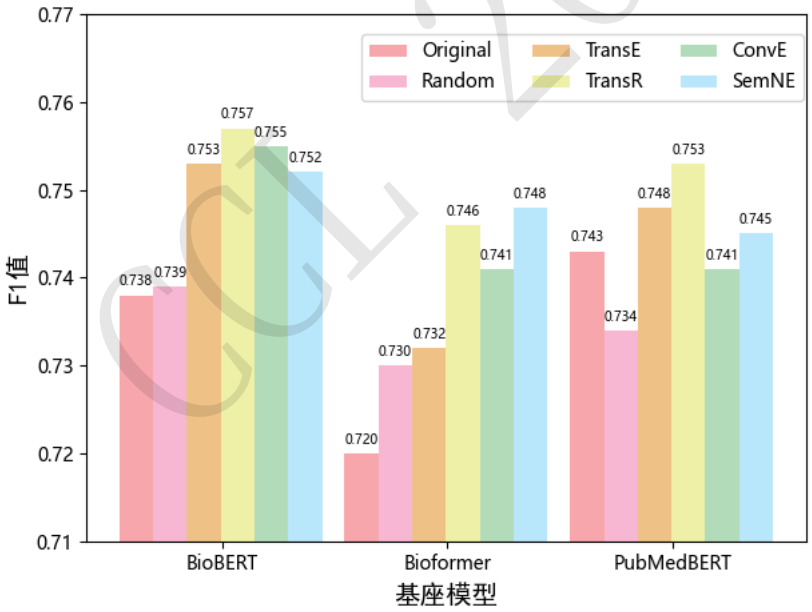

Figure 5: 不同知识表示方法在GSC+数据集上对模型的性能影响

从图中结果可以观察到：(1) 三种基座模型在添加不同方法构建的结构信息嵌入向量后，均提升了人类表型概念识别的性能。特别是在使用BioBERT作为基座模型时，并使用TransR作为结构信息向量提取方法时表现最佳；(2) 在不同模型上，使用TransR作为嵌入向量生成方法获得了最高的平均F1值0.752，SemNE的平均性能也达到了0.748。这表明模型可以从TransR和SemNE提取到的结构信息嵌入向量中学到最丰富的信息；(3) 为了验证使用不同方法提取结构信息向量的有效性，本文使用随机函数随机生成嵌入向量，并将其添加到三种

基座模型中。结果显示，相比于其他嵌入向量生成方法，随机生成的嵌入向量在三种基座模型上性能较差，进一步验证了使用不同方法进行结构信息向量提取可以提高表型概念的识别性能。

4.3.3 数据增强方法对模型性能的影响

本实验主要探索不同数据增强方法对模型性能的影响，本文选择基于BioBERT的PhenoTagger作为基线模型（Baseline），增加了三种常用的数据增强方法进行比较：1）SynonymAug：根据WordNet中的距离计算相似度从而选取同义词进行数据增强；2）RandomWordAug：对概念中的词汇随机选取一个其他单词进行替换进行数据增强；3）SpellingAug：引入拼写错误的进行数据增强。Our为本文中使用GPT进行增强的模型。在GSC+数据集上的实验结果如表4所示。

方法	GSC+数据集		
	Precision	Recall	F1
Baseline	0.766	0.712	0.738
SynonymAug	**0.774**	0.723	0.745
RandomWordAug	0.295	0.661	0.407
SpellingAug	0.754	0.725	0.739
Ours	0.768	**0.751**	**0.760**

Table 4: 数据集统计

从表4中可以观察到：（1）在使用WordNet同义词（SynonymAug）和拼写错误（SpellingAug）的数据增强时，相对于Baseline分别提升了0.7%和0.1%，但是在使用随机替换（RandomWordAug）的增强方法时，模型性能下降了约30%，其原因可能是随即替换引入了过多的噪声，导致模型无法进行识别；（2）使用GPT-4进行数据增强对比其余几种方法在性能上的提升更加明显，其原因可能是在使用GPT-4产生同义词的时候即考虑了术语概念也同时考虑到了术语的上下位关系，因此可以获得在语义上更加相似的增强数据。

4.4 与其他现存方法性能对比结果

前序实验分别验证了利用GPT-4进行数据增强和本体向量增强方法的有效性，本节进一步将两者进行合并，具体使用BioBERT作为基座模型，本体向量表示方法使用TransR，同时加入GPT-4增强的数据进行训练。在GSC+和ID-68两个数据集上将本文方法(Ours)与现有先进方法进行了比较，验证其有效性，结果如表5所示，使用领域内常用的NEN指标进行性能评估 (Luo et al., 2021)，平均F1值由GSC+数据集和ID-68数据集的F1值求平均得到。本文方法在NER和NEN两个子任务上更详细的实验结果请参考**附录C.本文方法详细结果**。

- OBO: 专门用于注释具有HPO表型异常的生物医学文本工具，使用基于词汇和上下文的匹配方法。
- MI(Monarch Initiative) (Shefchek et al., 2020): Monarch Initiative是一个集成数据和分析平台，将不同物种的表型与基因型联系起来，允许用户输入自由文本，并使用Monarch知识图中的术语对文本进行自动表型注释。
- Doc2hpo (Liu et al., 2019): 一个从临床文本中提取人类表型本体的方法，使用了Doc2hpo的集成引擎，该引擎结合了多个基于字符串的方法。
- NeuralCR (Arbabi et al., 2019): 一种基于CNN的人类表型概念识别模型。
- Txt2hpo(github.com/GeneDx/txt2hpo/tree/develop/txt2hpo): 一个用于从文本中提取HPO编码的表型的Python库。
- PhenoBERT (Feng et al., 2022): 一个以BERT为核心，在BERT前引入了一个两层的CNN模块的深度学习方法。

- GPT-3.5/GPT-4: Groza (2024)使用不同的Prompt在GPT-3.5和GPT-4上完成人类表型概念识别任务，本文选取两种模型最好的结果作为对比。
- PhenoTagger (Luo et al., 2021): 一种将词典匹配和深度学习方法结合的混合方法。

方法	GSC+数据集			ID-68数据集			平均F1
	Precision	Recall	F1	Precision	Recall	F1	
OBO	0.809	0.565	0.665	0.852	0.622	0.719	0.692
MI	0.757	0.605	0.673	0.879	0.438	0.584	0.629
Doc2hpo*	0.768	0.618	0.685	0.859	0.580	0.692	0.689
NeuralCR*	0.741	0.602	0.664	0.796	**0.794**	0.795	0.730
Txt2hpo	**0.825**	0.562	0.669	0.909	0.683	0.780	0.725
PhenoBERT	0.813	0.640	0.716	**0.946**	0.776	**0.852**	0.784
GPT-3.5*	0.410	0.410	0.410	-	-	-	-
GPT-4*	0.750	0.470	0.580	-	-	-	-
PhenoTagger	0.766	0.712	0.738	0.916	0.766	0.834	0.786
Ours(GPT)	0.768	**0.751**	0.760	0.912	0.780	0.841	0.801
Ours(EMB)	0.786	0.730	0.757	0.928	0.770	0.842	0.799
Ours(GPT+EMB)	0.783	0.744	**0.763**	0.924	0.781	0.846	**0.805**

Table 5: 与现存方法对比结果。其中GPT表示使用GPT-4扩充数据，EMB表示使用本体向量增强。*表示结果引用于原论文，加粗表示最高值。

从表5结果中可以观察到：(1) 在GSC+和ID-68两个数据集上，相比原始的PhenoTagger，本文提出的本体向量增强和GPT数据增强两种改进方法都取得了性能上的提升，将两种方法合并后获得了进一步的提升，在两个数据集上取得了最高平均F1值，为0.805，相比目前先进的基于深度学习的本体驱动方法PhenoTagger和PhenoBERT在平均F1值上分别提升了1.9%和2.1%，表明了本文方法的有效性和鲁棒性；(2) 在GSC+数据集上，与基于词典的方法相比，本文方法在准确率方面略有降低，但在召回率方面平均提升了近15%。而在ID-68数据集上，相较于基于词典的方法，本文方法在精确度和召回率方面平均都有超过8%的提升。模型在ID-68数据集上的效果优于在GSC+数据集上的效果，其原因可能是ID-68数据集仅包含68篇电子病历，其中概念数量为866，唯一概念数量为800；而GSC+数据集包含228篇生物医学文献摘要，其中概念数量为2122，唯一概念数量为1433。相比GSC+数据集，ID-68数据集规模较小而且表型概念类型较少，识别难度相对较低，因此模型在ID-68数据集上的效果明显优于在GSC+数据集上的效果；(3)在ID-68数据集上，本文方法略低于PhenoBERT方法，其原因可能是不同于数据集GSC+来源于生物医学文献，ID-68数据集由68篇电子病历组成。PhenoBERT在产生概念候选时，使用了Stanza工具中的"ner-i2b2"进行了电子病历实体识别，而本文方法只使用了n-gram产生候选。在电子病历上，PhenoBERT利用额外工具产生了更准确的实体候选，因此在电子病历文本ID-68上，本文方法略差于PhenoBERT（F值相差0.006），但根据实验结果在两个数据集上本文方法获得了最高的平均F1值，说明了本文方法的鲁棒性和泛化能力；(4) 直接使用大语言模型GPT3.5/4进行人类表型识别，结果低于词典方法，这可能受限于目前大语言模型对HPO中概念ID知识学习，概念识别这类任务对先进大语言模型仍具有巨大挑战，而本文使用GPT进行数据增强能够进一步提升模型性能。综上，本文提出的方法能够有效提升原始模型的识别性能，在两个数据集均获得了先进结果。

4.5 识别结果样例分析

为了深入验证本文方法的有效性，本文对数据增强最优模型和本体向量增强最优模型结果上进行了人工分析，选取了两个具体案例进行展示。

表6显示了在使用GPT增强远程监督训练集前后对同一条生物医学文本的识别结果。可以观察到，在使用GPT增强后，成功识别出了原有模型遗漏的概念，并且概念的分类和跨度均被正确识别。表7则展示了在引入结构信息前后对同一条生物医学文本的识别结果。在引入结构信息前，模型识别错误的概念，并且分类也出现错误。在当前使用的HPO版本

中，“HP:00009371”是“HP:00001156”的孙子节点，在引入结构信息后，成功识别出概念的位置和类别。这两个实例表明了本文方法的可行性。

输入文本：“... have orbital haemangiomatous cysts, ...”				
正确答案	78	107	orbital haemangiomatous cysts	HP:0001144
未使用GPT增强	\	\	\	\
GPT增强	78	107	orbital haemangiomatous cysts	HP:0001144

Table 6: 使用GPT增强后性能提升样例

输入文本：“Brachydactyly type A-1 (BDA1) was, in 1903...”				
正确答案	0	22	Brachydactyly type A-1	HP:0009371
未引入结构信息	0	13	Brachydactyly	HP:0001156
引入结构信息	0	22	Brachydactyly type A-1	HP:0009371

Table 7: 引入结构信息后性能提升样例

5 结论与展望

针对目前人类表型概念识别方法未充分使用本体信息的问题，本文提出从数据增强和模型增强两方面利用本体丰富信息提升模型识别性能。在数据增强方面，通过本体信息设计提示模板，利用先进大语言模型GPT-4产生同义词进行远程监督数据增强；在模型增强方面，提出本体向量增强的深度学习模型，通过设计知识表示向量融合本体结构信息提示模型性能。将两种方法融合后，在两个测试集上均取得了性能的进一步提升，验证了方法的有效性。

在未来工作中，我们将深入探索更先进的知识表示方法来充分利用除结构层次信息外的其他丰富信息，例如术语定义、术语评论等。同时，本文方法基于本体驱动的概念识别方法，不依赖于人工标注数据集，未来也将在更多生物医学本体上进行测试。此外，大语言模型展示出了巨大潜力，探索基于大语言模型的概念识别生成式框架也是我们未来工作的方向。

参考文献

Shams Anazi, Sateesh Maddirevula, Vincenzo Salpietro, Yasmine T Asi, Saud Alsahli, Amal Alhashem, Hanan E Shamseldin, Fatema AlZahrani, Nisha Patel, Niema Ibrahim, et al. 2017. Expanding the genetic heterogeneity of intellectual disability. *Human genetics*, 136:1419–1429.

Aryan Arbabi, David R Adams, Sanja Fidler, Michael Brudno, et al. 2019. Identifying clinical terms in medical text using ontology-guided machine learning. *JMIR medical informatics*, 7(2):e12596.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. ” O’Reilly Media, Inc.”.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems*, 26.

Wei Chen, Qiushi Wang, Zefei Long, Xianyin Zhang, Zhongtian Lu, Bingxuan Li, Siyuan Wang, Jiarong Xu, Xiang Bai, Xuanjing Huang, and Wei Zhongyu. 2023. Disc-finllm: A chinese financial large language model based on multiple experts fine-tuning. *arXiv preprint arXiv:2310.15205*.

Jiaxi Cui, Zongjian Li, Yang Yan, Bohua Chen, and Li Yuan. 2023. Chatlaw: Open-source legal large language model with integrated external knowledge bases. *arXiv preprint arXiv:2306.16092*.

Tim Dettmers, Pasquale Minervini, Pontus Stenetorp, and Sebastian Riedel. 2018. Convolutional 2d knowledge graph embeddings. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Li Fang, Qingyu Chen, Chih-Hsuan Wei, Zhiyong Lu, and Kai Wang. 2023. Bioformer: an efficient transformer language model for biomedical text mining. *ArXiv*.

Yuhao Feng, Lei Qi, and Weidong Tian. 2022. Phenobert: a combined deep learning method for automated recognition of human phenotype ontology. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 20(2):1269–1277.

Edward Fredkin. 1960. Trie memory. *Communications of the ACM*, 3(9):490–499.

Tudor Groza, Sebastian Köhler, Sandra Doelken, Nigel Collier, Anika Oellrich, Damian Smedley, Francisco M Couto, Gareth Baynam, Andreas Zankl, and Peter N Robinson. 2015. Automatic concept recognition using the human phenotype ontology reference and test suite corpora. *Database*, 2015:bav005.

Tudor Groza, Harry Caufield, Dylan Gration, Gareth Baynam, Melissa A Haendel, Peter N Robinson, Christopher J Mungall, and Justin T Reese. 2024. An evaluation of gpt models for phenotype concept recognition. *BMC Medical Informatics and Decision Making*, 24(1):30.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Clement Jonquet, Nigam Haresh Shah, and Mark A. Musen. 2009. The open biomedical annotator. *Summit on Translational Bioinformatics*, 2009:56 – 60.

Sebastian Köhler, Leigh Carmody, Nicole Vasilevsky, Julius O B Jacobsen, Daniel Danis, Jean-Philippe Gourdine, Michael Gargano, Nomi L Harris, Nicolas Matentzoglu, Julie A McMurry, et al. 2019. Expansion of the human phenotype ontology (hpo) knowledge base and resources. *Nucleic acids research*, 47(D1):D1018–D1027.

Sebastian Köhler, Michael Gargano, Nicolas Matentzoglu, Leigh C Carmody, David Lewis-Smith, Nicole A Vasilevsky, Daniel Danis, Ganna Balagura, Gareth Baynam, Amy M Brower, et al. 2021. The human phenotype ontology in 2021. *Nucleic acids research*, 49(D1):D1207–D1217.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. 2015. Learning entity and relation embeddings for knowledge graph completion. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29.

Cong Liu, Fabricio Sampaio Peres Kury, Ziran Li, Casey N. Ta, Kai Wang, and Chunhua Weng. 2019. Doc2hpo: a web application for efficient and accurate hpo concept curation. *Nucleic Acids Research*, 47:W566 – W570.

Manuel Lobo, Andre Lamurias, and Francisco M Couto. 2017. Identifying human phenotype terms by combining machine learning and validation rules. *BioMed Research International*, 2017.

Ling Luo, Shankai Yan, Po-Ting Lai, Daniel Veltri, Andrew Oler, Sandhya Xirasagar, Rajarshi Ghosh, Morgan Similuk, Peter N Robinson, and Zhiyong Lu. 2021. Phenotagger: a hybrid method for phenotype concept recognition using human phenotype ontology. *Bioinformatics*, 37(13):1884–1890.

Ling Luo, Jinzhong Ning, Yingwen Zhao, Zhijun Wang, Zeyuan Ding, Peng Chen, Weiru Fu, Qinyu Han, Guangtao Xu, Yunzhi Qiu, Dinghao Pan, Jiru Li, et al. 2024. Taiyi: a bilingual fine-tuned large language model for diverse biomedical tasks. *Journal of the American Medical Informatics Association: JAMIA*, pages ocae037–ocae037.

Peter N Robinson, Sebastian Köhler, Sebastian Bauer, Dominik Seelow, Denise Horn, and Stefan Mundlos. 2008. The human phenotype ontology: a tool for annotating and analyzing human hereditary disease. *The American Journal of Human Genetics*, 83(5):610–615.

Murray Shanahan, Kyle McDonell, and Laria Reynolds. 2023. Role play with large language models. *Nature*, 623(7987):493–498.

Kent A Shefchek, Nomi L Harris, Michael Gargano, Nicolas Matentzoglu, Deepak Unni, Matthew Brush, Daniel Keith, Tom Conlin, Nicole Vasilevsky, Xingmin Aaron Zhang, et al. 2020. The monarch initiative in 2019: an integrative data and analytic platform connecting phenotypes to genotypes across species. *Nucleic acids research*, 48(D1):D704–D715.

Maria Taboada, Hadriana Rodríguez, Diego Martínez, María Pardo, and María Jesús Sobrido. 2014. Automated semantic annotation of rare disease cases: a case study. *Database*, 2014:bau045.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Zhizheng Wang, Yuanyuan Sun, Xuyang Hu, Jiafeng Zhao, Zhihao Yang, and Hongfei Lin. 2021. A semantic network encoder for associated fact prediction. *IEEE Transactions on Knowledge and Data Engineering*, 34(11):5114–5125.

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023a. Document-level machine translation with large language models. *arXiv preprint arXiv:2304.02210*.

Xiao Wang, Weikang Zhou, Can Zu, Han Xia, Tianze Chen, Yuansen Zhang, Rui Zheng, Junjie Ye, Qi Zhang, Tao Gui, et al. 2023b. Instructuie: multi-task instruction tuning for unified information extraction. *arXiv preprint arXiv:2304.08085*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Wenxuan Zhou, Sheng Zhang, Yu Gu, Muhao Chen, and Hoifung Poon. 2023. Universalner: Targeted distillation from large language models for open named entity recognition. *arXiv preprint arXiv:2308.03279*.

6 附录

A.同义词生成Prompt构建

本文通过使用先进的大语言模型进行数据增强来缓解同义词信息不足的问题。通过设计详细的Prompt激发大语言模型的理解能力来生成同义词。本文在实验过程中发现HPO术语中有一部分术语具备明确的定义，而另一部分术语并未包含明确的定义。因此，在构建Prompt的过程中，本文分别针对这两种情况进行了处理。对于存在术语定义的概念，本文首先设定大语言模型扮演医学专家的角色，然后对定义中的症状、病因等因素进行分析，结合HPO的层次结构信息以及同义词与原始术语的语义一致性，通过示例构建同义词。而对于不包含明确定义的概念，则无需进行定义的分析步骤。具体Prompt如表8。

	Prompt
有定义概念	As a medical expert, you will be given the name of a human phenotype term along with its definition and subclassing relationships. Your task is to generate synonyms for the phenotype by following these steps: 1. Analyze the symptoms, causes and other significant factors in the definition of the phenotype; 2. Differentiate the term from its parent term by utilizing subclass relationships; 3. Ensure that the synonyms maintain the contextual meaning of the original name; 4. Provide the top three synonyms, ranked from high to low based on their similarity to the term. No need to include detailed analysis steps in the response. ## Example: Input: Name: "Multicystic kidney dysplasia" Definition: "Multicystic dysplasia of the kidney is characterized by multiple cysts of varying size in the kidney and the absence of a normal pelvicaliceal system. The condition is associated with ureteral or ureteropelvic atresia, and the affected kidney is nonfunctional." Subclass relationship: "Multicystic kidney dysplasia is a subclass of Renal cyst." Response: 1. "Multicystic renal dysplasia" 2. "Multicystic dysplastic kidney" 3. "Multicystic kidneys" ##
无定义概念	As a medical expert, you will be given the name of a human phenotype term and its subclassing relationships. Your task is to generate synonyms for the phenotype by following these steps: 1. Differentiate the term from its parent term by utilizing subclass relationships; 2. Ensure that the synonyms maintain the contextual meaning of the original name; 3. Provide the top three synonyms, ranked from high to low based on their similarity to the term. No need to include detailed analysis steps in the response. ## Example: Input: Name: "Multicystic kidney dysplasia" Subclass relationship: "Multicystic kidney dysplasia is a subclass of Renal cyst." Response: 1. "Multicystic renal dysplasia" 2. "Multicystic dysplastic kidney" 3. "Multicystic kidneys" ##

Table 8: 同义词生成Prompt构建

第二十三届中国计算语言学大会论文集，第553页-第567页，太原，中国，2024年7月25日至28日。
Proceedings of the 23rd China National Conference on Computational Linguistics, pages 553-567, Taiyuan, China, July 25-28, 2024.

B.同义词数量在GSC+数据集上对模型的性能影响

本文使用GPT-4对不包含同义词信息的术语产生3个同义词，在对同义词的使用数量进行探索的过程中分别尝试添加1个，2个，以及3个同义词，基于三种基座模型BioBERT、Bioformer和PubMeBERT在GSC+数据集上进行性能对比。实验结果如图6所示。

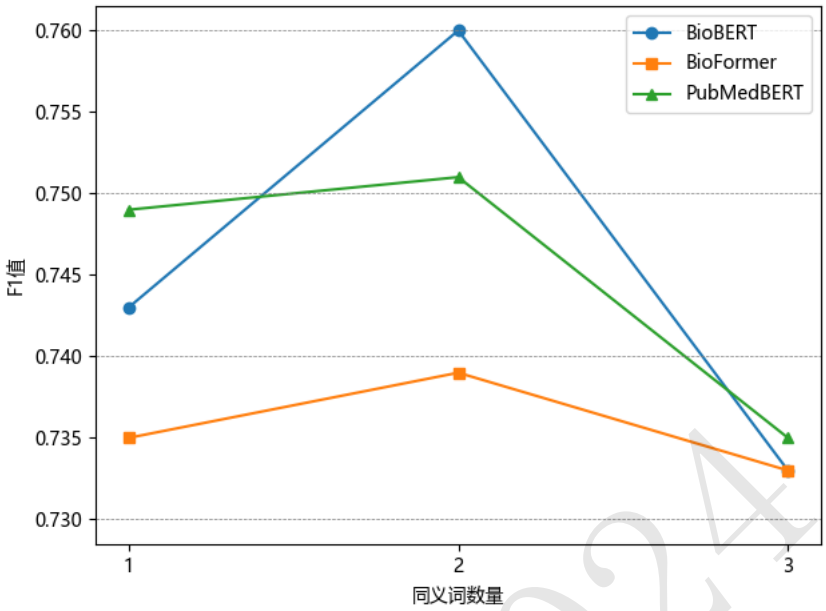

Figure 6: 同义词数量在GSC+数据集上对模型的性能影响

实验结果表明：在三种基座模型上，均为添加2个同义词时获得了最佳的性能。其原因可能是GPT-4生成2个同义词最为准确，因此添加2个同义词可以学习到最丰富的语义信息，而添加1个和3个同义词时出现了信息不足和干扰信息过多的情况。因此后续实验均选择使用2个同义词。

C.本文方法详细结果

本文方法可以在NER和NEN子任务上分别进行评价，本文最后使用的GPT+EMB融合后模型在GSC+和ID-68数据集两个数据集上的NER和NEN任务的效果如表9所示。

子任务	GSC+数据集			ID-68数据集		
	Precision	Recall	F1	Precision	Recall	F1
NER	0.941	0.848	0.880	0.964	0.837	0.888
NEN	0.783	0.744	0.763	0.924	0.781	0.846

Table 9: 本文方法详细结果

第二十三届中国计算语言学大会论文集，第553页-第567页，太原，中国，2024年7月25日至28日。
卷1：主会论文