计算机科学
COMPUTER SCIENCE

基于提示学习的生物医学关系抽取方法

文坤建, 陈艳平, 黄瑞章, 秦永彬

引用本文

文坤建, 陈艳平, 黄瑞章, 秦永彬. 基于提示学习的生物医学关系抽取方法[J]. 计算机科学, 2023, 50(10): 223-229.

WEN Kunjian, CHEN Yanping, HUANG Ruizhang, QIN Yongbin. Biomedical Relationship Extraction Method Based on Prompt Learning [J]. Computer Science, 2023, 50(10): 223-229.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.220900108

基于提示学习的生物医学关系抽取方法

文坤建　陈艳平　黄瑞章　秦永彬
贵州大学公共大数据国家重点实验室　贵阳 550025
贵州大学计算机科学与技术学院　贵阳 550025
(gs.kjwen21@gzu.edu.cn)

摘　要　在非结构化生物医学文本数据中提取出实体之间的关系,对生物医学的信息化发展有着重大意义,同时也是自然语言处理领域的研究热点。目前,在生物医学数据中正确地提取出实体间的关系面临着两个难点:1)由于在生物医学数据中实体单词大多由复合词、未知词组成,模型难以学习到实体内部的语义特征;2)由于生物医学带标注数据较少,而神经网络的参数量较大,使得神经网络容易过拟合。因此,文中提出了基于提示学习的生物医学关系抽取方法,增加了一种针对实体的注解标签,来对实体进行提示以达到实体语义增强以及联系上下文信息的目的。此外,在传统提示调优方法的基础上,文中使用连续性模板来缓解人工设计模板所带来的性能偏差,同时结合深度前缀控制 attention 的深度提示能力,使模型在处理较少数据的情况时仍能取得良好的效果。

关键词:关系抽取;生物信息抽取;提示调优
中图法分类号　TP391

Biomedical Relationship Extraction Method Based on Prompt Learning

WEN Kunjian,CHEN Yanping,HUANG Ruizhang and QIN Yongbin
State Key Laboratory of Public Big Data,Guizhou University,Guiyang 550025,China
College of Computer Science and Technology,Guizhou University,Guiyang 550025,China

Abstract　Extracting the relationship between entities from unstructured biomedical text data is of great significance for the development of biomedical informatization. At the same time,it is also a research hotspot in the field of natural language processing. At present,there are two difficulties in correctly extracting the relationship between entities in biomedical data. One is that in biomedicine,entity words are mostly composed of compound words and unknown words,which makes it difficult for the model to learn the semantic features inside the entity. Second,because there are few biomedical band labeling data and the amount of parameters of neural network is large,the neural network is prone to overfitting. Therefore,a biomedical relationship extraction method based on prompt learning is proposed in this paper. In this paper,an annotation label for entities is added to prompt entities to enhance entity semantics and contact context information. In addition,based on the traditional prompt optimization method,this paper uses the continuity template to alleviate the performance deviation caused by the manual design of the template. At the same time,combined with the depth prefix to control the depth prompt ability of attention,the model can still achieve good results when dealing with a small amount of data.

Keywords　Relation extraction,Biological information extraction,Prompt-tuning

1　引言

蛋白质之间的相互作用(Protein-protein Interaction,PPI)是生物医学实体间非常重要的关系,其中涉及了细胞生长、代谢途径等重要信息。识别蛋白质之间的相互作用具有广泛的科研价值和经济效益,可用于支持药物设计和疾病机制研究等任务。近年来,随着生物医学领域信息化的推进,相关的资料、文献、数据等数字化文本信息呈现出指数级增长的趋势[1]。生物医学文献中蕴含着丰富的、前沿的生物医学知识,是相关从业人员重要的知识来源。但仅依靠人工的方式从中提取信息费时费力。因此,自动从非结构化文本中提取蛋白质关系,对生物医学信息化发展有着重大意义,同时也是自然语言处理(Natural Language Processing,NLP)领域的研究热点。

到稿日期:2022-09-12　返修日期:2022-12-07
基金项目:国家自然科学基金(62166007)
This work was supported by the National Natural Science Foundation of China(62166007).
通信作者:陈艳平(ypench@gmail.com)

PPI关系抽取是根据生物医学文献中的句子内容，对两个蛋白质实体之间的相互作用关系进行分类。PPI关系可以定义为形如〈实体1-关系-实体2〉的三元组。在如图1所示的句子中，存在三对相互作用关系：〈ykuD-True-SigK〉，〈ykuD-False-T4〉，〈SigK-True-T4〉。关系为True时表示实体间存在相互作用，反之则不存在。

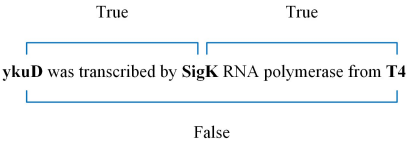


图1 生物关系抽取三元组

Fig. 1 Biological relation extraction triples

传统提取PPI的方法有基于规则的方法、基于特征工程的方法和基于核函数的方法。虽然这些方法取得了一定的效果，但都非常依赖于人工的方式去设计对应的规则以及特征，使得方法的通用性较低。近年来，基于深度学习的方法[2]成为了关系抽取任务的主流，例如卷积神经网络(Convolutional Neural Network，CNN)[3]和循环神经网络(Recurrent Neural Network，RNN)[4]。该方法能在没有人工设计的特征和规则的情况下，自动地从训练数据中学习到单词的连续向量(即分布式表示)，并将其作为任务的特定特征。尽管基于深度学习的方法取得了极大的进展，但许多研究发现，由于现有生物关系抽取中的训练标注数据较少，很难为神经网络中众多的参数提供足够的训练样本，因此容易产生过拟合的问题[5]。

目前，将下游任务重新表示为与预训练语言模型(Pre-trained Language Model，PLM)源任务相近的提示调优方法(Prompt-tuning，PT[6])成为了关系抽取任务中新的范式。例如，在情感分析问题上，使用PT分类句子“I like the Disney films very much”时，可以构建一个模板“It was [*MASK*]”来将该任务形式化为PLM的预训练源任务。通过PLM的Masked Language Modeling(MLM)任务[7]，可以根据上下文信息得出[*MASK*]的预测词。然后，使用预测词对句子分类，如“great”为积极，“bad”为消极。由于PT使用的是PLM的预训练源任务，引入的额外参数很少，因此即使在训练量较小的情况下仍能取得良好的效果。但目前传统的PT大多采用人工的方式构建模板，模板的优劣又直接影响模型的表现，使得模型性能浮动较大。

针对现有模型的不足，本文提出了融合连续性模板与深度提示前缀的提示调优模型(Continuity Template and Depth Prompt prefix Model，CTDP)。在传统PT的基础上，使用具有连续语义且可训练的单词替换模板中的离散单词，经过训练后能在连续空间中自动搜寻得出最优的状态，以此来缓解人工设计方式带来的性能偏差。在此基础上，结合深度提示前缀的注意力增强能力，使模型即使在数据量很小的情况下仍能取得不错的效果。此外，本文观察到生物医学数据中存在许多由复合词或未知词组成的实体，这将严重削弱模型对句子语义的理解，因此本文提出了一种注解式标签，通过对实体进行实体特点的解释，使得模型更好地理解实体语义，同时将标签也纳入模板的构造中，以起到联系上下文的作用。本文的贡献如下：

1)针对生物医学数据中实体大多由复合词、未知词组成的问题，本文提出了注解式标签来对实体进行实体特点或类型的提示，以达到缓解该问题的作用。

2)鉴于人工设计模板过程繁琐，本文使用连续性模板来降低其设计难度，并与深度提示前缀相结合，提出了在数据量较少的情况下仍能取得良好效果的CTDP模型。

3)将本文的CTDP模型在IEPA，HPRD50和LLL数据集上进行验证，并取得了目前最好的效果。

2 相关工作

传统的机器学习大多采用浅层架构的方式来进行关系抽取，如基于规则的方法[8-9]、基于特征的方法[10]以及基于核函数的方法[11]。但这些方法存在需要领域专家去设计具体的规则、特征或核函数的问题，在不同领域之间的通用性较差。目前主流的关系抽取方法则是使用各种深层网络堆叠的方式，来自动地从原始文本中提取高阶抽象表示，例如CNN[3]RNN[4]、LSTM[12]和注意机制[13]。

为了更好地学习句子表示，使用自监督学习从原始文本中学习知识的PLM得以提出，如ELMo[14]和BERT[7]等。PLM被广泛用于将句子嵌入到分布式表示中，而后再堆叠其他结构来进行微调PLM。Li等[15]提出了一种多粒度语义融合方法来进行生物医学关系提取，使用PLM将句子编码为词嵌入表示，可以有效地对句子的全局语义进行编码。同时，采用多通道策略对单词的局部语义进行编码，使同一个单词在句子中具有不同的表示形式。

近年来，PT受到了相当大的关注，并取得了重大的成功。通过使用语言提示作为上下文，PT可以更好地激发PLM的潜能。PT中模板构造的优劣对PT的性能有着很大的影响。Han等[16]提出了基于规则的方式设计模板，将模板中的提示内容拆分为几个子提示，构建好子提示后再将它们组合成为整体，使模板的构建更具效率。Gao等[17]采用seq-to-seq模型生成提示的候选模板，然后使用每个模板进行PT，并验证它们在开发集上的有效性。Liu等[18]提出了使用前缀提示的PT，通过放弃传统PT中的模板，使得模型能够支持序列标注这样的任务。Liu等[19]提出在模板中使用可训练的连续提示，通过将模板纳入模型参数中，使其可以随着模型训练，最终在连续空间中搜寻得出最优的提示结果。

3 融合连续性模板与深度提示前缀的提示调优方法

传统的PT中，构造模板时采用离散的单词来构造一个有上下文关系的文字提示句子，如图2中的Typical Prompt所示。本文方法则使用具有连续性语义且可训练的单词替换模板中固定的词，同时在句子与模板中增加注解式标签来增强句子中的实体语义以及联系上下文信息，如图2中的Our Method所示。

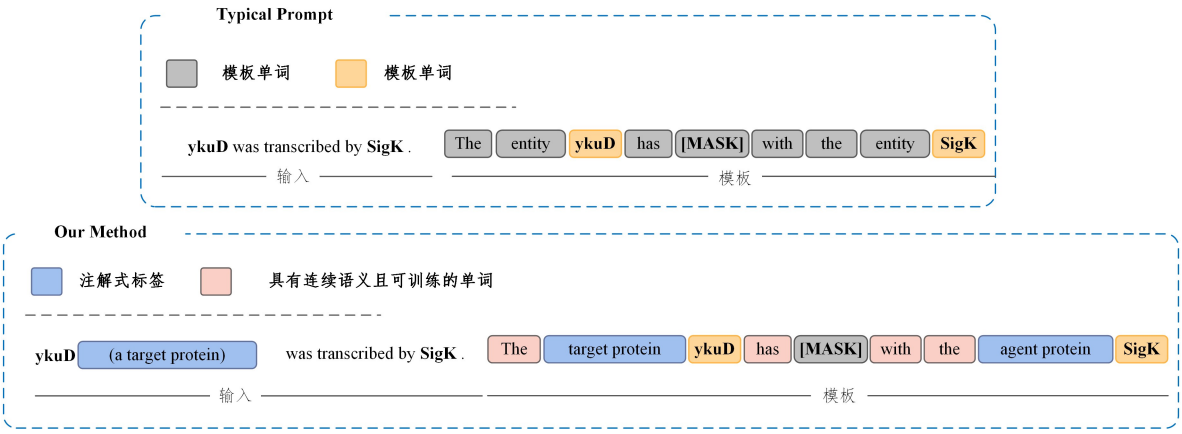


图 2 典型的 prompt 与本文方法(电子版为彩图)

Fig. 2 Typical prompt and our method

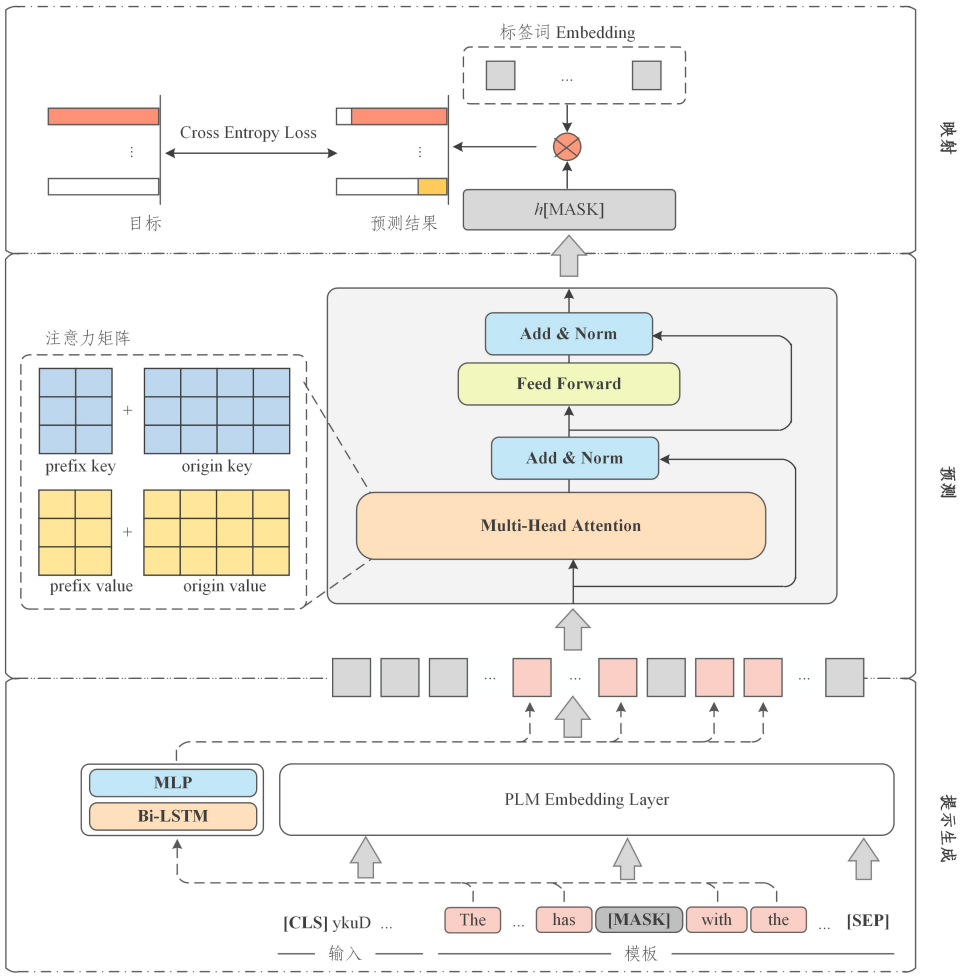


图 3 CTDP 模型结构

Fig. 3 CTDP model structure

本文将模型分为 3 个部分：提示生成、预测和映射。首先在提示生成阶段，根据原始输入的句子生成模板，并和句子拼接组合成整体，再经过处理后得到词嵌入表示；然后在预测阶段，经过注意力运算获得句子的语义，得出模板中[*MASK*]位置的隐藏层向量；最后在映射阶段，将隐藏层向量映射为每种关系分类的概率，得出模型的预测结果，如图 3 所示。

3.1 提示生成

本文将提示生成分为两部分内容：注解式标签的生成与连续性模板的生成。

3.1.1 注解式标签

针对生物医学数据集的特殊性，本文提出了针对实体的注解式标签，如图 2 中的蓝色部分所示。通过该标签来对实体进行实体特点的提示，以实现实体语义的强化。同时在构建模板时，也将注解式标签融入模板之中，以起到联系上下文的作用。

本文在构造注解式标签时采用了两种策略，一种是根据实体类型来构建标签，另一种是根据实体位置来构建标签。例如，当存在实体类型为“target”时，构建出标签信息为

"(a target protein)"。而当不存在实体类型时,根据实体位置将实体分为头实体与尾实体,例如将头实体的标签构建为"(a head protein)"。各数据集采用注解式标签构造的情况如表1所列。

表1 各数据集采用注解式标签构造

Table 1 Each data set is constructed with annotated labels

dataset	entity type\position	tag content
LLL	target	(a target protein)
	agent	(a agent protein)
	agent/target	(a agent or targetprotein)
IEPA	first entity	(a head protein)
	second entity	(a tail protein)
HPRD50	first entity	(a head protein)
	second entity	(a tail protein)

3.1.2 连续性模板

本文使用两种 Embedding 操作来将句子编码到向量表示,模板中具有连续性语义且可训练的单词(如图2中的粉色部分)使用的自定义 Embedding 操作记为$Embed_1(\cdot)$,而其他单词使用的 PLM Embedding 层记为$Embed_2(\cdot)$。

对连续性模板的实施并没有深入到 PLM 结构中做出修改。首先使用人工的方式确定出一套初始模板后,将初始模板经过 PLM Embedding 层编码为词嵌入表示,最后将初始模板的词嵌入作为自定义 Embedding 层的初始状态。

对由 N 个单词组成的句子 $\boldsymbol{x}=[x_i]_{1\leqslant i\leqslant N}$,定义实体1为 $\boldsymbol{s}=[x_{s_{\text{start}}},\cdots,x_{s_{\text{end}}}]$,实体2为 $\boldsymbol{o}=[x_{o_{\text{start}}},\cdots,x_{o_{\text{end}}}]$,如图2输入中的加粗部分所示。$\boldsymbol{T}_s$,$\boldsymbol{T}_o$ 代表对应实体1与实体2的注解式标签,如图2中的蓝色部分所示。则模板构建为:

$$f(\boldsymbol{s},\boldsymbol{o})=[c_1\cdots\boldsymbol{T}_s,\boldsymbol{s}\cdots[MASK]\cdots\boldsymbol{T}_o,\boldsymbol{o}\cdots\boldsymbol{c}_{|\mathbb{C}|}] \tag{1}$$

其中,省略部分由 $\boldsymbol{c}\in\mathbb{C}$ 组成,而 $\mathbb{C}$ 为模板中具有连续性语义且可训练的单词集合,如图2中的粉色部分所示。可以得到输入到模型时的样本 $\boldsymbol{x}_{\text{prompt}}=[\boldsymbol{x},f(\boldsymbol{s},\boldsymbol{o})]$。将样本 $\boldsymbol{x}_{\text{prompt}}$ 编码为词嵌入表示的过程可表述为如下公式:

$$h(\boldsymbol{x}_{\text{prompt}})=[h(\boldsymbol{x}),h(f(\boldsymbol{s},\boldsymbol{o}))] \tag{2}$$

$$h(\boldsymbol{x})=Embed_2(\boldsymbol{x}) \tag{3}$$

其中,$h(\cdot)$表示对句子逐词进行词嵌入编码。对 $f(\boldsymbol{s},\boldsymbol{o})$进行词嵌入编码时,先将 $\mathbb{C}$ 经过自定义 Embedding 层编码为词嵌入表示,再通过 Bi-LSTM 来增加单词之间的连续性,之后使用 MLP 来做非线性映射,最后使用 $\mathbb{C}$ 的连续向量来表示 $f(\boldsymbol{s},\boldsymbol{o})$中属于 $\mathbb{C}$ 的单词,不属于的则使用$Embed_2(\cdot)$对其编码。设 $f(\boldsymbol{s},\boldsymbol{o})$中的单词用 $\boldsymbol{v}$ 表示,该编码过程的表达式如下:

$$h(f(\boldsymbol{s},\boldsymbol{o}))=\begin{cases}Embed_1(v), & v\in\mathbb{C}\\ Embed_2(v), & v\notin\mathbb{C}\end{cases} \tag{4}$$

3.2 预测

对每个 Multi-Head Attention[13] 的注意力矩阵进行扩展,具体为通过在 attention 机制中的 ***key*** 和 ***value*** 矩阵前拼接上可训练的前缀矩阵,来影响模型的注意力方向,以达到提示效果。

前缀矩阵初始是随机产生的,原因在于深度提示前缀需要嵌入到 Transfor-mers[13] 模型的每一层中,而 Transformers[13] 模型的每一层所起到的作用大不相同,对深度提示前缀进行初始化的方案太过复杂。深度提示前缀的方案虽然采用随机初始化的方式进行,但是在经过简单的训练之后,仍能学习到对 Transformers[13] 的每一层都有益的前缀提示。其计算过程如下:

$$\boldsymbol{K}=[\boldsymbol{K}_{\text{prefix}};K_{\text{origin}}] \tag{5}$$

$$\boldsymbol{V}=[\boldsymbol{V}_{\text{prefix}};V_{\text{origin}}] \tag{6}$$

$$Attention(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{\mathrm{d_K}}}\right)\boldsymbol{V} \tag{7}$$

其中,[;]表示矩阵拼接,$\boldsymbol{K}_{\text{prefix}}$,$\boldsymbol{V}_{\text{prefix}}$ 为前缀矩阵,$\boldsymbol{Q}$,$\boldsymbol{K}_{\text{origin}}$,$\boldsymbol{V}_{\text{origin}}$ 为 attention 原始矩阵。

3.3 映射

每一个分类对应存在一个标签词,设关系类别集为 Y,则对应的标签词集为 Y_w。对应样本 $\boldsymbol{x}_{\text{prompt}}$ 的关系类别为 y,对应的标签词为 y_w。在样本 $\boldsymbol{x}_{\text{prompt}}$ 输入模型经过运算后,会得到 $f(\boldsymbol{s},\boldsymbol{o})$中 $[MASK]$ 处的隐藏层向量 $\boldsymbol{h}_{[MASK]}$,通过计算 $\boldsymbol{h}_{[MASK]}$ 与标签词集中词的词嵌入向量的相似度得出对应分类的概率,其过程为:

$$P(y|\boldsymbol{x}_{\text{prompt}})=P([MASK]=y_w|\boldsymbol{x}_{\text{prompt}}) \tag{8}$$

$$P(y|\boldsymbol{x}_{\text{prompt}})=\text{softmax}(\boldsymbol{h}_{[MASK]}\cdot Embed_2(Y_w)^{\mathrm{T}}) \tag{9}$$

因此,可得出优化目标为:

$$loss=CE(P(y|\boldsymbol{x}_{\text{prompt}}),y) \tag{10}$$

4 结果与分析

4.1 数据集与评估

本文使用 PPI 语料库中的 IEPA[20], HPRD50[21] 和 LLL[22] 数据集来测试模型性能。由于数据集中数据的数量较少,且数据集中不存在规定的数据划分方式,因此本文采用 k 折交叉验证的方式来进行测试。将每一份数据作为测试集,将其他数据作为训练集,在每个数据集的5份数据上轮流选取测试集来进行测试,并取5次结果的平均值作为模型的最终性能。

在评估时,本文采用传统的 Precision/Recall/F1-score (P/R/F)作为性能的评估指标,其中 Precision 和 Recall 的定义如下:

$$Precision=\frac{TP}{TP+FP} \tag{11}$$

$$Recall=\frac{TP}{TP+FN} \tag{12}$$

其中,*Precision* 指输出中有多少阳性样本是正确的,*Recall* 指输出中有多少阳性样本是正确预测的。*F1-score* 是 *Precision* 和 *Recall* 的调和平均值,其计算式如下:

$$F1\text{-}source=\frac{2\times Precsision\times Recall}{Precision+Recall} \tag{13}$$

本文使用 BioBert-v1.1[23] 作为预训练语言模型,使用的超参数有最大句长为150,批次大小为16,训练轮次为50,学习率为 2×10^{-5},并使用 AdamW 优化器进行优化。

4.2 结果与讨论

本文以几种典型的方法为基准来评估 CTDP 模型的性能。DRCNN[24] 是一种基于残差卷积神经网络的方法,通过深化神经网络结构来进行 PPI 提取。tLSTM+tAttn[24] 建立在依赖解析树上。RNN+CNN[25] 结合了递归和卷积神经

网络来提取 PPI。Mul-semantic fusion CNN[15]对句子的全局语义和局部语义进行编码后融合以提取 PPI。R-BERT[26]使用实体信息增强 PLM 的编码信息，来进行关系抽取。PTR[16]基于规则的方式来设计模板的 PT 关系抽取方法。表 2 列出了每个模型在 PPI 语料上的表现，其中加粗显示的数据为最优值。

表 2 本文方法与其他对比方法在 PPI 语料上的性能比较

Table 2 Performance comparison of our method with other comparative methods on PPI corpus

（单位：%）

Models	LLL[22]			HPRD50[21]			IEPA[20]		
	P	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*
DRCNN[24]	80.5	87.2	83.2	74.9	82.8	77.7	71.6	80.6	75.5
RNN+CNN[26]	72.5	87.2	76.50	64.3	65.8	63.4	69.6	82.7	75.1
tLSTM+tAttn[25]	84.8	84.3	84.2	81.7	82.3	81.3	78.6	78.7	78.5
Mul-semantic fusion CNN[15]	91.8	93.4	92.5	78.9	**90.9**	84.5	81.8	80.6	81.2
R-BERT[27]	88.3	92.9	90.5	**85.5**	83.8	84.6	81.1	83.0	81.8
PTR[16]	87.9	96.0	91.6	80.3	82.6	81.1	77.5	85.17	81.0
Our Method	**94.3**	**96.0**	**95.0**	83.8	90.0	**86.2**	**82.3**	**88.5**	**85.1**

在 3 个数据集中，LLL[22]所包含的数据数量是最少的，对于微调预训练语言模型的方法来说，在此数据集上训练的过程是最为困难的，由于数据的数量较少，因此传统的微调方法无法将参数调整到良好的状态，但本文方法在此数据集上仍拥有 95%的良好性能。

在 HPRD50[21]数据集上，与 Mul-semantic fusion CNN 方法相比，本文方法的召回率略低于该方法，但其所得到的精确度更高，因此最终的 F1-source 仍然是本文方法更高。

在 IEPA[20]数据集上，本文方法的 F1-source 提升最为明显，与 Mul-semantic fusion CNN 的方法相比提高了 3.9%。

4.3 消融实验

为了证明 CTDP 模型的每个组成部分的贡献，本文对它们进行了消融研究。根据模型的 3 个组成部分，设定 3 个组成因素：1)注解式标签(label)；2)深度提示前缀(prefix)；3)连续性模板(template)；并逐步去掉某些模块来验证其效果，结果如表 3 所列。

表 3 逐步去掉各模块后性能的变化

Table 3 Performance changes after gradually removing each module

	LLL	HPRD50	IEPA
our model	**95.02**	**85.10**	**86.21**
- prefix	93.30$_{\downarrow 1.72}$	84.20$_{\downarrow 0.9}$	85.39$_{\downarrow 0.82}$
- template	92.51$_{\downarrow 2.51}$	80.22$_{\downarrow 4.88}$	82.45$_{\downarrow 3.76}$
- label	91.56$_{\downarrow 3.46}$	83.79$_{\downarrow 1.31}$	85.16$_{\downarrow 1.05}$
- template - label	82.33$_{\downarrow 12.69}$	81.25$_{\downarrow 3.85}$	80.16$_{\downarrow 6.05}$
- prefix - label	89.37$_{\downarrow 5.65}$	83.57$_{\downarrow 1.53}$	84.33$_{\downarrow 1.88}$

对于去掉连续性模板的情况，由于没有模板便无法进行 MLM 任务，因此在进行这一实验时，本文将映射部分改为与传统微调方法一致，即使用隐藏层向量来做最后的类别分类。为保障实验的公平性，在同一数据集中所使用的参数均是固定的。

当去掉深度提示前缀之后，CTDP 模型在每个数据集上 F1 性能下降的幅度在 3 个影响因素中是最小的，这正表明了深度提示前缀能强化 Transformers[13]的特征提取能力。随着训练数据量的增加，Transformers[13]模型对特征的提取能力也会有所提升，因此去掉深度提示前缀后的负面效果也会逐渐下降。

当去掉连续性模板之后，模型的性能在 HPRD50[21]和 IEPA[20]数据集上的下降幅度更大，在 LLL[22]数据集上的下降幅度小于去掉注解式标签的情况，在 LLL[22]中的情况应与注解式标签的构造方式相关联，但从结果来看，连续性模板在模型中仍然起到了非常重要的作用。

当去掉注解式标签之后，模型的性能在每个数据集上都有所下降，在 LLL[22]数据集上下降程度最大，本文认为其原因在于，在构建每个数据集中的注解式标签时，对于 LLL[22]数据集，采用的是与实体类型相关联的策略，而对于另外两个数据集则采用了与实体位置相关联的策略，因为 LLL[22]数据集中的解释性标签与实体类型相关联，每一条数据之间更有区分度，使模型能更好地学习到句子语义，增强模型的识别能力。

当将连续性模板与注解式标签都去掉而只使用深度提示前缀时，模型的性能在各个数据集上均出现了大幅下降，由此可见模型中每一部分都有着重要的作用。而将深度提示前缀与注解式标签都去掉，只使用连续性模板时，模型性能下降的幅度小于只使用深度提示前缀的情况。

4.4 小样本训练

为验证模型在数据稀少情况下的训练成果，本文在 IEPA[20]，HPRD50[21]，LLL[22]这 3 个数据集上展开了实验。控制测试数据不变，将训练数据分别缩小为 40%，60%，80%来进行实验，实验结果如图 4 所示。

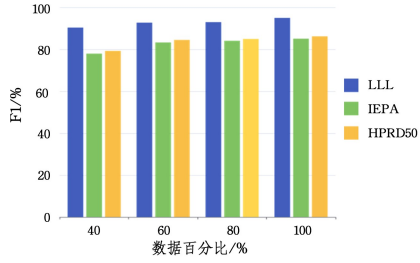


图 4 小样本训练结果

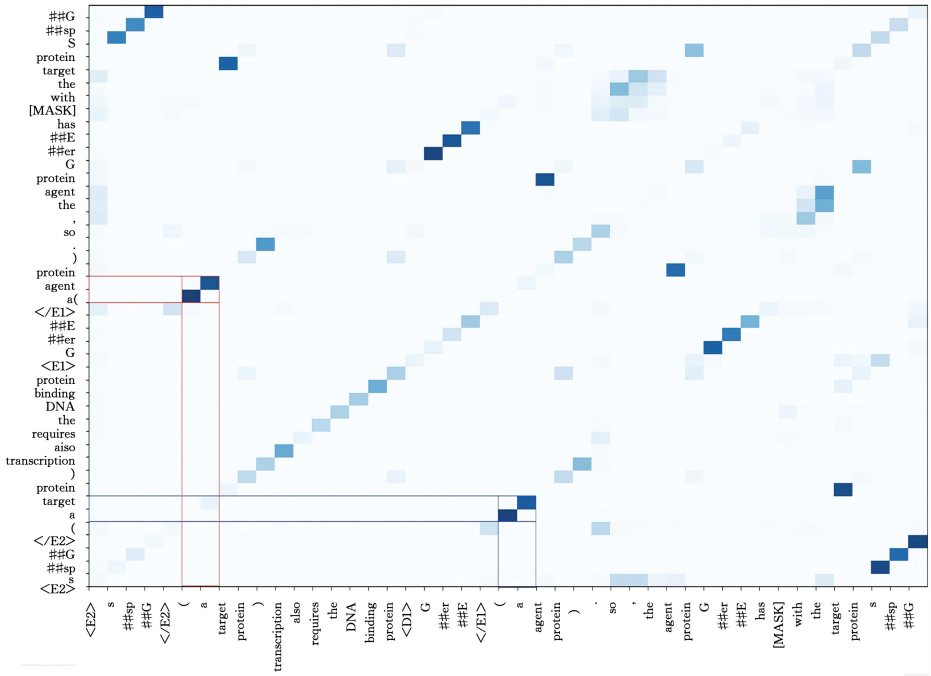
Fig. 4 Small sample training results

在 3 个数据集上均是在取训练数据为 100%时 F1 性能取得最高，随着训练数据的减少，性能开始逐渐下降，但下降趋势较为平稳。即使在只使用 40%的训练数据下，模型仍然能取得较为良好的结果，这印证了 CTDP 模型对数据稀少的

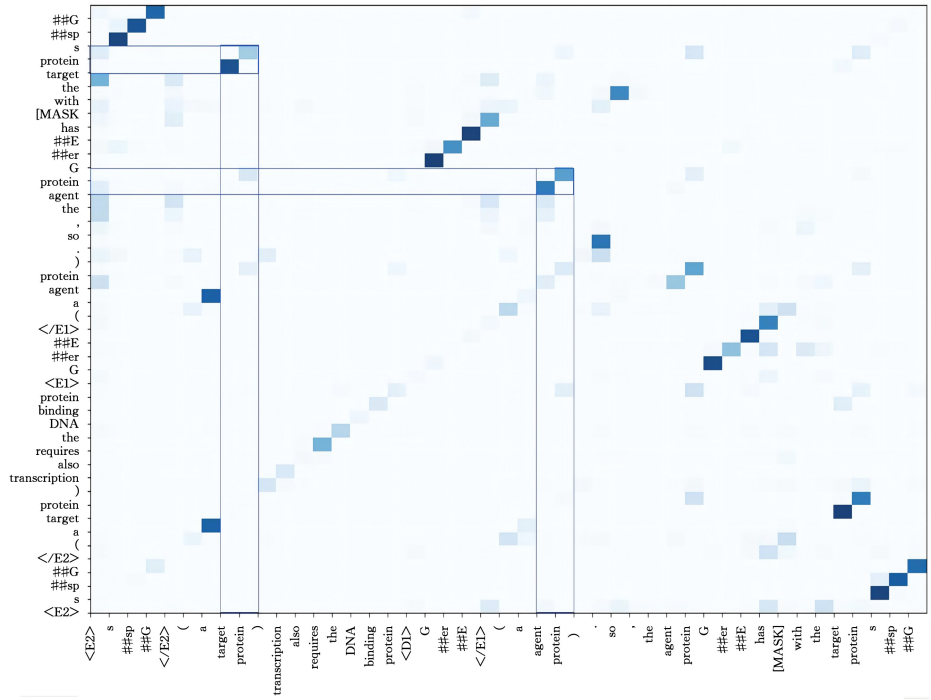
情况有较强的鲁棒性。

4.5 attention 可视化分析

为验证模型中注解式标签的作用，本文对模型训练后的 attention 进行了可视化分析，结果如图 5 所示。在图 5(a)和图 5(b)中，句子中的实体与模板中的实体之间都有很好的相互关注度，这印证了本文模板构造的合理性。在图 5(a)中，在模板中的注解式标签与在句子中和其对应的注解式标签有很好的相互关注度，在图 5(b)中，实体 1 的注解式标签与实体 2 的注解式标签有较高的相互关注度，这印证了本文提出的标签具有联系上下文的能力。



(a)句子中注解式标签相互关注



(b)句子与模板中注解式标签相互关注

图 5 attention 可视化

Fig. 5 attention visualization

结束语 本文将生物关系抽取与提示调优相结合，提出了融合连续性模板与深度提示前缀的提示调优模型。该模型从将下游任务转化为与预训练语言模型训练时一致的方式出发，通过加入连续性模板、深度提示前缀以及作用于句子中实体上的注解式标签，来提高模型提取句子语义的能力。本文在 PPI 的 3 个数据集上展开实验，结果表明，与最新的方法相比，本文方法在性能上有显著的提升，这验证了该方法的可行性。在未来的工作中，我们将探索该方法在通用领域内的

实现效果,以及探索更多模板上的创新。

参 考 文 献

[1] WEXLER P. The US. National Library of Medicine's Toxicology and Environmental Health Information Program[J]. Toxicology,2004,198(1/2/3):161-168.

[2] LECUN Y,BENGIO Y,HINTON G. Deep learning[J]. Nature,2015,521(7553):436-444.

[3] KIM Y. Convolutional neural networks for sentence classification[J]. arXiv:1408.5882,2014.

[4] SUTSKEVER I,VINYALS O,LE Q V. Sequence to sequence learning with neural networks[J/OL]. Advances in Neural Information Processing Systems,2014,27. https://proceedings.neurips.cc/paper/2014/hash/a14ac55a4f27472c5d894ec1c3c743d2-Abstract.html.

[5] BELKIN M,HSU D,MA S,et al. Reconciling modern machine-learning practice and the classical bias-variance trade-off[J]. Proceedings of the National Academy of Sciences,2019,116(32):15849-15854.

[6] SCHICK T,SCHÜTZE H. Exploiting cloze questions for few shot text classification and natural language inference[J]. arXiv:2001.07676,2020.

[7] DEVLIN J,CHANG M W,LEE K,et al. Bert:Pre-training of deep bidirectional transformers for language understanding[J]. arXiv:1810.04805,2018.

[8] BLASCHKE C,ANDRADE M A,OUZOUNIS C A,et al. Automatic extraction of biological information from scientific text:protein-protein interactions[C]// ISMB. 1999,7:60-67.

[9] ONO T,HISHIGAKI H,TANIGAMI A,et al. Automated extraction of information on protein-protein interactions from the-biological literature[J]. Bioinformatics,2001,17(2):155-161.

[10] KAMBHATLA N. Combining lexical,syntactic,and semantic features with maximum entropy models for information extraction[C]// Proceedings of the ACL Interactive Poster and Demonstration Sessions. 2004:178-181.

[11] BUNESCU R C,MOONEY R J. A shortest path dependency kernel for relation extraction[C]// Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing. 2005:724-731.

[12] HOCHREITER S,SCHMIDHUBER J. Long short-term memory[J]. Neural Computation,1997,9(8):1735-1780.

[13] VASWANI A,SHAZEER N,PARMAR N,et al. Attention is all you need[J/OL]. Advances in Neural Information Processing Systems,2017,30. https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

[14] PETERS M E,NEUMANN M,IYYER M,et al. Deep contextualized word representations[J]. arXiv:1802.05365,2018.

[15] LI Y,CHEN Y,QIN Y,et al. Protein-protein interaction relation extraction based on multigranularity semantic fusion[J]. Journal of Biomedical Informatics,2021,123:1532-0464.

[16] HAN X,ZHAO W,DING N,et al. Ptr:Prompt tuning with rules for text classification[J]. arXiv:2105.11259,2021.

[17] GAO T,FISCH A,CHEN D. Making pre-trained language models better few-shot learners[J]. arXiv:2012.15723,2020.

[18] LIU X,JI K,FU Y,et al. P-Tuning v2:Prompt Tuning Can Be Comparable to Fine-tuning Universally Across Scales and Tasks[J]. arXiv:2110.07602,2021.

[19] LIU X,ZHENG Y,DU Z,et al. GPT understands,too [J]. arXiv:2103.10385,2021.

[20] DING J,BERLEANT D,NETTLETON D,et al. Mining MEDLINE:abstracts,sentences,or phrases? [M]// Biocomputing 2002. 2001:326-337.

[21] FUNDEL K,KÜFFNER R,ZIMMER R. RelEx—Relation extraction using dependency parse trees[J]. Bioinformatics,2007,23(3):365-371.

[22] NÉDELLEC C. Learning language in logic-genic interaction extraction challenge[C]// 4. Learning Language in Logic Workshop(LLL05). ACM — Association for Computing Machinery,2005.

[23] LEE J,YOON W,KIM S,et al. BioBERT:a pre-trained biomedical language representation model for biomedical text mining[J]. Bioinformatics,2020,36(4):1234-1240.

[24] ZHANG H,GUAN R,ZHOU F,et al. Deep residual convolutional neural network for protein-protein interaction extraction[J]. IEEE Access,2019,7:89354-89365.

[25] AHMED M,ISLAM J,SAMEE M R,et al. Identifying protein-protein interaction using tree lstm and structured attention[C]//2019 IEEE 13th International Conference on Semantic Computing(ICSC). IEEE,2019:224-231.

[26] ZHANG Y,LIN H,YANG Z,et al. A hybrid model based on neural networks for biomedical relation extraction[J]. Journal of Biomedical Informatics,2018,81:83-92.

WEN Kunjian, born in 1998, postgraduate. His main research interests include biological information extraction and so on.

CHEN Yanping, born in 1980, Ph.D, associate professor. His main research interests include artificial intelligence and natural language processing.

(责任编辑:喻藜)