计算机科学
COMPUTER SCIENCE

面向工艺实体识别的双向神经概率转换器

李瑞婷, 王裴岩, 王立帮, 杨丹清忻

引用本文

李瑞婷, 王裴岩, 王立帮, 杨丹清忻. 面向工艺实体识别的双向神经概率转换器[J]. 计算机科学, 2024, 51(6A): 230700206-8.

LI Ruiting, WANG Peiyan, WANG Libang, YANG Danqingxin. Bidirectional Neural Probabilistic Transducer for Process Text Entity Recognition [J]. Computer Science, 2024, 51(6A): 230700206-8.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230700206

面向工艺实体识别的双向神经概率转换器

李瑞婷　王裴岩　王立帮　杨丹清忻
沈阳航空航天大学计算机学院　沈阳 110136
(1473326715@qq.com)

摘　要　工艺实体识别旨在识别出产品制造中所遵照或是产生的文本中蕴含的零件、材料、属性和属性值等实体。目前,工艺等领域实体识别大多加入词典或正则规则等领域实体先验知识,修正神经网络模型识别结果或是生成预识别特征加入模型中。但上述方法未能实现领域实体识别的先验知识与神经网络模型统一建模,领域知识的加入没有减小模型训练代价,仍需大量标注数据。为解决上述问题,提出了面向工艺实体识别的双向神经概率转换器(Bi-NPT),将工艺实体识别先验知识建模为正则规则,然后将正则规则转化为参数化的概率有限状态转换器,使得模型在训练前带有实体识别的先验知识,同时具有可训练性。通过在标注数据上的训练,模型能够习得正则规则未覆盖实体的识别能力。实验结果表明,提出的 Bi-NPT 在未训练的情况下与正则规则实体识别效果相当,这表明未经过训练的初始模型即携带了实体识别知识。在小样本条件下,Bi-NPT 优于 PER,Template-based BART 和 NNShot 方法;在充足样本条件下,Bi-NPT 优于 BiLSTM 与 TENER 等方法。

关键词:工艺文本;实体识别;正则规则;概率有限状态转换器
中图分类号　TP391

Bidirectional Neural Probabilistic Transducer for Process Text Entity Recognition

LI Ruiting,WANG Peiyan,WANG Libang and YANG Danqingxin
School of Computer Science,Shenyang Aerospace University,Shenyang 110136,China

Abstract　Process text entity recognition aims to recognize entities such as parts,materials,attributes and attribute values from texts generated or associated with the manufacturing process of products. Recently,in most domain-specific entity recognition tasks,such as process domain,prior knowledge in the form of dictionaries or rules is used to adjust neural network model results or generate pre-recognized features to incorporate into the model. However,these methods do not realize the integration of domain entity recognition knowledge and neural network models. Furthermore,the addition of domain knowledge does not reduce the training cost of the model and still need a large amount of labeled data. To address these challenges,this paper proposes a bidirectional neural probabilistic transducer(Bi-NPT) for process text entity recognition. This approach models the domain-specific prior knowledge for process text entity recognition as regular rules,and then converts these rules into a parameterized probabilistic finite state transducer. This method makes the model carry entity recognition prior knowledge before training,while being trainable. The model acquires the ability to recognize entities not covered by the regular rules by training on labeled data. Experimental results demonstrate that the proposed Bi-NPT performs comparably to regular rule-based entity recognition without training,suggesting that the untrained initial model already has possess entity recognition knowledge. Additionally,Bi-NPT outperforms other methods such as PER,Template-based BART,NNShot in few-shot and BiLSTM,TENER in rich-resource scenarios.

Keywords　Process text,Entity recognition,Regular rules,Probabilistic finite state transducer

1　引言

工艺实体识别是对产品制造中所遵照或是产生的工艺标准、工艺大纲、工艺规范、指导书等文本中蕴含的零件、属性、属性值、工艺辅料和孔结构等进行识别[1-2]。现今,基于深度学习的方法是实体识别的主要方法,如 BiLSTM[3],TENER[4] 及 FLAT[5] 等。此类方法将实体识别建模为序列标注问题,通过训练神经网络模型,从具有一定规模的实体标注语料中学习获得实体构词规律与实体语境规律。然而,专业领域的实体识别往往以该领域的知识为依据,兼顾其语言规律特性,使得其具有一定的难度。因此,如何在神经网络中加入领域实体特有知识是领域实体识别的关键问题。

解决上述问题一般有两种途径。第一种是在深度学习模型识别后,使用词典与规则对识别结果再次修正,如 Ma 等[6] 的化学领域实体识别,Feng 等[7] 的军事领域实体识别以及 Zhu[8] 的维吾尔语实体识别。此类方法虽然能够提升实体识别效果,但是不能修正无规则或是词典未覆盖的实体结果。另一种是首先使用词典或规则对实体进行预识别,将预识别结果作为额外特征加入深度学习模型中,如 Jia 等[2] 的工艺文本命名实体识别。此类方法通过预识别特征加入知识对模型训练的指导,不仅能够提升规则与词典覆盖的实体的识别效果,而且没有规则与词典的实体效果也能得到提升。

基金项目:辽宁省应用基础研究计划(2022JH2/101300248);全国科技名词审定委员会科研项目(YB2022015);国家自然科学基金(U1908216)
This work was supported by the Applied Basic Research Program of Liaoning Province(2022JH2/101300248),Research Programs of China National Committee for Terminology in Science and Technology(YB2022015) and National Natural Science Foundation of China(U1908216).
通信作者:王裴岩(wangpy@sau.edu.cn)

然而，上述两种方法中，神经网络模型与领域知识相互独立建模，规则与词典等先验性领域知识的加入没能降低神经模型对训练数据量的要求，训练神经模型的标注语料规模依然很大。与通用领域识别人名、地名与机构名3类实体相比，工艺文本实体类粒度更细(16类)。在标注同样数量语句的情况下，平均每类实体实例量更少。也就是说，若工艺文本与通用领域文本达到相同的实例量，则需要标注更多的语料。工艺领域文本的强专业性，使得语料标注更为困难。

为解决上述问题，本文提出了一种双向神经概率转换器(Bi-NPT)，用于工艺领域实体识别任务。通过正则规则建模领域实体识别先验知识，将正则规则转化为有限状态转换器并从中提取模型参数，使模型在训练前带有实体识别的先验知识。并且，所提出的Bi-NPT具有可训练性，通过在标注数据上的训练，能够获得对正则规则未覆盖实体的识别能力。具体地，首先使用正则规则定义实体字符级特征与实体标签的映射关系，得到正则规则集；再将正则规则集转化为有限状态转换器，建模实体字符级特征之间的状态跳转逻辑。从有限状态转换器中构建Bi-NPT参数，包括起始状态参数向量、终止状态参数向量和两个状态转移参数张量。为了避免正则规则之间的标签发生冲突，模型设置了标签优先级层，由上述参数矩阵计算并更新前后向隐状态，再经过正则规则优先级层得到实体标签的得分矩阵，最后解码得到每个输入字符对应的实体标签。

通过实验，未训练的Bi-NPT与正则规则实体识别效果相当，这表明Bi-NPT在构建模型参数时能有效建模实体识别的先验知识。在小样本条件下，Bi-NPT优于当前表现较优的PER[2]，Template-based BART[9]与NNShot[10]方法；在充足样本条件下，Bi-NPT优于BiLSTM[3]与TENER[4]等方法。此外，通过改变正则规则规模可以发现，正则规则的规模会通过模型参数量影响其学习能力，正则规则越多，Bi-NPT的识别效果越好。

2 相关工作

现有的专业领域实体识别模型大多数仍基于深度学习方法，该方法通过大规模的高质量数据训练得到文本的向量表示，与人工构建特征或规则的方法相比包含更多的语义信息。由于专业领域标注数据不足，不足以完全支撑深度学习模型的训练，因此有相关工作使用数据增强方法弥补专业领域小样本训练数据的匮乏。Liu等[11]针对农业领域提出了一种融合规则的深度学习模型WPD-RA，该模型采用轻量级动态词向量模型ALBERT与BiLSTM-CRF模型相结合的策略，针对某些实体类别数据较少的问题，提出数据增强方法，通过相似词替换来补充句子语义，从而有效提高在小样本的情况下农业领域命名实体识别的效果。

此外，为降低深度模型对标注数据的需求，研究人员还关注了将额外表征作为模型输入特征的方法，该方法同样能够提升实体识别模型的性能。Cui等[12]面向中文电子病历文本，使用卷积神经网络提取汉字图像特征并与五笔字型编码进行融合作为高级语义信息，指导FLAT模型训练与预测。Zhang等[13]面向教育领域，使用字、词和位置信息指导BiGRU-CRF训练，使模型更好地界定实体边界。Liu等[14]针对军事领域标注文本不足的问题，结合实体识别技术，提出了BERT-BiLSTM-CRF的模型，该模型以字、字位置、语义块及词性作为模型输入特征，通过BERT网络迁移学习，获得通用领域语义编码特征，再利用BiLSTM解码军事语义特征，最后通过CRF实现序列预测，并通过实验证明了方法的有效性。Jia等[2]提出了一种融入工艺领域知识的神经网络命名实体识别方法，该方法利用领域词典与规则预识别出部分实体作为实体预识别特征，提出了一种神经网络模型CNN-BiLSTM-CRF，通过CNN网络利用预识别实体整体特征指导字序列标注模型的训练与预测。实验表明该方法不仅能够提高词典及规则覆盖的实体识别效果，还能够提高其他类实体的识别效果，优于对比方法。

虽然深度学习引领了近年来的技术热潮，但由于基于符号主义的规则系统具备良好的可解释性，因此仍然有着稳固的地位。然而，该方法作为不可训练的先验知识，在数据资源丰富的情况下难以达到与神经网络相近的效果。近年来，如何使规则更好地融入神经网络，将规则与神经网络一体化成为了一个重要的研究方向。对于命名实体识别任务，目前没有开展相关研究，但在其他任务上的研究已经开展。

Peng等[15]将加权有限状态自动机与递归神经网络结合，用于文本分类任务，提出了一个循环隐藏状态更新函数，将此函数看作加权有限状态自动机集合的正向计算，从而模拟有理递归过程。Lin等[16]提出了神经有限状态传感器，用于语音识别任务。模型中有限状态传感器的弧权重取决于使用该弧的上下文，每条路径的权值由神经网络训练后给出，因此神经网络可以沿着一条路径捕获状态之间的依赖关系。Rastogi等[17]提出了一个混合架构FST-LSTM，用于词形还原任务。该方法通过在不同的上下文中对相同的FST弧分配不同的权重，并使用LSTM自动提取确定这些权重的特征，将有限状态转导方法与神经网络方法结合，以加权有限状态自动机的形式定义对齐的输出串上的概率分布。上述方法虽然融合了符号规则系统与神经网络，但模型必须在标记数据上训练，且融入的规则系统不能由文本中获取到的规则直接转换。为了解决该问题，Jiang等[18]面向意图检测任务提出了FA-RNN模型，将句式与其意图标签的对应关系构建为规则，进而将规则通过有限自动机转化为神经网络，可用于零样本任务，还可以利用标记数据进行训练，以提高预测精度。Jiang等[19]在上述研究基础上，提出了用于语义槽填充任务的FST-RNN模型，将先验知识定义为“提示词-结果”的一对一映射关系，使用有限状态转换器替换有限自动机，完成正则规则到神经网络的转化，从而适应序列任务。

航空制造领域的实体粒度更细，且构词规律复杂。不同于上述方法，本文面向实体识别任务，用正则规则描述工艺领域实体的字符级构词特征，将正则规则转化为有限状态转换器加以描述字符特征之间的跳转逻辑；再将构词特征及其跳转逻辑转化为神经网络模型参数，约束模型对实体标签的推理过程。

3 双向神经概率转换器

3.1 工艺领域规则实体的正则规则定义方法

为了适应模型的字符级序列输入输出和本文所使用语料

的“BIOE”标注体系，本节使用正则规则定义字符级捕获组与实体标签的映射关系，用连接符“〈:〉”分隔。由于每条正则规则只能描述实体的一种字符级特征，而一类实体可能包含多种字符级特征，且一条标注数据通常包含多类实体，因此在正则规则的编写过程中，除已知实体标签外还引入“OO”作为待定实体标签，以满足数据与正则规则一对多的映射需求。此外，为了使正则规则更加适应工艺领域规则实体的特点，使用特殊符号定义正则规则中需要概括的字符集合，正则规则中含有特殊含义的符号如表 1 所列。

根据实体的构词特征编写正则规则，构词特征具体分为两类：实体编号标准和实体提示词特征。实体编号标准指该实体的编号符合某一类标准定义。例如零件编号实体“NAS0985”，该实体是符合国家标准的一个铆钉编号，虽然一个零件编号只对应一个或一种零件实体，但“NAS”是描述符合该标准的所有编号，因此可以作为该标准下零件编号实体的正则提示词。实体提示词特征指部分非编号实体具有一定的提示词特征，该特征也可作为正则的编写依据。例如属性值实体“−0.072 MPa”，该实体用于描述工艺作业中的真空度属性。虽然工艺作业的各个环境对真空度的要求不固定，但“MPa”作为描述真空度的固定单位，也可作为正则提示词。规则实体及字符级正则规则定义举例如表 2 所列。

表 1　正则规则符号及含义

Table 1　Symbols in regular rules and their meanings

符号	含义
$	通配符，表示任意字符
Λ	数字字符(0-9)
?	匹配前面的子表达式 0 或 1 次
÷	大写字母(A-Z)
+	匹配前面的子表达式 1 或多次
*	匹配前面的子表达式 0 或多次
|	指明两项之间的一个选择

表 2　正则规则举例

Table 2　Examples of regular rules

实体编号标准	实体类型	PART_ID
	实体语境	铆钉 **NAS1321AD10E-24** 的增量变为 0.8mm(1/32in.)
	正则规则	$〈:〉OO* N〈:〉B-PART_ID A〈:〉I-PART_ID S〈:〉I-PART_ID Λ〈:〉I-PART_ID* ÷〈:〉I-PART_ID* Λ〈:〉I-PART_ID* ÷〈:〉I-PART_ID* -〈:〉I-PART_IDΛ〈:〉I-PART_ID* Λ〈:〉E-PART_ID$〈:〉OO*
实体提示词特征	实体类型	ATTRIBUTE_VA
	实体语境	零件抽真空至真空度大于**−0.072 MPa** 后，施加罐压 0.315 MPa±0.035 MPa(45 psi±5 psi)。
	正则规则	$〈:〉OO* -〈:〉B-ATTRIBUTE_VA 0〈:〉I-ATTRIBUTE_VA. 〈:〉I-ATTRIBUTE_VA Λ〈:〉I-ATTRIBUTE_VA* M〈:〉I-ATTRIBUTE_VA P〈:〉I-ATTRIBUTE_VA a〈:〉E-ATTRIBUTE_VA $〈:〉OO*

3.2　有限状态转换器的转化

为每个规则实体编写相应的正则规则后，得到工艺领域规则实体的正则规则集。给定输入序列，可通过正则规则集判定其是否符合每条正则规则的过滤逻辑来得到相应的标签。为了获得实体的字符级特征的状态跳转逻辑从而构建模型所需参数，需要将正则规则转化为有限状态转换器。以“工艺文件”实体为例，其正则规则到有限状态转换器的转化及识别过程如图 1 所示。

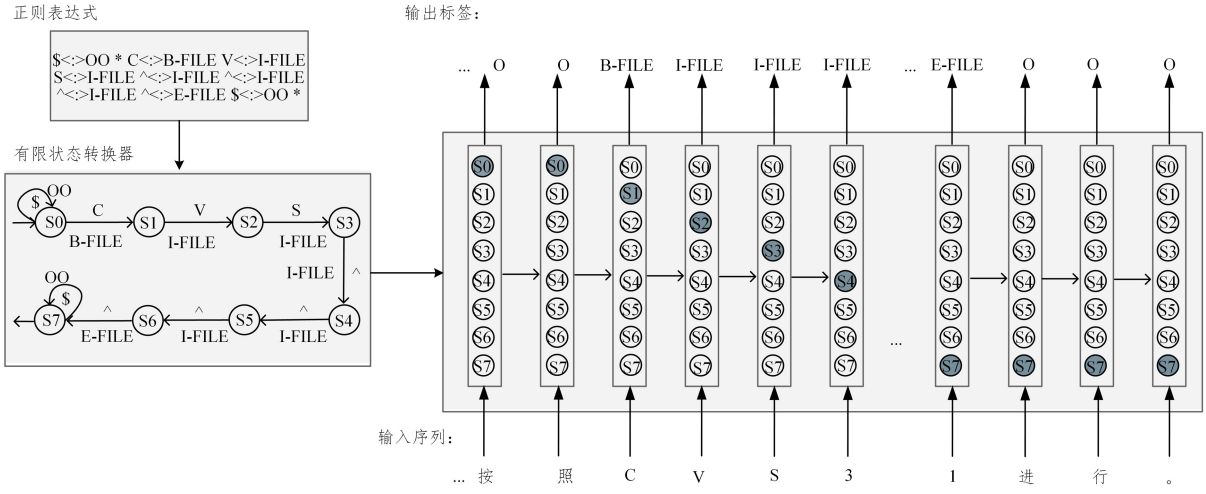

图 1　正则规则转化为有限状态转换器的实体识别示例

Fig. 1　Example of entity recognition for finite-state transducer

接下来详述正则规则转化为有限状态转换器的过程与有限状态转换器的识别过程。

Sakuma 等[20]已经严格证明，一条带有捕获组的正则规则可转化为一个有限状态转换器。有限状态转换器可表示为：

$$A=(Q,\Sigma,\Gamma,\delta,S,F)$$

其中，Q 是非空有限状态集；Σ 是输入状态转换器的非空有限字符集；Γ 是状态转换器的非空输出有限标签集；δ 是转移函数，$|\delta|=|\Sigma|\times|\Gamma|\times|Q|\times|Q|$；$S$ 是起始状态的非空有限集，$S\subseteq Q$；F 是终止状态的非空有限集，$F\subseteq Q$。

首先，将有限状态转换器视为具有词集 $\Sigma\times\Gamma$ 的有限状态自动机，使用 Thompson 算法[21]从正则规则中构建自动机；再使用 Hopsroft 算法[22]得到最小化自动机。将 Σ 作为自动机的输入词集，Γ 作为自动机的输出词集，从而得到有限状态转换器。将每条正则规则转化成有限状态转换器后，Γ 表示标签集合，该集合的元素除用于模型输出的实体标签外，

还包括一个待定标签。由于要保证每条正则规则到实体的字符级特征映射唯一，且不能影响输入序列的其他实体的正确识别，因此待定标签的作用是保证当输入序列完全符合一个转换器的转换条件时，除正则规则捕获组以外的序列子集仍可以匹配其他转换器，并输出正确的实体标签。$\delta(\gamma, x_i)=\upsilon$，$l_j$ 表示有限状态转换器中的状态 γ，在接受来自输入序列的字符 x_i 后，输出标签 l_j，并转移到下一个状态 υ。对于有限状态转换器的输入序列，该转换器从起始状态开始，按序逐一接受输入序列的字符，并根据转移函数依次转移至下一个状态，每次转移均有一个输出；当且仅当到达终止状态，即转换器完全接受一条输入序列时，完成状态转换并释放完整的输出序列。

3.3 双向神经概率转换器

双向神经概率转换器是一种可训练的概率模型，其初始转换概率由正则规则定义。模型训练后，每一步的转换概率由模型通过标注数据学得，且转换条件同时捕获当前时刻上下文的转移特征。该模型将实体标签的求解过程转化为有限状态转换器捕获输入序列完成状态转换条件的概率计算过程。本节将从双向神经概率转换器的参数构建和实体标签求解两部分进行介绍。

3.3.1 双向神经概率转换器的参数构建

给定有限长的输入序列 $X=(x_1, x_2, \cdots, x_p)$ 和输出标签集合 $l=(l_1, l_2, \cdots, l_q)$，可从有限状态转换器 A 中构建双向神经概率转换器的起始状态向量 $\boldsymbol{s}\subseteq\mathbb{R}^m$、终止状态向量 $\boldsymbol{e}\subseteq\mathbb{R}^m$ 和状态转移四阶张量 $\boldsymbol{W}\in\mathbb{R}^{p\times q\times m\times m}$。由于四阶张量在运算时具有较高的时间复杂度和空间复杂度，参考 Jiang 等[18-19]的工作，将一个表示状态转移的四阶张量 $\boldsymbol{W}$ 拆分为两个三阶张量：$\boldsymbol{W}_i\in\mathbb{R}^{p\times m\times m}$ 和 $\boldsymbol{W}_o\in\mathbb{R}^{q\times m\times m}$。其中，$\boldsymbol{W}_i$ 表示输入字符触发状态转换过程，$\boldsymbol{W}_o$ 表示由状态转换并输出实体标签的过程。对于输入字符 x_i、输出标签 l_j、转移前状态 γ 和转移后状态 υ，$\boldsymbol{W}_i$ 和 $\boldsymbol{W}_o$ 满足式(1)。从有限状态转换器中构建 4 个参数张量，如图 2 所示。

$$\boldsymbol{W}[x_i, l_j, \gamma, \upsilon]=\boldsymbol{W}_i[x_i, \gamma, \upsilon]\times\boldsymbol{W}_o[l_j, \gamma, \upsilon] \tag{1}$$

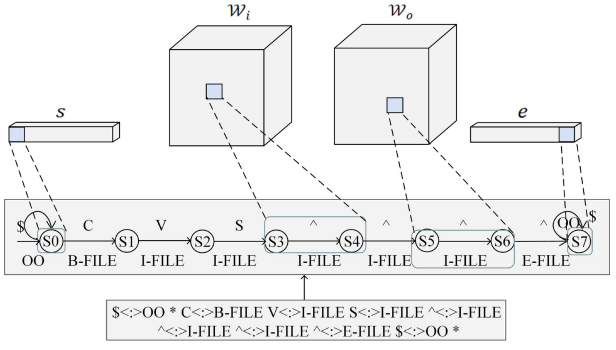

图 2 Bi-NPT 的参数构建过程

Fig. 2 Parameter construction process of Bi-NPT

3.3.2 双向神经概率转换器的实体标签求解

对于给定的输入序列 $X=(x_1, x_2, \cdots, x_p)$，Bi-NPT 需找到输出标签序列 $L=(l_1, l_2, \cdots, l_q)$，其中 L 的得分即 Bi-NPT 中匹配输入 X 和输出 L 的所有接受路径的权重之和。

Francisco 等[23]已经证明，找到给定句子的最高得分输出是 NP 困难的，因此在求解输出标签时，不再考虑其前后的输出得分，使用近似推理的方法，令每个输出标签的得分独立于整条输出序列。

双向神经概率转换器可视作双向循环的序列模型，其初始参数由上述 4 个参数张量定义，训练后参数可表示转换器中的状态转换概率以及转换过程中的输入输出概率，模型对标签的求解过程如图 3 所示。

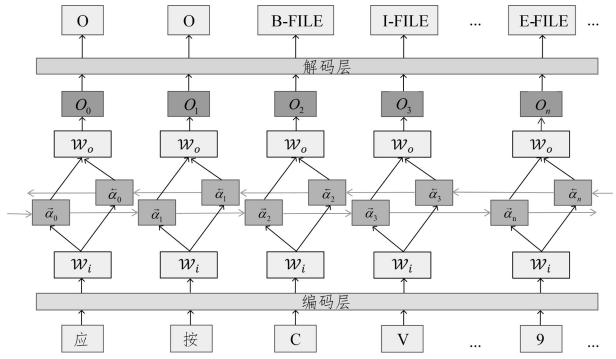

图 3 双向神经概率转换器求解过程

Fig. 3 Bi-NPT solving process

双向神经概率转换器处理序列输入时，该模型按时间步展开，序列中每个时间步的输入对应模型的时间步输入。对于某时间步的输入 $x_t\in\mathbb{R}^d$，其前向传播隐状态更新方式如式(2)－式(5)所示：

$$\vec{\alpha}_0=\boldsymbol{s}^{\mathrm{T}} \tag{2}$$

$$\vec{\alpha}_t=\sum_{m_1}\boldsymbol{W}_i[x_t]_{m_1,m_2}\,\vec{\alpha}_{t-1} \tag{3}$$

$$\vec{\alpha}_t=\underset{m_1\in Q}{\operatorname{argmax}}\,\vec{\alpha}_t(m_1, m_2) \tag{4}$$

$$\vec{\alpha}_t=relu(\vec{\alpha}_t) \tag{5}$$

其中，$\boldsymbol{W}_i[x_t]_{m_1,m_2}$ 表示 x_t 在双向神经概率转换器中的转移前状态，$\vec{\alpha}_t\in\mathbb{R}^d$，$relu$ 是激活函数。

同理，可获得 t 时刻双向神经概率转换器中的反向传播隐状态 $\overleftarrow{\alpha}_t$，其中 $\overleftarrow{\alpha}_0=\boldsymbol{e}^{\mathrm{T}}$。$x_t$ 对应输出标签的得分矩阵求解，可看作是计算 x_t 在双向神经概率转换器中，作为一个状态到达另一个状态的捕获内容的概率。该概率共同考虑 x_t 的前后向隐状态和有限状态转换器的转移特征，计算过程如式(6)所示：

$$\boldsymbol{O}_t=\sum\boldsymbol{W}_o(\sum(\sum\vec{\alpha}_t\overleftarrow{\alpha}_t)\boldsymbol{W}_i[x_t]) \tag{6}$$

由于正则规则的捕获组的子集也可能被正则规则捕获，因此正则规则定义的标签之间可能存在冲突。为了避免上述冲突，本模型在解码层加入优先级修正函数 $priority$，令同一捕获组的输出标签中，“I-”优先于“E-”和“B-”，实体之间的优先级由专家指定。因此 x_t 对应的标签分数概率向量更新如式(7)所示：

$$\boldsymbol{O}_t'=priority(\boldsymbol{W}_p\boldsymbol{O}_t+\boldsymbol{b}_p) \tag{7}$$

其中，$\boldsymbol{W}_p$ 和 $\boldsymbol{b}_p$ 分别是优先级层的权重和偏置。

此时的输出标签中，仍存在“通配符”标签，o_t' 表示 x_t 的不确定性。当所有标签概率全部低于不确定性时，x_t 的对应标签解码为标签“O”，否则忽略 o_t'，直接解码搜索概率最大的标签。对标签得分矩阵解码搜索如式(8)、式(9)所示：

$$\boldsymbol{O}_t''=(\max_{i\in(0,l)}(o_0', o_t'), o_1', o_2', \cdots, o_{l-1}') \tag{8}$$

$$l_t=\underset{0\leqslant i\leqslant l-1}{\operatorname{argmax}}\,\boldsymbol{O}_t''(i) \tag{9}$$

待遍历所有输入序列后，将每个时间步的 x_t 对应得分矩阵按下标计入得分矩阵，使用交叉熵损失函数度量预测值与真实值的差异，如式(10)所示：

$$loss=-\frac{1}{m}\sum_{t=1}^{m}\sum_{j=1}^{l}l_l^{(t)}o_j''^{(t)} \tag{10}$$

其中，m 表示样本数，l_t 表示 t 时刻的真实输出标签，$o_j''^{(t)}$ 表示模型对输入 x_t 对应标签 j 的预测概率。

4 实验分析

本章详细描述了实验数据集和正则规则的构建过程并进行了实验统计分析，通过实验得到如下结论：1）零样本下，Bi-NPT 能够有效建模先验知识，部分实体的识别效果优于正则匹配；2）Bi-NPT 能够使用标注数据训练，在少样本与充足样本条件下具有相对优势；3）正则规则的规模通过模型参数量影响双向神经概率转换器的学习能力，正则规则越多，Bi-NPT 训练后的识别效果越好。

4.1 数据集构建

本实验以中文工艺规范文本实体识别语料作为数据集，该语料来源于制造领域工艺规范文档，涉及装配、复材加工、数控加工与普通机械加工等工艺领域。该语料由 3 名领域专家人工标注，实体标注之间的 Kappa 系数[24]达到 0.68，表明了实体标注的较高一致性，同时也表明了语料的可靠性。语料共 4424 条数据，包含 16 类 16383 个实体，统计信息如表 3 所列。与通用领域相比，中文工艺规范文本实体识别语料规模更小，实体类粒度更细，且“零件编号”和“属性值”等实体构词规律复杂。如何在少样本场景下使模型区分实体的复杂特征并正确识别实体，是中文工艺规范文本实体识别的难点。

表 3 实验数据统计信息

Table 3 Experimental data statistics

实体类型	实体标签	实体示例 1	实体示例 2	数量		
				训练集	验证集	测试集
零件编号	PART_ID	NAS1252	MIL-DTL-24308	132	14	56
数量	PART_NU	3 个	两支	89	17	25
图注	FIGURE_NOTE	图 6-20	图 7-80(a)	267	45	84
表注	TABLE_NOTE	表 7-1	表 6.2	189	41	66
属性值	ATTRIBUTE_VA	90°±1/2°	228.6 mm～355.6 mm (9 in～14 in)	924	132	301
工艺文件	FILE	CVS3001	Q/CRG0078	384	54	118
工艺辅料编号	ACCESSORY_ID	CZM5275	CMS-SL-104B-2	596	91	189
零件	PART	弹簧调节螺钉	双头螺柱	1697	199	511
多余物	REDUNDANT	切屑	毛刺	111	19	34
材料	MATERIAL	玻璃纤维	铝合金	304	53	124
属性	ATTRIBUTE	孔壁表面粗糙度	最大压印厚度	1724	224	531
工具	TOOL	夹钳	铣刀	968	154	308
操作	OPERATION	钻孔	打磨	1297	212	424
工艺辅料	ACCESSORY	润滑剂	密封剂	1278	186	397
部位	PART_AR	扩孔钻出口面	防雨蚀保护罩内表面	956	141	299
孔结构	HOLE	锪窝孔	圆柱孔	279	45	94

将数据集按 7∶1∶2 的比例随机划分为训练集、验证集和测试集，分别用于模型的训练、验证和测试。为了验证不同数据条件下的模型识别效果，考虑到实体在工艺文本中的分布特点，分别以实体数量和句子百分比为数据抽取标准，各训练集统计信息如表 4 所列。其中，N-shots 表示该训练集按实体数量抽取，保证训练数据内每类实体各有 N 个；N%-sen 表示该训练集按句子百分比抽取，抽取训练集中 N% 的句子，同时保证抽取的数据中每类实体数量占全训练数据中该类实体数量的百分比在[N−5%，N+5%]区间内。

表 4 训练集统计信息

Table 4 Statistics of training data sets

训练集	句子数	实体总数	训练集	句子数	实体总数
5-shots	27	80	50%-sen	1547	5560
20-shots	92	320	60%-sen	1856	6661
50-shots	246	800	70%-sen	2165	7821
10%-sen	309	1129	80%-sen	2475	8972
20%-sen	618	2228	90%-sen	2784	10055
30%-sen	928	3342	100%-sen	3094	11195
40%-sen	1 237	4492	—	—	—

4.2 评价指标

本文使用准确率 P(Precision)、召回率 R(Recall)、F1 值来评估各模型实体识别效果，评估方法如式(11)－式(13)所示：

$$P=\frac{\text{模型正确识别实体个数}}{\text{模型识别实体总数}}\times 100\% \tag{11}$$

$$R=\frac{\text{模型正确识别实体个数}}{\text{文本实体总数}}\times 100\% \tag{12}$$

$$F1=\frac{2\times P\times R}{P+R}\times 100\% \tag{13}$$

4.3 正则规则与模型设置

正则规则按 100%-sen 训练集中出现的规则实体编写，由工艺文本语料的标注专家完成编写工作。专家对表 5 中 7 类规则实体进行特征总结，考虑到正则规则应用效率，只保留覆盖 3 个及以上实体的正则规则，得到正则规则共 24 条。为了评估正则规则效率，使用训练集 100%-sen 进行正则匹配，用评价指标 R 作为正则规则覆盖率。正则规则的数量统计、评价结果和模型参数统计如表 5 所列。

表 5 正则规则及模型参数统计

Table 5 Statistics of regular rules and model parameters

实体类型	正则规则数量	实体特征数量	正则规则覆盖率/%	有限状态转换器状态数
零件编号	4	41	92.11	36
数量	2	2	21.43	6
图注	1	2	98.85	6
表注	1	2	87.26	6
属性值	10	215	71.31	86
工艺文件	3	3	92.11	21
工艺辅料编号	3	3	97.54	20
共计	24	268	84.21	181
双向神经概率转换器参数量			66018120	

表 6 列出了双向神经概率转换器的最佳参数设置。采用批量梯度下降作为优化算法，使用 Adam 优化器训练模型，并

设置初始学习率为 0.0001。

表 6 实验参数设置

Table 6 Experimental parameter settings

参数	值
学习率	0.0001
训练批大小	8
迭代次数	30

4.4 实验方法与结果分析

4.4.1 零样本实体识别

为了验证 Bi-NPT 能够有效利用正则规则的先验知识，将 Bi-NPT 不进行标注数据的训练，直接对 7 类规则实体进行测试。此外，为了比较 Bi-NPT 与正则规则识别效果的差异，还使用正则规则对 7 类规则实体进行匹配，结果如表 7 所列。

表 7 零样本下 Bi-NPT 与正则规则实体识别的实验结果

Table 7 Experimental results of Bi-NPT and regular rules in zero-shot

实体类型	Bi-NPT			正则匹配		
	P	*R*	*F*1	*P*	*R*	*F*1
零件编号	72.97	48.21	58.06	97.22	62.5	76.09
数量	14.29	20.00	16.67	19.51	32.00	24.24
图注	98.77	95.24	96.97	98.80	97.62	98.20
表注	100	95.45	97.67	90.28	98.48	94.20
属性值	69.72	58.14	63.41	79.37	33.22	46.84
工艺文件	96.36	89.83	92.98	89.83	89.83	89.83
工艺辅料编号	97.33	96.30	96.81	97.34	96.83	97.08
微平均	83.51	76.04	79.60	87.20	69.01	77.05

Bi-NPT 与正则匹配的结果不完全一致，这是因为 Bi-NPT 的参数虽然由正则规则转化的有限状态转换器构建得到，模型学习前各参数只有 0 或 1，但其在对实体标签进行打分后，还需要通过优先级函数来解决标签之间的冲突问题，因此识别结果不完全相同。对 Bi-NPT 训练前的识别效果进行分析，Bi-NPT 对零件编号、数量等实体的识别效果劣于正则匹配，但对表注、属性值等实体的识别效果优于正则匹配。由于表注、属性值等实体在数据中的分布远多于零件编号与数量，因此 Bi-NPT 的微平均效果与正则匹配相比提升了 2.55%。Bi-NPT 在训练前已经有效建模了正则规则所携带的实体识别知识，不经过标注数据的训练便可直接应用，与传统深度学习方法相比，具有明显优势。

4.4.2 少样本和充足样本实体识别

本文将按照实体数量抽取的训练样本作为少样本条件，将按照句子百分比抽取的训练样本作为充足样本条件，分布训练 Bi-NPT。选用当前效果较突出的深度学习模型与本文方法进行了对比实验，对工艺文本的 16 类实体进行评估，结果如表 8 所列。

表 8 少样本与充足样本下的模型对比实验

Table 8 Experimental results(F1s) in few-shot and rich-resource settings

模型	少样本条件(*F*1)			模型	充足样本条件(*F*1)		
	5-shots	20-shots	50-shots		10%-sen	50%-sen	100%-sen
Bi-NPT	29.22	28.01	29.98	Bi-NPT	41.18	55.66	60.45
PER	7.02	16.47	25.18	BiLSTM	42.64	54.49	59.61
Template-based BART	5.3	12.52	15.68	PER	41.11	60.23	65.33
NNShot	15.47	27.2	29.5	TENER	26.91	51.39	58.25
QaNER	22.44	34.94	39.43	FLAT	43.51	57.98	64.11

BiLSTM[3]是当前较为主流的，可用于处理序列任务的深度学习模型；PER[2]是专门面向工艺大纲所提出的识别方法。与工艺大纲不同，本文使用的工艺规范语料规模更小，且实体类型与工艺大纲实体相比存在差异。Template-based BART[9]，NNShot[10]和 QaNER[25]是专门面向小样本的实体识别方法。其中，Template-based BART 是基于提示模板的方法，将实体识别看作生成式问题；NNShot 引入最近邻分类器，采用最近邻原理计算词的相似度，在向量空间中选最近词的类别进行标注；QaNER 将提示模板改进为问答框架，并使用了掩码语言模型用于填充模板。TENER[4]和 FLAT[5]是当前效果较好的命名实体识别模型，TENER 带有方向与相对位置信息的 atteniton[26]机制，使用 Transformer 对信息建模用于 NER 任务；FLAT 利用 lattice 信息并引入了相对位置编码，利用字词的跨度信息提升识别效率。为了比较模型本身的推理性能，我们弱化了解码器对识别结果的影响，上述所有模型的解码均使用 Softmax 解码方法。PER 模型中使用的正则规则和 Bi-NPT 预定义所用正则规则相同。

在少样本条件下，Bi-NPT 较优的识别性能主要来自正则规则携带的规则实体构词特征。当每个实体类型只提供 5 个样本时，Bi-NPT 优于所对比的其他方法，表示本文方法在极少样本下具有优势。在 20-shots 和 50-shots 数据下 Bi-NPT 的效果不及 QaNER，这是因为 QaNER 额外使用掩码语言模型用于填充模板，所以与本文所使用的实体构词规则相比，QaNER 的模板可覆盖更多的实体。样本增加至 50-shots 时，Bi-NPT 的识别效果仍明显好于 PER，Template-based BART 和 NNShot 模型。

在充足样本条件下，Bi-NPT 优于 TENER 模型。在 50%-sen 和 100%-sen 数据下，Bi-NPT 与 PER 和 FLAT 相比效果较差。这是由于 PER 和 FLAT 通过充足标注数据训练后，提取到了比 Bi-NPT 领域知识更多的特征。PER 使用卷积网络提取预识别实体的特征，可提取的特征比实体的构词特征更丰富；同理，FLAT 除字编码外，还引入了词汇信息，因此在充足样本条件下识别效果更好。Bi-NPT 的效果与 BiLSTM 相当，甚至在 50%-sen 和 100%-sen 数据下超过了

BiLSTM，可见该模型具有与深度神经网络等同的学习能力，Bi-NPT 模型具有很好的研究空间与发展前景。

综上，本文提出的 Bi-NPT 模型在少样本条件下，与面向少样本实体识别的 QaNER 模型具有可比性，甚至在极少样本下表现更优，且性能优于现有的深度学习模型；而在充足样本条件下，能够充分利用标注数据的训练来扩充自身性能，与深度学习模型相比具有同等竞争力。当前专用于小样本条件的模型在训练数据较多的情况下难以捕捉数据间的丰富特征，而适用于较多训练数据的模型在小样本条件下易发生过拟合；而 Bi-NPT 可灵活适应任何训练数据量情况，与其他模型相比具有相对优势。

此外，本文对模型本身的学习能力进行了微观分析，选用 5-shots，20-shots，50-shots，10%-sen，50%-sen 和 100%-sen 分布训练 Bi-NPT，分别对正则规则覆盖的实体和未覆盖的实体进行评估，结果如图 4 所示。可以看出，无论是少样本还是充足样本条件下，Bi-NPT 都可以从样本中学习，提升实体识别的能力。训练过程中，正则规则定义的实体构词特征转化为 Bi-NPT 状态间转换概率的表征，模型向着梯度下降的方向优化整体参数，即使样本极少，Bi-NPT 对规则未覆盖实体的识别性能仍得到提升。使用 5-shots 和 20-shots 数据训练模型时，Bi-NPT 在测试集上的性能低于训练前，这是因为 Bi-NPT 使用训练数据进行训练后，通过在验证集上的表现来判断是否将其作为最优模型，测试集和验证集虽然按实体比例划分，但仍具有样本差异，即模型在训练集和验证集上优化得到的效果没有在测试集上体现。

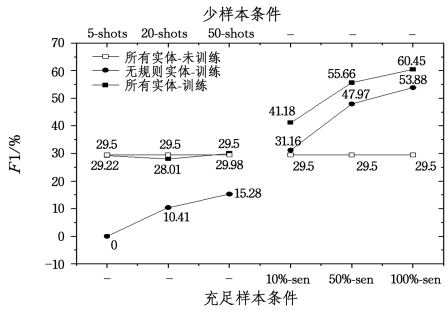

图 4　Bi-NPT 实体识别结果的微观分析

Fig. 4　Microscopic analysis of Bi-NPT entity recognition results

4.4.3　正则规则规模实验

本文将正则规则的规模作为自变量，分别取零规则、一条正则规则和完整正则规则预定义的 Bi-NPT 模型，使用按句子百分比抽取的 10 组训练样本进行训练，比较不同规模正则规则初始化的模型的学习能力。其中零规则表示每类实体只有一条空正则规则，该规则指定了实体标签，但未指定实体的字符级特征，即使用通配符“$”配合实体标签，完成模型的初始化过程；一条规则指每类实体经规则合并后得到的一条规则；完整规则指先前编写的全部规则，实验结果如图 5 所示。

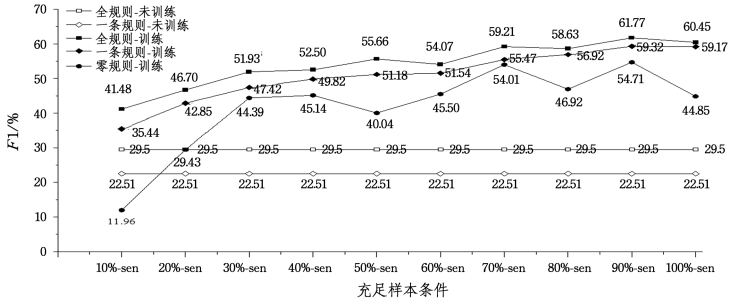

图 5　Bi-NPT 在不同正则规则规模下的实体识别情况

Fig. 5　Entity identification with/without rules with different data gradients of Bi-NPT

可以看出，对于每一组训练数据，Bi-NPT 训练后的识别效果与训练前相比都有提升，说明 Bi-NPT 具有从标注数据中学习的能力。Bi-NPT 的学习能力与识别效果受正则规则数量的影响，正则规则数量越多，模型的识别效果越好，其中全规则的 Bi-NPT 识别效果最好。这是因为更多的正则规则可以带给模型更多的先验知识，并为模型提供更多的参数支持训练。

结束语　当前实体识别领域的研究往往关注深度学习方法，而对于专业领域而言，领域内的先验知识没有得到高效利用，先验知识在模型训练前无法发挥作用。因此，本文提出了面向工艺领域实体识别的双向神经概率转换器(Bi-NPT)，使用正则规则定义领域内实体构词特征与实体标签的映射关系，将正则规则转化为有限状态转换器并从中构建 Bi-NPT 所需参数。Bi-NPT 在未训练的情况下，就拥有识别实体的能力，并且通过在标注数据上训练，识别正则规则覆盖范围外的实体。本方法高效利用了先验知识，并有效降低了模型对数据的需求。实验表明，该模型在训练前具备正则规则的性能，且可以通过标注数据进行训练，效果与深度神经网络模型相当，无论是与面向小样本模型相比还是与深度学习模型相比，都具有相对优势。此外，本文进一步分析了正则规则的规模对模型学习效果的影响，为未来工作提供了可尝试的方向。

在未来工作中，我们将尝试在模型的学习过程中动态增减参数量，使模型的参数量适于训练数据量，避免模型参数不足或过大而造成学习不稳定的现象。此外，我们还将尝试将动态规划加入模型的解码过程中，使模型的输出更加合理，从而提升模型的性能。

参 考 文 献

[1] ZHANG N N，WANG P Y，ZHANG G P. Named Entity Deep Learning Recognition Method for Process Operation Description Text[J]. Computer Applications and Software，2019，36(11)：188-195，261.

[2] JIA M，WANG P Y，ZHANG G P，et al. Named Entity Recognition for Process Text[J]. Journal of Chinese Information Processing，2022，36(3)：54-63.

[3] SIAMI-NAMINI S，TAVAKOLI N，NAMIN A S. The perfor-

mance of LSTM and BiLSTM in forecasting time series[C]// 2019 IEEE International Conference on Big Data (Big Data). IEEE, 2019: 3285-3292.

[4] YAN H, DENG B C, LI X N, et al. TENER: adapting transformer encoder for named entity recognition[J]. arXiv: 1911.04474, 2019.

[5] LI X N, YAN H, QIU X P, et al. FLAT: Chinese NER using flat-lattice transformer[C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 6836-6842.

[6] MA J H, WANG L Q, YAO S. Named Entity Recognition for Chemical Resource Text[J]. Journal of Zhengzhou University (Natural Science Edition), 2018, 50(4): 14-20.

[7] FENG Y T, ZHANG H J, HAO W N. Named Entity Recognition for Military Text[J]. Journal of Computer Science, 2015, 42(7): 15-18.

[8] ZHU S L. Deep Learing Based Uyghur Named Entities Recognition[J]. Journal of Computer Engineering And Design, 2019, 40(10): 2874-2878, 2890.

[9] CUI L, WU Y, LIU J, et al. Template-Based Named Entity Recognition Using BART[C]// Findings of the Association for Computational Linguistics: ACL (IJCNLP 2021). 2021: 1835-1845.

[10] YANG Y, KATIYAR A. Simple and Effective Few-Shot Named Entity Recognition with Structured Nearest Neighbor Learning[C]// Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020: 6365-6375.

[11] LIU H B, ZHANG D M, XIONG S F, et al. Named Entity Recognition of Wheat Diseases and Pests fusing ALBERT and Rules[J]. Journal of Frontiers of Computer Science and Technology, 2023, 17(6): 1395-1404.

[12] CUI S G, CHEN J Y, LI X H. Named Entity Recognition for Chinese Electronic Medical Record by Fusing Semantic and Boundary Information[J]. Journal of University of Electronic Science and Technology of China, 2022, 51(4): 565-571.

[13] ZHANG Z W, CHEN J Y, GAO K N, et al. SVR-BIGRU-CRF Based Chinese Named Entity Recognition for Education Domain[J]. Journal of Chinese Information Processing, 2022, 36(7): 114-122.

[14] LIU W P, ZHANG B, CHEN W R, et al. Military Named Entity Recognition Based on Transfer Representation Learning[J]. Command Information System and Technology, 2020, 62(2): 68-73.

[15] PENG H, SCHWARTZ R, THOMSON S, et al. Rational Recurrences[J]. arXiv: 1808.09357, 2018.

[16] LIN C, ZHU H, GORMLEY M R, et al. Neural finite-state transducers: Beyond rational relations[C]// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). 2019: 272-283.

[17] RASTOGI P, COTTERELL R, EISNER J. Weighting Finite-State Transductions With Neural Context[C]// Proceedings of NAACL-HLT. 2016: 623-633.

[18] JIANG C Y, ZHAO Y G, CHU S B, et al. Cold-start and interpretability: Turning regular expressions into trainable recurrent neural networks[C]// Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020: 3193-3207.

[19] JIANG C Y, JIN Z J, TU K W. Neuralizing Regular Expressions for Slot Filling[C]// Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021: 9481-9498.

[20] SAKUMA Y, MINAMIDE Y, VORONKOV A. Translating regular expression matching into transducers[J]. Journal of Applied Logic, 2012, 10(1): 32-51.

[21] THOMPSON K. Programming techniques: Regular expression search algorithm[J]. Communications of the ACM, 1968, 11(6): 419-422.

[22] HOPCROFT J. An n log n algorithm for minimizing states in a finite automaton[M]// Theory of machines and computations. Academic Press, 1971: 189-196.

[23] FRANCISCO C, COLIN D. Computational complexity of problems on probabilistic grammars and transducers[C]// International Colloquium on Grammatical Inference, 2000: 15-24.

[24] COHEN J. A coefficient of agreement for nominal scales[J]. Educational and Psychological Measurement, 1960, 20(1): 37-46.

[25] LIU A T, XIAO W, ZHU H, et al. QaNER: Prompting question answering models for few-shot named entity recognition[J]. arXiv: 2203.01543, 2022.

[26] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. arXiv: 1706.03762, 2017.

LI Ruiting, born in 1999, postgraduate, is a student member of CCF (No. O9179G). Her main research interests include artificial intelligence and knowledge engineering.

WANG Peiyan, born in 1983, senior engineer, is a member of CCF (No. 33066M). His main research interests include naturallanguage processing, machine learning and knowledge engineering.