文章编号:1007-5321(2020)05-0084-07 **DOI**:10.13190/j.jbupt.2020-032

基于数据增强的中文医疗命名实体识别

王蓬辉, 李明正, 李 思
(北京邮电大学 人工智能学院, 北京 100876)

摘要: 由于缺乏大量已标注数据,在中文医疗命名实体识别中,主要利用外部资源来改善医疗实体识别的性能,这需要大量的时间和有效的规则加入外部资源. 为了解决标注数据不足的问题,提出了一种基于生成对抗网络的数据增强算法,自动生成大量标注数据,提高医疗实体识别的性能. 实验结果表明,该算法在性能方面优于实验中的基准模型,证明了该算法在医疗实体识别上的有效性.

关 键 词: 命名实体识别; 数据增强; 序列生成对抗网络
中图分类号: TP181 **文献标志码**: A

Data Augmentation for Chinese Clinical Named Entity Recognition

WANG Peng-hui, LI Ming-zheng, LI Si
(School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing 100876, China)

Abstract: Chinese clinical named entity recognition plays an important role in recognizing medical entities contained in Chinese electronic medical records. Limited to lack of large annotated data, most of existing methods concentrate on employing external resources to improve the performance of clinical named entity recognition, which require lots of time and efficient rules. To solve the problem of lack of large annotated data, data augmentation using sequence adversarial generative network is used to generate more various data depending on entities and non-entities in the training set. Experiments show that when using generated data to expand training set, the proposed named entity recognition system has achieved competitive performance compared with state-of-art methods, which shows the effectiveness of our data augmentation method.

Key words: named entity recognition; data augmentation; generative adversarial network

作为信息提取的基本任务,命名实体识别(NER,named entity recognition)在过去几年受到了广泛关注. 对于通用领域,例如新闻领域,研究人员主要关注 3 种基本实体类型,即人名、地名、组织机构名. 目前,命名实体识别任务在通用领域中已存在大量标注数据,现有方法利用神经网络提取高效的信息特征,达到了良好的实体识别性能[1-2]. 在基于神经网络的命名实体识别过程中,基于字符的长短期记忆网络(LSTM, long-short term memory network)[1]常用于获取丰富的上下文信息,卷积神经网络(CNN, convolutional neural networks)[2]用于提取字符级语义表示.

在中文医疗文本领域,医疗文本中包含了大量医疗领域特有的实体,如疾病和解剖结构. 这些特

收稿日期: 2020-03-24
基金项目: 国家自然科学基金项目(61702047)
作者简介: 王蓬辉(1996—), 男, 硕士生.
通信作者: 李 思(1985—), 女, 副教授, E-mail: lisi@bupt.edu.cn.

有实体需要具有医疗知识的人员才可以做准确标识，这使得在医疗领域通常缺乏大量的标注数据. 因此，采用神经网络的方法难以获取高效的信息特征，这使得医疗实体识别的准确度不高. 近期，中文医疗命名实体识别的研究工作取得了一定的进展. 这些工作主要从 2 个方面来改善医疗命名实体识别的性能. 一方面，修改深度模型的结构. 多任务的学习[3]被用于综合考虑医疗文本中不同类别的实体的识别结果；注意力机制[4]通过赋予不同的字符以不同的权重，来提取文本中更加重要的信息. 另一方面，引入外部资源，以帮助提高命名实体识别的准确性. 外部的实体字典通常以额外特征的形式集成到模型中[5]. 此外，扩大训练集也是直接改善性能的方法[6]. 但是，通过人工标注大量的数据集不切实际. 因为在医疗领域，对医疗数据进行准确的标注需要专业的医疗人员，这会花费大量的时间和人力.

通过分析中文医疗文本数据，研究人员发现，医疗文本的表达具有一定的规律性，多数医疗实体总是出现在相似的非实体后面. 其中，非实体指的是文本中那些不属于预定义的实体类别字符. 如图 1 所示，“下腹部”和“中下腹”出现在相同的非实体部分“患者因”后面. 因此，笔者采用了序列生成网络，通过学习医疗文本中实体和非实体部分的关系，来自动生成数据，扩大训练集，达到提高医疗命名实体识别性能的目的.

	非实体	实体	非实体
句子	患者因	下腹部	隐痛不适于2014年4月6日就诊我院门诊
标签	O O O	B-ANA I-ANA I-ANA	O O O O O O O O O O O O O O O O O O O O

	非实体	实体	非实体
句子	患者因	中下腹	痛入我院普外科治疗
标签	O O O	B-ANA I-ANA I-ANA	O O O O O O O O O

图 1　在医疗数据集中的一些实例

笔者提出一种基于序列生成对抗网络的数据增强算法用于命名实体识别(DA-NER，data augmentation for named entity recognition)模型，以缓解命名实体识别任务中标注数据不足的情况. DA-NER 模型的本质是通过对抗生成的方式进行数据增强，以减轻大量标注数据缺乏的影响. 在 DA-NER 模型中，有 2 个问题需要解决. 一是生成的文本数据是离散序列，很难将梯度从判别器传递到生成器. 笔者借鉴了强化学习[7]的思路，将判别器的输出作为奖励来指导生成器. 二是生成器仅生成句子，没有对应的标签无法用于扩大训练集. DA-NER 模型采用了序列生成变换，解决了生成数据的标签问题.

笔者的主要贡献包括：① 提出了 DA-NER 模型，学习训练集中的实体和非实体之间的关系，自动生成标注数据，缓解了缺乏大量已标注医疗数据的影响. ② 在医疗数据集和其他领域数据集上的实验结果表明，该模型不仅能提升医疗命名实体识别的性能，还可以应用到其他领域.

1　相关工作

在研究方法上，早期的命名实体识别任务主要集中在人工设计特征和规则上，以实现高性能的命名实体识别. Zhang 等[8]在进行中文命名实体识别任务中，设计了 7 种和人名相关的规则来识别中文人名. Chen 等[9]则引入了三元组和二元组的特征. 近几年，随着深度学习的不断发展，神经网络成为解决命名实体识别任务的主流方法，不仅解决了人工设计特征的问题，还提升了命名实体识别的性能. 在英文命名实体识别方面，Collobert 等[10]采用了 CNN 和条件随机场(CRF，conditional random field)获得了比人工设计特征更好的模型效果. 在中文命名实体识别上，Zhang 等[11]的 Lattice 模型不依赖分词信息，但是可以更加高效地利用词的信息. Zhu 等[12]将 CNN 与注意力机制进行了结合.

在研究领域上，早期的命名实体研究工作主要专注在新闻等通用领域上，其中 He 等[13]采用 CRF 方法在中文新闻领域的命名实体进行了识别. 近年来，随着医疗信息化的高速发展，针对医疗数据的命名实体研究工作受到了广泛的关注. Wu 等[14]采用了无监督学习的方式从未标注的医疗数据中增强字符表示，然后再结合神经网络应用于命名实体识别. Wang 等[15]采用图像特征和语音特征来增强字符表示，改善了医疗命名实体识别的性能. 这些方法都在一定程度上解决了医疗命名实体识别数据缺乏的问题.

与之前的方法不同，针对医疗数据缺乏的问题，笔者采用了序列生成对抗网络，通过学习训练集中实体和非实体部分之间的关系，生成多样化的数据，提升命名实体识别的性能.

2　序列生成变换

在以往的文本生成[7]任务中，字符常被视为组成一个完整序列的基本单元. 生成器通过学习字符之间的关系，生成一个字符序列$\{c_1, c_2, \cdots, c_n\}$，$c_i \in$

v,其中 v 是数据集中所有字符形成的字典. 但是,在命名实体识别数据上如果采用这样的方式进行数据生成,生成器只能生成一些没有标签的序列. 如果采用监督学习的方式,生成数据无法直接用于扩大训练数据集,进而提升命名实体识别性能. 为此,笔者对生成器的生成过程进行了变换,以解决标签问题.

笔者对医疗文本进行了分析. 正如在图 1 所示,结合命名实体识别任务,当考虑句子的实体标签时,医疗文本"患者因下腹部隐痛不适 3 月于 2014-04-06-就诊我院门诊"包括 3 个部分:解剖部位实体"下腹部"、非实体部分"患者因"、"隐痛不适 3 月于 2014-04-06 就诊我院门诊". 根据字符的标签,基于字符的文本序列可以看成包含实体和非实体部分的序列. 那么,在命名实体识别数据的生成过程中,生成字符的过程可以看成生成实体和非实体部分的过程. 如果采用训练集中的数据来训练生成器生成数据,由于生成数据的实体和非实体部分都来自训练集,那么就可以采用字符串匹配找到其对应的标签,这样就解决了生成数据的标签问题.

为了用实体和非实体部分来表示训练集中的所有句子,首先得构建一个包含实体和非实体的集合,给定训练集

$$S = \{X_{1:T}, Y_{1:T}\}$$

其中 X_1 和 Y_1 为数据集中的句子及其标签;然后初始化集合 $B = \{\}$,对于训练集中的 $\{X_t, Y_t\}$,根据标签找到其中的实体和非实体部分,然后把它们加入集合.

具体来说,假设一个文本序列 $\{c_1, c_2, c_3, c_4, c_5, c_6\}$ 的标签是 $\{O, O, B\text{-}PER, I\text{-}PER, O, O\}$,可以将 c_1c_2, c_5c_6 归为非实体部分,c_3c_4 归为实体部分,然后将它们和对应的标签加入集合中.

3 模型

3.1 模型框架

DA-NER 模型结构如图 2 所示. 该模型包含生成器、判别器和 NER 模型 3 个主要部分.

3.2 生成器

生成器的目标是学习训练集中实体和非实体之间的隐藏关系,然后生成可以欺骗判别器的数据,用于扩大训练集. 在生成序列时,采用从左向右生成序列的策略,因为这种生成方式符合汉语的习惯. 生成器从初始状态 l_0 开始生成序列,直到生成序列

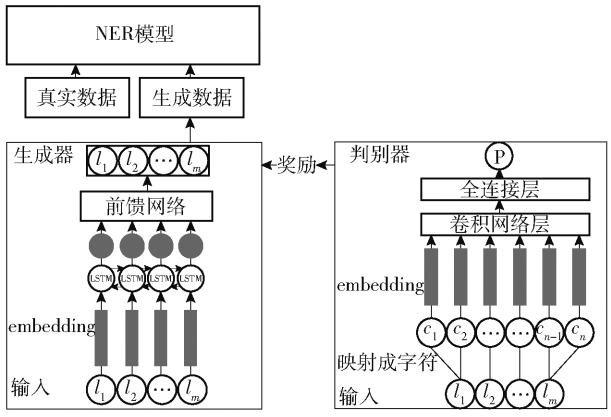

图 2 DA-NER 整体模型结构

的长度等于设定的长度才停止. 其中生成序列的基本单元来源于训练集中的实体和非实体部分.

在生成器中,先随机初始化一个向量字典,用于查询每个输入基本单元所对应的数值化矩阵. 生成器在每个时刻的输出具有一定的时序关系. 因此,笔者采用 LSTM 作为生成器,来建立输出单元之间的时序关系. 具体来说,假设生成器的最终输出序列是 $\{l_1, l_2, \cdots, l_m\}$,其中 m 是设定的生成序列长度,那么当生成器要预测 i 时刻的输出时,有

$$\boldsymbol{h}_i^g = F(\boldsymbol{h}_{i-1}^g, l_{i-1}) \tag{1}$$

$$p(l_i \mid l_0, l_1, \cdots, l_{i-1}) = \mathrm{softmax}(W_f \boldsymbol{h}_i^g + b_f) \tag{2}$$

其中:$\boldsymbol{h}_{i-1}^g$ 为 $i-1$ 时刻生成器中 LSTM 的隐层输出,l_{i-1} 为 $i-1$ 时刻生成器的输出,F 为生成器中 LSTM 模块,W_f 和 b_f 为前馈网络可训练的参数权重,采用 $\boldsymbol{h}_{i-1}^g$ 来初始化 i 时刻的 LSTM 是为了引入前一时刻的信息. 然后前馈网络将获取的隐层状态信息映射成所有可能的输出单元的概率. softmax 函数用于将输出的概率归一化. 最终选择概率最高的单元作为输出.

在生成对抗网络中,梯度更新对于维持生成器和判别器之间的平衡十分重要. 但是,在文本生成中,由于生成数据是离散的文本序列,这使得梯度更新无法在生成器和判别器进行传播. 为了解决这个问题,笔者借鉴序列生成对抗网络(SeqGAN, sequence generative adversarial nets)[7]的思想,采用强化学习的方法解决梯度更新的问题,通过判别器的分数来指导生成器的训练过程. 图 3 展示了 SeqGAN 的结构. 判别器作为决策网络,状态是生成器每次生成的数据,奖励是判别器的输出. 判别器每次接受一个完整的序列输入,给出分数,来给予生成器一定的奖励.

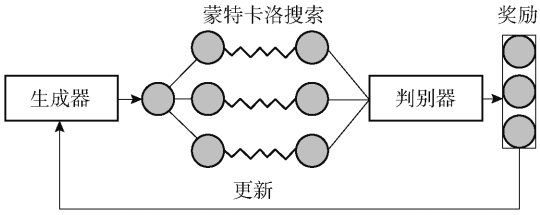

图 3　Seq-GAN 模型结构

在生成序列过程中，仅考虑之前时刻状态的影响是不够的，当前时刻的输出对于整个输出序列的影响也需要考虑进去. 因此，笔者通过蒙特卡洛搜索中的 roll-out 策略，对之后时刻的输出状态进行采样，来考虑当前 i 时刻的输出对于整个输出序列的影响. 对于 i 时刻之后的输出进行了 K 次采样，有

$$[(l_1,l_2,\cdots,l_m)^1,(l_1,l_2,\cdots,l_m)^2,\cdots,(l_1,l_2,\cdots,l_m)^K]=\mathrm{MC}[(l_1,l_2,\cdots,l_i),K] \tag{3}$$

其中：l_i 为当前时刻的输出，m 为设定的最大输出序列长度，K 为蒙特卡洛搜索的采样次数，MC 为蒙特卡洛搜索方法.

这样，生成器每个时刻产生一个输出单元，都会通过采样后形成完整的输出序列，然后判别器对完整序列进行判断，给出当前时刻的输出的分数，指导生成器，而不是当生成器输出一个完整的序列之后再进行判断. i 时刻生成器获得的奖励为

$$R_i=\frac{1}{K}\sum_{k=0}^{K}D[(l_1,l_2,\cdots,l_m)^K] \tag{4}$$

其中 D 为判别器函数. 生成器的目标函数是最大化期望，有

$$J(\theta)=\sum_{i=1}^{m}E_{l_{1:i}\sim G_\theta}[\mathrm{lb}G_\theta(l_i|l_{0:i-1})R_i] \tag{5}$$

其中：$E_{l_{1:i}\sim G_\theta}$是对生成器输出的序列 $l_{1:i}$概率求期望；G_θ 为生成器函数，$G_\theta(l_i|l_{0:i-1})$为生成器在输出序列为 $l_{0:i-1}$下输出 l_i 的概率.

3.3　判别器

判别器可以视为一个文本分类模型，接收一个完整的文本序列，然后判断该序列是否为真实的数据. CNN 因其良好的性能常常被用来构建文本分类网络，故笔者采用了 CNN 构建判别器.

判别器主要由 CNN 和全连接网络组成，其具体网络结构见图 2. 具体来说，给定输入序列为$\{l_1,l_2,\cdots,l_m\}$，判别器首先将序列中的每个单元即实体或者非实体部分映射成对应的字符序列$\{c_1,c_2,\cdots,c_n\}$. 之所以将序列映射成字符序列是因为字符序列包含更加丰富的字符级别信息，有利于判别器判断序列的真假.

接着，通过字符嵌入，可以将每个字符映射成对应的向量. 同时生成的数据中还包括字符的实体标签信息，判别器可以结合每个字符的标签信息来判断字符序列在命名实体识别任务中的合理性. 在结合了数据的标签信息后，每个字符的表示方式为

$$\boldsymbol{x}_i=[\boldsymbol{e}_{c_i}:\boldsymbol{t}_i] \tag{6}$$

其中：$\boldsymbol{e}_{c_i}$为字符 c_i 的字向量，$\boldsymbol{t}_i$ 为字符 c_i 对应的标签. CNN 用于提取输入序列的局部特征为

$$\boldsymbol{h}_i^{\mathrm{d}}=\boldsymbol{W}_{\mathrm{CNN}}^{\mathrm{T}}[\boldsymbol{x}_{i-\frac{w-1}{2}}:\boldsymbol{x}_{i+\frac{w-1}{2}}]+\boldsymbol{b}_{\mathrm{CNN}}^{\mathrm{T}} \tag{7}$$

其中：$\boldsymbol{W}_{\mathrm{CNN}}^{\mathrm{T}}$和 $\boldsymbol{b}_{\mathrm{CNN}}^{\mathrm{T}}$为卷积核的参数，$w$ 为卷积核的窗口大小. 再采用最大池化操作得到序列为

$$\boldsymbol{o}=\max\{\boldsymbol{h}_1^{\mathrm{d}},\boldsymbol{h}_2^{\mathrm{d}},\cdots,\boldsymbol{h}_m^{\mathrm{d}}\} \tag{8}$$

最后，全连接网络用于将最终的序列表征映射到判断序列为真假的概率. 与生成器的目标相反，判别器的目标是能够正确判断输入序列是否为真实数据，具体地，判别器的目标函数为

$$\max_{D_\varphi}E_{p_{\mathrm{data}}}[\mathrm{lb}D(l_1,l_2,\cdots,l_m)]+E_{l_1,l_2,\cdots,l_m\sim G_\theta}[1-\mathrm{lb}D(l_1,l_2,\cdots,l_m)] \tag{9}$$

其中 D_φ和 G_θ 分别为判别器和生成器的参数.

3.4　NER 模型

借鉴前人的命名实体识别工作[16]，LSTM 在命名实体识别中常用于提取文本特征，CRF 则用于特征提取之后的解码过程. 笔者也采用 LSTM + CRF 的模型结构作为 NER 模型. 与之前模型的不同之处在于，NER 模型的输入不仅包括真实数据，还包括生成器的生成数据.

给定输入序列$\{c_1,c_2,\cdots,c_n\}$，NER 模型首先在预训练的字向量词典中找到每个字符对应的数值化向量$\{\boldsymbol{e}_1,\boldsymbol{e}_2,\cdots,\boldsymbol{e}_n\}$. 然后 LSTM 用于获取每个字符上下文相关的隐层状态信息. 前馈神经网络则将 LSTM 输出的隐层状态映射成每个字符的标签概率. CRF 用于建立输出标签之间的相关性和解码得到最后的标签输出. 其中，标签序列的输出概率定义为

$$P(Y|X)=\frac{\exp[s(X,Y)]}{\sum_{Y'}\exp[s((X,Y'))]} \tag{10}$$

其中：s 为 CRF 中的特征函数，X 为输入序列，Y 为真实的标签序列，Y'为任意输出的标签序列.

在模型训练中，NER 模型的目标是最大化真实标签序列的概率，对应的损失函数为

$$L = -\sum_{i=0}^{N} \mathrm{lb}P(Y_i \mid X_i) \tag{11}$$

其中:N 为训练集中的句子总数;$P(Y_i \mid X_i)$ 为输入文本序列 X_i 时,NER 模型的预测标签序列为 Y_i 的条件概率.

4 实验

4.1 数据统计

在实验过程中,使用了 4 个数据集来验证 DA-NER 数据增强算法在命名实体识别任务上的有效性. 为了验证数据增强算法在医疗文本上的有效性,在医疗命名实体识别数据集 CCKS 2019① 和 CMID② 进行了实验,同时为了进一步探究该算法是否适用于其他领域,选取了常用的中文命名实体识别数据集 Weibo NER[16] 和 Resume[11],并且在这些数据上进行了实验. 4 个数据集的统计结果见表 1.

表 1 4 个数据集的数据统计

数据集	数据集划分	句子数/文档数	实体类型
CCKS 2019	训练集	1 000	疾病,症状,影像检查,解剖部位,药物,手术
	开发集	100	
	测试集	400	
CMID	训练集	9 803	疾病,症状,影像检查,解剖部位,药物,手术
	开发集	–	
	测试集	2 451	
Resume	训练集	3 821	人名,国籍,种族,教育地理位置,组织机构,专业,头衔
	开发集	463	
	测试集	477	
Weibo NER	训练集	1 350	人名,地理位置,地缘政治,组织机构
	开发集	270	
	测试集	270	

4.2 性能评估

在医疗数据上,采用 2 个基准模型来验证数据增强方法的有效性:一种是基于字符的模型,使用的是基于字符的 LSTM + CRF 结构[11];一种是采用 Bert[17] 作为字符的预训练字向量的模型.

从表 2 可知,通过数据增强的方式,DA-NER 模型取得了比基准模型更好的结果. 在不使用 Bert 的情况下,DA-NER 模型在 CCKS 2019 数据集上的 F 值达到了 81.76%,在 CMID 数据集上的 F 值达到了 57.12%,分别比基准模型高 0.8% 和 0.68%. 在使用 Bert 的情况下,DA-NER 模型在 CCKS 2019 数据集上的 F 值达到了 83.40%,在 CMID 数据集上的 F 值达到了 59.31%,分别比基准模型高 0.65% 和 0.96%.

表 2 在医疗数据集上的实验结果

数据集	方法	准确率	召回率	F 值
CCKS 2019	Baseline[11]	81.09	80.82	80.96
	DA-NER	81.73	81.79	81.76
	Baseline + Bert	81.82	83.69	82.75
	DA-NER + Bert	82.58	84.24	83.40
CMID	Baseline[11]	57.97	54.99	56.44
	DA-NER	58.86	55.49	57.12
	Baseline + Bert	58.01	58.69	58.35
	DA-NER + Bert	58.22	60.45	59.31

这些实验结果验证了数据增强的方法在医疗数据集上的有效性.

此外,笔者针对不同长度的输出单元对医疗命名实体识别性能的影响进行了研究. 图 4 所示为采用 DA-NER 模型在 CCKS 2019 数据集上取得的实验结果.

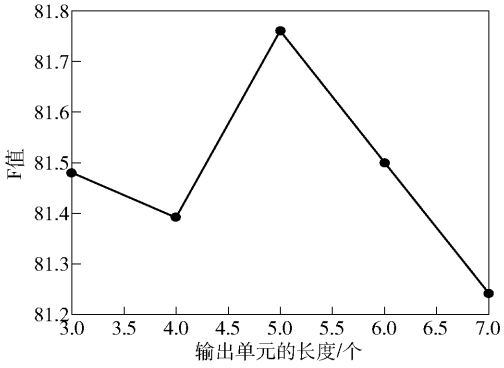

图 4 不同长度的输出单元下性能的比较

可以看出,设置不同长度的输出单元会对命名实体识别的性能造成影响,输出单元设置过长或者过短都会造成实体识别性能的下降,当设置输出单元的长度为 5 时,医疗命名实体识别的性能最佳.

4.3 扩展实验

为了探究笔者提出的数据增强方法是否还适用其他领域,研究人员还在 Weibo NER 和 Resume 数据集上进行了实验,其中基于字符的 LSTM + CRF 模型[11] 作为基准模型. 在现有的先进命名实体识别系统中,笔者选择 Lattice 模型[11] 和 CAN-NER 模型[12] 与 DA-NER 模型进行了对比,探究数据增强算

① http://www.ccks2019.cn/

② https://github.com/liutongyang/CMID

法与现有的先进命名实体识别方法是否具有可比性.

采用 DA-NER 模型在 Weibo NER 和 Resume 数据集上的实验分别取得了 59.42% 和 95.28% 的 F 值,性能不仅超越了基准模型,而且与 Lattice 模型和 CAN-NER 模型相比也有提升,如表 3 所示.

表 3　DA-NER 模型在 Weibo NER 和 Resume 上的实验结果

数据集	模型	准确率	召回率	F 值
Weibo NER	Lattice model[11]	–	–	58.79
	CAN-NER[12]	–	–	59.31
	Baseline[11]	–	–	56.75
	DA-NER	69.01	52.17	59.42
Resume	Lattice model[11]	94.81	94.11	94.46
	CAN-NER[12]	95.05	94.82	94.94
	Baseline[11]	94.53	94.29	94.41
	DA-NER	95.22	95.34	95.28

5　实例分析

为了进一步分析数据增强方法的有效性,笔者分析了不同数据集上的真实数据和生成数据,如表 4 所示. 在 CCKS 2019 和 CMID 数据集上,真实数据中“胃”、“直肠癌”实体,在生成数据中变成了“下腹”、“肛瘘”实体,可见,数据增强的方法在生成数据时可以生成多样性的实体部分的数据. 在 CCKS 2019 和 CMID 数据集上,生成数据中的非实体部分“隐痛”替代了“可见癌转移”,生成数据中的非实体部分“怎么治疗”替代了“想买点药吃”,说明数据增强的方法在生成数据时也能生成多样性的非实体部分数据. 同样地,在 Weibo NER 上,真实数据中的非实体部分“裸婚时代”在生成数据中变成了“30 年前”,在 Resume 上,真实数据中的“刘昊维”,在生成数据中是“黄忠和”,表明数据增强方法在其他数据集上也能产生多样化的数据.

表 4　DA-NER 模型在数据集上的生成数据和真实数据的例子

数据集	生成数据	真实数据
CCKS 2019	下腹壁不均匀增厚伴周围多发小淋巴结	胃壁不均匀增厚伴周围多发小淋巴结
	网膜淋巴结隐痛	网膜淋巴结可见癌转移
CMID	最近检查出有高血压,怎么治疗	最近检查出有高血压,想买点药吃
	肛瘘通过哪些方法筛查	直肠癌通过哪些方法筛查
Weibo NER	30 年前,刘易阳说:细节打败爱情原来是真的	裸婚时代刘易阳说:细节打败爱情原来是真的
Resume	黄忠和先生,现任公司监事会职工代表监事	刘昊维先生,现任公司监事会职工代表监事

上述分析说明,数据增强的方法在生成数据时可以生成多样化的句子,不仅表现在实体的多样化,还表现在非实体部分多样化,以此来扩大训练数据集,提高命名实体的识别性能.

6　结束语

笔者提出了一种基于序列生成对抗网络的数据增强算法,即 DA-NER 模型,通过扩大训练集,达到提高命名实体识别性能的目的. 实验结果表明,DA-NER 模型可以在不使用外部资源的情况下,生成更加多样化的训练数据,来扩大数据集,不仅在医学领域,而且在其他领域也能提高命名实体的识别性能. 未来工作考虑尝试在更大数据集中进行,并和外部知识库进行结合以提高精度.

参考文献:

[1] Dong Chuanhai, Zhang Jiajun, Zong Chengqing, et al. Character based LSTM-CRF with radical-level features for Chinese named entity recognition[C]//Natural Language Understanding and Intelligent Applications - 5th Conference on Natural Language Processing and Chinese Computing(NLPCC). Kunming: Springer Press, 2016: 239-250.

[2] Ma Xuezhe, Hovy E. End-to-end sequence labeling via Bi-directional LSTM-CNNs-CRF[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL). Berlin: ACL, 2016: 1064-1074.

[3] Wang Xuan, Zhang Yu, Ren Xiang, et al. Cross-type biomedical named entity recognition with deep multi-task learning[J]. Bioinformatics, 2019, 35(10): 1745-1752.

[4] Li Luqi, Zhao Jie, Hou Li, et al. An attention-based deep learning model for clinical named entity recognition of Chinese electronic medical records[J]. BMC Med Inf & Decision Making, 2019, 19(5): 4.

[5] Wang Qi, Zhou Yangming, Tong Ruan, et al. Incorporating dictionaries into deep neural networks for the Chinese

clinical named entity recognition[J]. J Biomed Informatics, 2019: 92.

[6] Cui Zongyong, Zhang Mingrui, Cao Zongjie et al. Image data augmentation for SAR sensor via generative adversarial nets[J]. IEEE Access, 2019, 7: 42255-42268.

[7] Yu Lantao, Zhang Weinan, Wang Jun, et al. Sequence generative adversarial nets with policy gradient[C]// Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI). San Francisco: AAAI, 2017: 2852-2858.

[8] Zhang Suxiang, Qin Ying, Wen Juan, et al. Word segmentation and named entity recognition for sighan bakeoff3[C]// Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing. Sydney: ACL, 2006: 158-161.

[9] Chen Aitao, Peng Fuchun, Shan Roy, et al. Chinese named entity recognition with conditional probabilistic models[C]// Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing. Sydney: ACL, 2006: 173-176.

[10] Collobert R, Weston J, Bottou L, et al. Natural language processing (almost) from scratch[J]. Journal of Machine Learning Research, 2011, 12: 2493-2537.

[11] Zhang Yue, Yang Jie. Chinese NER using lattice LSTM [C]// Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL). Melbourne: ACL, 2018: 1554-1564.

[12] Zhu Yuying, Wang Guoxin. CAN-NER: convolutional attention network for Chinese named entity recognition [C]// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT). Minneapolis: ACL, 2019: 3384-3393.

[13] He Jingzhou, Wang Houfeng. Chinese named entity recognition and word segmentation based on character[C]// Third International Joint Conference on Natural Language Processing (IJCNLP). Hyderabad: ACL, 2008: 128-132.

[14] Wu Yonghui, Jiang Min, Lei Jianbo, et al. Named entity recognition in Chinese clinical text using deep neural network[C]// eHealth-enabled Health - Proceedings of the 15th World Congress on Health and Biomedical Informatics. Sao Paulo: IOS Press, 2015: 624-628.

[15] Wang Yifei, Ananiadou S, Tsujii J. Improve Chinese clinical named entity recognition performance by using the graphical and phonetic feature[C]// International Conference on Bioinformatics and Biomedicine(BIBM). Madrid: IEEE Press, 2018: 1582-1586

[16] Peng Nanyun, Dredze M. Named entity recognition for Chinese social media with jointly trained embeddings [C]// Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP). Lisbon: ACL, 2015: 548-554.

[17] Devlin J, Chang Mingwei, Lee K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[C]// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies(NAACL-HLT). Minneapolis: ACL, 2019: 4171-4186.