文章编号：1003-0077(2019)06-0072-08

面向新类型人名识别的数据增强方法

宋希良[1,2]，韩先培[1]，孙 乐[1]

(1. 中国科学院 软件研究所 中文信息处理实验室，北京 100190；2. 中国科学院大学，北京 100049)

摘 要：人名识别常被作为命名实体识别任务的一部分，与其他类型的实体同时进行识别。当前使用 NER 方法的人名识别依赖于训练语料对特定类型人名的覆盖，在遇到新类型人名时性能显著下降。针对上述问题，该文提出了一种基于数据增强(data augmentation)的方法，使用新类型人名实体替换的策略来生成伪训练数据，该方法能够有效提升系统对新类型人名的识别性能。为了选择有代表性的特定类型人名实体，该文提出了贪心的代表性子类型人名选择算法。在使用 1998 年《人民日报》数据自动生成的伪测试数据和人工标注的新闻数据的测试结果中，多个模型上人名识别的 F_1 值分别提升了至少 12 个百分点和 6 个百分点。

关键词：人名识别；Data Augmentation；新类型人名

中图分类号：TP391　　**文献标识码**：A

Data Augmentation Method for New Type Person Named Entity Recognition

SONG Xiliang[1,2], HAN Xianpei[1], SUN Le[1]

(1. Chinese Information Processing Laboratory, Institute of Software, Chinese Academy of Sciences, Beijing 100190, China; 2. University of Chinese Academy of Sciences, Beijing 100049, China)

Abstract: Person name recognition tasks are often performed as part of the named entity recognition (NER) tasks, along with other types of entities. Currently, person name recognition method relies on the coverage of the training corpus for a particular type of person name, and the performance is significantly degraded when a new type of person name is encountered. To address this issue, we propose a method namesd Data Augmentation. In this method, we generate pseudo training data by replacing the common person name entities in training data with new specific types of entities. This method can effectively improve the recognition performance of the system for new types of person names. We propose a greedy representative subtype name selection algorithm which can select typical person name of a specific type. We conduct experiments on two test data sets: one is pseudo test data set based on the People's Daily data in 1998 and the other is manually labeled news data. The F_1 measure of the recognition result is increased by at least 12% and 6%, respectively.

Keywords: person name recognition; data augmentation; new type of person name

0 引言

命名实体识别(named entity recognition, NER)是信息抽取中的基础任务，旨在从无结构的文本中识别出人名、地名和组织结构名等类型的实体。人名识别可以作为命名实体识别任务的一部分，使用命名实体识别的方法与其他类型的实体同时进行识别；其也可以作为一项单独的任务，使用基于规则、词典、统计以及混合的方法进行识别。

基于规则的方法需要领域专家总结并维护大量的规则，需要相应的人力成本；基于词典的方法只能识别词典范围内的实体，泛化性能差。除纯粹基于规则和词典的方法外，当前的人名识别方法依赖于训练语料对特定类型人名的覆盖，在遇到新类型人名时识别性能显著下降。例如，《人民日报》语料中的人名大部分都是汉族人名，系统在遇到维吾尔人名、日本人名和苏俄人名等子类型人名时性

收稿日期：2019-01-27　**定稿日期**：2019-02-28

基金项目：国家自然科学基金(61433015，61572477，61772505)；中国科协青年人才托举工程(YESS20160177)

能很差。

针对上述问题重新构建训练语料是一种耗时耗力的任务。有研究者利用 Wikipedia 的结构自动标注数据用于训练数据，如 DBpedia Spotlight①、TagMe②、AIDA③ 等，但其存在两个问题：一是其实体来源于 Wikipedia 实体集合，大部分是比较常见的实体，且其类型和规模不能进行扩展；二是其文本来自 Wikipedia 文本，训练得到的模型对其他类型的文本（如来自社交媒体的文本）的性能会变差。也有研究者提出基于词典的数据标注方法，文献[1]提出了两种使用词典进行标注训练的方法，一种是使用生语料库词典匹配的训练方法（DMC Training），另一种是使用生语料库自动标注加词典增强与标注语料库相结合的训练方法（DECAC Training）。此类方法也存在两个问题：一是这种类似于远距离监督学习的方法会产生标注噪声，词典中的实体在不同的上下文中可能不是实体或者与词典中的实体类型不一致；二是对于不在词典中的真实实体则会被错误地标注成非实体。

此外，针对子类型人名识别问题，有学者针对不同的人名子类型，总结该子类型人名特点，使用规则、词典、统计以及相混合的方法构建特定于该类型的人名识别系统。文献[2-3]使用统计和规则相结合的方法来识别日本人名以及音译人名，文献[4]利用统计的方法结合总结的维吾尔人名的构成规则来进行维吾尔族人名的识别，文献[5]针对音译人名的发音特点，将中文拼音与对应外语的字符串映射到国际音标字母表（international phonetic alphabet，IPA）然后基于发音相似度进行音译人名的识别。该类型方法的准确率较高，但移植到其他人名子类型的灵活性比较差。

针对以上问题，本文提出了一种基于数据增强（data augmentation）的方法，通过简单的新类型实体替换策略来生成伪训练数据，有效提升系统对不同新类型人名的识别性能。对每种子类型人名，本文提出了贪心的代表性子类型人名选择算法来选择有代表性的该类型人名实体的子集，让模型自动学习该子类型的人名构成特点，无需特定于该类型人名的先验知识。

本文的组织结构如下：第 1 节主要介绍相关工作；第 2 节主要介绍本文所采用的数据增强方法；第 3 节主要介绍实验设置、实验结果和实验结论；最后一节介绍本文的结论和未来工作。

1　相关工作

人名识别可以作为 NER 任务的一部分，可以采用 NER 的方法进行识别，此类方法通常将任务建模成序列标注任务，使用的统计模型有隐马尔可夫模型 HMM[6]、条件随机场模型 CRF[7-8] 以及深度学习模型[9-12] 等。对特定的子类型人名识别时，当前的方法主要采用针对特定于该子类型的特点，利用该子类型先验知识，建立特定于该子类型人名的模型。文献[13]使用基于语义角色标注的方法，利用中国人名及上下文中的不同角色作用，来进行中国人名识别，该方法依赖于训练语料中对中国人名以及上下文的角色覆盖情况。也有学者针对人名的构成特点，使用混合策略的方法进行人名的识别。文献[5]利用外国人名的中文音译名的发音特点，提出了基于中文和外文相似度的外文翻译人名的识别方法；文献[3,14]提出了 CRF 模型初筛—人名可行度模型确认—上下文规则筛选—局部统计算法进行边界纠正—全文扩散未识别人名的统计和规则相结合的线式方法，规则主要来源于基于错误驱动的转换学习和基于人名的边界纠正规则。以上的方法的性能均依赖于标注数据，标注数据的覆盖情况、质量及规模决定了模型的性能。

与本文相关的一项评测任务是 WNUT2017④ 组织的评测任务“Emerging and Rare entity recognition”，该任务旨在从最新出现的文本如社交媒体文本中，识别出比较稀少或未出现过的实体。该评测任务的潜在要求是待识别的实体在训练数据中没有出现过或者出现的次数极少，因此该任务限制了训练数据中实体的覆盖度。参与评测任务队伍模型的主要框架是基于词、字符的 Bi-LSTM-CRF 模型，各个队伍主要进行了两个方向的探索：一是实体本身信息的探索如实体、组块、词典等更深层次的信息；二是实体上下文信息的探索，如全局上下文信息和局部上下文信息。

也有学者研究不同领域间的命名实体识别问题，即领域适应问题。其假设与本文稍有不同：假设源领域和目标领域的上下文分布不同，但实体的类别标

① https://www.dbpedia-spotlight.org/

② https://tagme.d4science.org/tagme/

③ https://gate.d5.mpi-inf.mpg.de/webaida

④ http://noisy-text.github.io/2017/emerging-rare-entities.html

签相同或类似。根据目标领域中的数据是否有标注，可以分为两种类型的任务：第一种，目标领域没有标注数据，只有大量未标注数据；第二种，目标领域有少量标注数据和大量未标注数据。第一种任务主要采用的是无监督领域适应方法，文献[15-16]利用主题模型如LSA和LDA将特征映射到潜在语义空间，以此来进行领域适应；文献[17]使用迭代训练的方式，在大量的未标注目标数据集上训练模型。在第二种任务中，针对目标领域的少量标注数据，不同的研究者提出了不同的使用方式：文献[18]与主题模型类似，将源领域和目标领域的特征组织成层次的树状结构，然后在训练目标领域模型时，使用源领域模型的先验知识，以此来进行领域适应；文献[19-20]在领域之间共享全部或部分架构源，但其目的是使用源领域训练的参数对目标领域参数进行初始化，以此来利用神经网络学习到的源领域的语义先验知识。

2 数据增强方法

新类型人名实体的识别性能依赖于训练数据对这些人名的覆盖，而通用领域的人名训练数据往往不包含新类型的人名实体，或者仅包含很少量的新类型人名实体。包含新类型人名的真实训练数据不易获取，但其人名集合及通用领域人名训练数据容易获取。本文假设不同类型的人名出现的上下文分布一致，基于此假设提出了数据增强(data augmentation，DA)方法，通过获取不同人名子类型的词典，以简单的替换策略来自动生成符合语法和假设的新类型人名标注实例。

2.1 人名上下文条件独立性假设

在不同的人名子类型出现的上下文中，有很多通用的上下文，例如，人名后面可以有表示动作性的词语，前面可以有表示头衔的词语。这些比较通用的上下文对判定候选词是否是人名提供必要的信息；另一方面，在不同的人名子类型出现的上下文中，也有特定于其子类型的上下文，例如，日本人名的上下文可以是包含与日本有关的地名、组织机构名等，这些上下文对区分该人名的子类型提供重要的信息。特定人名子类型的上下文只能通过包含该子类型人名的真实标注数据获取，由于缺乏真实的标注数据，这一类信息很难获取到，但通用的上下文信息可以在通用领域的标注数据中获取。在本文中，不同子类型的人名统一标注为PER，不区分其子类型，同时为了充分利用通用人名标注数据来自动生成子类型人名标注数据，本节提出了人名上下文条件独立性假设。该假设指不同类型的人名出现的上下文分布一致，具体地，给定出现人名实体的上下文，其出现的人名子类型与上下文无关。更形式地，给定人名左上下文 C_{left}，以及右上下文 C_{right}，其出现不同人名子类型 per_i，per_j 的概率相等，如式(1)所示。

$$P(\text{per}_i \mid C_{\text{left}}, C_{\text{right}}) = P(\text{per}_j \mid C_{\text{left}}, C_{\text{right}}), \quad \forall i \neq j \tag{1}$$

例如，“今天上午，{PER} 出席了会议，并做了大会报告。”中的PER可以是任何人名子类型的实体。

本文在Chinese Gigaword第二版中的新华社语料[①]中，统计了不同类型的人名及其上下文的分布情况。本文在使用Stanford NER工具进行标注以后，将人名分为汉族人名、欧美人名、日本人名、苏俄人名及新疆维吾尔族人名(维族人名)，计算各种子类型人名之间的JS散度，计算结果如表1所示。从表1中可以看出，各种子类型人名之间的JS散度值比较小，而与整个语料的JS散度值比较大，这在一定程度上验证了本文提出的上下文条件独立性假设。

表1 各种人名子类型上下文以及语料分布的JS距离

	语料上下文	汉族人名	欧美人名	日本人名	苏俄人名	维族人名
语料上下文	0.000	0.162	0.201	0.228	0.250	0.239
汉族人名	0.162	0.000	0.130	0.160	0.161	0.162
欧美人名	0.201	0.130	0.000	0.149	0.053	0.104
日本人名	0.228	0.160	0.149	0.000	0.169	0.180
苏俄人名	0.250	0.161	0.053	0.169	0.000	0.109
维族人名	0.239	0.162	0.104	0.180	0.109	0.000

2.3 新类型人名选择与词典获取

本文选择的人名的子类型是维族人名子类型、日本人名子类型以及苏俄人名子类型。维吾尔人名虽然属于中国人名，但由于维吾尔人有自己的独立语言，人名数量多、规律性差、随意性大、结构成分复杂、歧义性较大，识别起来存在着一定困难[4]，且因其相关的研究比较少，因此本文选择了维吾尔人名子类型作为新类型人名之一。文献[21]对人名做了深入统计：在3.8万个欧美人名、4.4万个苏俄人名

① https://catalog.ldc.upenn.edu/LDC2005T14

和 1.5 万个日本人名的实体名列表上，300 个高频欧美人名用字覆盖了 98.75%的欧美人名，200 个高频苏俄人名覆盖了 99.32%的苏俄人名，而 1 000 个高频日语人名用字覆盖了 94.19%的日本人名。相比欧美人名，日本人名用字相对比较广，姓氏比较多，且还有许多与地名重合的部分，识别起来更具有挑战性，因此本文选择了日本人名子类型作为第二种新类型人名。苏俄人名的识别相关研究比较少，本文选择将其作为第三种新类型人名进行识别。

新类型人名词表存储了新类型人名使用的字符和词语的分布和组合情况，主要用于生成新类型人名训练语料。本文采用多种策略，从互联网上获取对应的子类型人名实体表，主要来源于现成的人名词表，如搜狗词库①；双语人名词表，如新华社世界人名翻译大辞典[22]；对应子类型人名的垂直网站，如新疆地区政府网站、教育网站的公示信息等。

2.3　训练实例生成

获取新类型人名词典以后，需要使用该词典与通用人名的标注语料生成新类型人名的标注数据。当新类型的人名资源不容易获取时，获取到的新类型人名词典的规模比较小，这时可以使用该新类型人名词典的全部人名实体生成标注数据，这样不会使得学习到的模型在标注时倾向于该类型人名实体。本文获取到的三种新类型人名实体的规模与训练语料标注实例在同一数量级上，使用全部的词典会使得新产生的标注数据中该人名子类型的标注实例频率远超于其他类型（如通用人名）实例，而且使用过多的词典，会使得上下文重复出现，模型在训练数据上出现过拟合现象。因此产生新类型标注数据时，需要选择新类型词典中的一个子集，该子集能够有效地代表整个新类型词典，使用其产生更合理的标注数据。

本节从两个方面研究了代表整个字典的子集选择方法。一方面，选择的子集的词汇能尽可能多地覆盖整个词典的词汇。这里的词汇可以是字符，也可以是分词的子词。本文研究了使用字符时的子集覆盖情况，发现维族人名和苏俄人名的子集的规模仅仅在几百时就能覆盖整个集合，日本人名子集规模稍微大一些，但其规模相对于训练实例的数量都太小，而使用字符覆盖不能保证子词的覆盖，这对基于词粒度的人名识别系统的帮助是有限的。本文使用的人名实体识别模型是基于词粒度的，因此选择使用子词的覆盖度进行子集选择，同时在分词时本文选择最小粒度的分词器，这样尽可能地减少分词错误，减少子集选择的规模。给定子集的规模大小 N，本文定义的覆盖度如式(2)所示。

$$\text{Coverage} = |U_{\text{dict}} \cap U_N| / |U_{\text{dict}}| \tag{2}$$

其中，U_N 表示规模为 N 的子集的子词集合，U_{dict} 表示整个新类型词典的子词的集合，$|U|$ 表示集合 U 的元素数量。

另一方面，词典中的子词有的是高频子词，包含该子词的人名也是相对高频人名，包含词典中的长尾子词人名是相对低频人名。为了使得子集能够包含更多的高频子词，同时也覆盖长尾的子词，选择的子集子词分布要尽可能地与整个词典子词分布接近。本文使用 KL 距离描述子集和全集的子词分布相似程度，如式(3)所示。

$$\text{KL}(p \| q) = \sum_{x \in U_{\text{dict}}} p(x) \log \frac{p(x)}{q(x)} \tag{3}$$

其中，p 是全集的子词分布，q 是子集的子词分布。

上述的两个目标相互影响，同时优化比较困难。为了简化问题，本文提出了贪心的代表性子类型人名选择算法，将以上两个过程分开，分别使用贪心的替代算法。首先设定子集的初始规模 N，使用贪心的策略最大化字符覆盖度；之后设定子集增大的最大规模为 $\alpha N(\alpha>1)$，使用贪心的策略最大化子集和全集的子词分布相似度。上述过程的算法描述如算法 1 所示。

Algorithm 1：代表性子类型人名选择算法

Input：特定新类型人名实体集合 U_{dict}，覆盖度最大候选集合数 N，候选子集最大数 $\alpha N(\alpha>1)$
Output：候选新类型人名子集 $U_{\alpha N}$；
候选子类型子集大小 count←0；
候选子类型子集 $U_{\alpha N}$←[]；
新类型人名全集剩余集合 U_{left}←U_{dict}
While count＜N **do**
　从集合 U_{left} 选择一个人名 person，最大化覆盖度 Coverage；
　$U_{\alpha N} = U_{\alpha N} \cup \{\text{person}\}$；
　$U_{\text{left}} = U_{left} \setminus \{\text{person}\}$；
　count＝count＋1；
End
While count＜αN **do**
　从集合 U_{left} 选择一个人名 person，最大化 KL($p \| q$)减少量；
　If Kl($p \| q$)＞0 **then**
　　$U_{\alpha N} = U_{\alpha N} \cup \{\text{person}\}$；
　　$U_{\text{left}} = U_{\text{left}} \setminus \{\text{person}\}$；
　　count＝count＋1；
　End
End
返回 $U_{\alpha N}$

① https://pinyin.sogou.com/dict/

3 实验及其结果

3.1 实验设置

3.1.1 数据及其处理

本文选用1998年《人民日报》1—6月份的语料：1—4月的数据作为基本的训练集，5—6月的数据作为基本的测试集。对这些数据使用ansj[①]细粒度分词器进行分词。在本文中，只考虑人名类别的实体，将地点和组织机构类型的实体忽略。原始训练和测试数据中的人名出现次数和人名出现句子数的统计信息如表2所示。构造的三种新类型人名词典，其规模都在3万以上，如表3所示[②]。

表2 训练测试数据人名统计信息

数据类型	人名出现句子数	人名出现次数
训练数据	51 018	81 562
测试数据	24 002	39 201

表3 新类型人名词典规模

人名子类型	词典规模
维族人名	44 483
日本人名	186 484
苏俄人名	35 146

本文使用两种类型的测试数据：一种是使用数据增强的方式自动生成的包含新类型人名的伪测试数据，其中的子类型人名从不在训练数据中出现的人名词典中随机选择；另一种是从三种新类型人名的新闻网站获取并人工标注的真实数据。这些新闻网站包括天山网[③]、俄罗斯卫星通讯社[④]、日本新闻网[⑤]等。按照三种新类型人名等比例标注，去掉大量不包含新类型人名的句子以及噪声数据，共标注536句，人名实体出现540次。

本文设定每一种新类型人名子集的最大实体数为包含通用人名句子数的1/3，使用覆盖度策略选择子类型人名时，N大小设置为能够覆盖90%该子类型人名词典子词的最小子词数。

3.1.2 模型

为了更好地适应不同人名子类型的识别，提高模型的通用性，本文选择了不同的模型。为了减少模型本身带来的误差，我们选择了两类模型三种实现方法。第一类模型是传统的CRF模型，本文选择了Stanford CRF[8]和CRFsuite[23]两种实现方法；第二种是基于深度学习的模型，本文选择的是Anago[12]实现方法。其中Stanford CRF使用其在OntoNotes数据集上调优的特征集合。CRFSuite的特征模板为窗口为6的上下文字符、子词以及前缀、后缀、长度等。Anago的词向量来自Wikipedia中文数据训练的200维度词向量，同时使用了基于字符的向量，其他参数默认。本文实验中均没有使用词性特征。

3.1.3 对比实验设置

本节使用《人民日报》的原始训练数据分别训练三个模型，以此作为基线系统，记作Base，使用新类型人名增强的训练数据训练的三个模型作为对比系统，记作DA。将训练得到的系统分别在伪测试数据和真实标注数据上进行测试，在模型后面分别使用fake和mannual作标记。评价的指标为人名实体的准确率(P)、召回率(R)和F_1值。

此外，为了对比覆盖度和分布相似性两个因素对实验结果的影响，本文构造了“低覆盖度—高分布相似性(LCHD)”和“高覆盖度—低分布相似性(HCLD)”的两组训练和测试数据集，使用上述的三个模型进行实验。

3.2 实验结果及分析

将基线系统和对比系统在增强的测试数据集上进行测试，伪测试数据中保留了部分原有的通用人名数据。表4展示了三种方法分别在不同语料上的人名识别结果。其中Base表示使用原始的《人民日报》数据训练得到的基线系统，DA表示使用增强的数据训练得到的对比系统。“模型(fake)”和“模型(mannual)”分别表示模型在伪测试数据中的测试结果和在真实标注数据上的测试结果。表5展示了三种词典子集选择策略的实验结果。

总体而言，在伪测试数据和人工标注的新闻数据的测试结果中，人名识别的性能均有显著提升，F_1值分别提升了至少12个和6个百分点。从表4可以看出，在伪测试数据中，三种模型F_1值均提升了12个以上百分点，其中CRFSuite模型提升最高，约20个百分点，其次是Stanford CRF，提升15个

① https://github.com/NLPchina/ansj_seg

② 部分人名来自https://github.com/wainshine/Chinese-Names-Corpus并进行了过滤。

③ http://www.ts.cn/

④ http://sputniknews.cn/

⑤ http://www.ribenxinwen.com/

表 4　三种模型与基线系统对比实验结果

	Base			DA		
	P	R	F_1	P	R	F_1
CRFSuite(fake)	0.799	0.707	0.750	**0.952**	0.946	**0.949**
Stanford CRF(fake)	0.753	0.891	0.790	0.945	0.949	0.947
Anago(fake)	**0.807**	**0.834**	**0.821**	0.945	**0.952**	**0.949**
CRFSuite(mannual)	**0.782**	0.685	0.730	0.802	**0.811**	0.806
Stanford CRF(mannual)	0.733	**0.751**	**0.747**	**0.812**	0.805	**0.809**
Anago(mannual)	0.678	0.638	0.658	0.796	0.771	0.783

表 5　三种词典子集选择策略实验结果

	CRFSuite			Stanford CRF			Anago		
	R	F_1	P	R	F_1	P	R	F_1	
Gold	0.802	**0.811**	**0.806**	0.812	**0.805**	**0.809**	**0.796**	**0.771**	**0.783**
LCHD	0.802	0.805	0.804	**0.823**	0.788	0.805	0.760	0.689	0.723
HCLD	**0.805**	0.807	**0.806**	0.816	0.766	0.791	0.710	0.631	0.668

百分点，Anago 的 F_1 提升 12 个百分点。实验结果表明，在对新类型人名不进行人工标注的情况下，使用新类型人名词典基于数据增强方法生成的伪训练数据，能够充分利用通用人名标注数据的标注结果，显著提升新类型人名的识别性能。

三个模型在伪测试数据上的测试性能接近，但在《人民日报》原有数据集上进行训练的基线模型中，Anago 的 F_1 值最高，其次是 Stanford CRF 模型。在人工标注的真实数据集实验中，三个模型的 F_1 值均提升了 6 个百分点以上，Anago 提升了 12 个百分点，CRFSuite 和 Stanford CRF 分别提升了 7 个和 6 个百分点。真实测试数据的性能整体上要低于在伪测试数据上的性能，主要是由于真实数据来源于最新的新闻数据，与 1998 年《人民日报》行文风格差异很大，人名实体的上下文分布也不完全一致。

基于覆盖度和分布相似性策略选择的两组训练数据训练的三组模型，在伪标注数据集上的性能与基于两种因素的选择结果类似，F_1 相差在 0.5 个百分点以内。其在真实的标注数据集的测试结果如表 5 所示。总体而言，考虑两种因素的选择策略(Gold)性能最佳。CRFSuite 模型的 F_1 值在三组选择策略中非常接近；Stanford CRF 模型的 F_1 也比较接近，基于高分布相似性策略(LCHD)要比基于高覆盖度策略(HCLD)的 F_1 值高 1.4 个百分点；Anago 模型的 F_1 值在三种策略中差异比较大，Gold 策略比 LCHD 高 6 个百分点，比 HCLD 高 11.5 个百分点。通过实验样例分析，虽然实验中 LCHD 是低覆盖度策略，但在真实数据集测试时，与 HCLD 策略的覆盖度差异很小，但后者选择的词表与真实分布相反：HCLD 策略更倾向于选择真实分布中低频的词，这使得模型已覆盖的词典学习存在偏差。通过其他额外的实验分析，我们发现通过高覆盖率选择初始的子集后，通过均匀分布选择剩余的词典词语的策略，实验性能也能接近 Gold 策略的结果。

在本文实验中，基线方法的 F_1 比较高的原因有三点：①《人民日报》语料中含有一些日本人名和音译人名实体；②本文使用的分词器的分词粒度比较小，很多新类型人名实体被分词器分成了单字词，这些单字词在原始的人民日报语料中已经被覆盖了一部分；③使用替换策略增强数据方法生成的测试数据实体的上下文与训练数据分布一致，模型可以根据上下文信息获取部分实体的类别信息。

4　结论与展望

本文介绍了利用新类型人名词典增强训练数据的方法，提出了贪心的代表性子类型人名选择算法，用于解决训练数据不覆盖新类型人名时模型不能有效识别这些人名实体的问题。实验对比了在伪测试数据和真实测试数据下的识别结果，本文提出的方法对识别结果均有显著提高。

目前本文只考虑了人名实体类型，没有考虑其

他实体类型,在未来工作中,我们将探索多种实体类型的数据增强方法,以进一步提高模型对不同实体的各种子类型的识别能力。此外,本文中所选择的词典的子集规模相对比较大,没有深入探究产生最佳性能的最小的词典子集规模,在未来工作中,我们将继续研究选择词典子集的最小规模,以及影响该规模的因素。

参考文献

[1] 刘洋.资源受限场景的命名实体识别方法[D].北京:中国科学院大学硕士学位论文,2016.

[2] Liang Y,Zhu Y. A hybrid approach for Chinese pronunciation-translated person names recognition[C]//Proceedings of International Conference on Audio,Language and Image Processing. IEEE,2008:1305-1310.

[3] 王祖兴,吕钊,顾君忠. 基于混合方法的中文人名识别研究[J]. 计算机工程与应用,2015,51(8):211-217.

[4] 加日拉·买买提热衣木,吐尔根·依布拉音,艾山·吾买尔. 基于统计和规则混合策略的维吾尔人名识别研究[J]. 新疆大学学报(自然科学版),2014(3):319-324.

[5] Busemann S, Zhang Y. Identifying foreign person names in Chinese text[C]//Proceedings of the International Conference on Language Resources and Evaluation. European Language Resources Association,2008: 2556-2563.

[6] Zhou G D, Su J. Named entity recognition using an HMM-based chunk tagger[C]//Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics,2002: 473-480.

[7] Lafferty J D,Mccallum A,Pereira F C N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data[C]//Proceedings of Eighteenth International Conference on Machine Learning. Morgan Kaufmann Publishers Inc,2001:282-289.

[8] Finkel J R, Grenager T, Manning C. Incorporating non-local information into information extraction systems by gibbs sampling[C]//Proceedings of the 43rd annual meeting on association for computational linguistics. Association for Computational Linguistics, 2005: 363-370.

[9] Huang Z,Xu W,Yu K. Bidirectional LSTM-CRF models for sequence tagging[J]. arXiv preprint arXiv: 1508.01991,2015.

[10] Chiu J P C,Nichols E. Named entity recognition with bidirectional LSTM-CNNs[J]. Transactions of the Association for Computational Linguistics, 2016, 4: 357-370.

[11] Ma X,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. Association for Computational Linguistics,2016:1064-1074.

[12] Lample G,Ballesteros M,Subramanian S,et al. Neural architectures for named entity recognition[C]//Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics. Association for Computational Linguistics,2016: 260-270.

[13] 张华平,刘群. 基于角色标注的中国人名自动识别研究[J]. 计算机学报,2004,27(1): 85-91.

[14] Wang Z, Zhu X, Lu Z. A context-aware automatic Chinese transliterated person names recognition approach[C]//Proceedings of the 8th International Conference on Semantics, Knowledge and Grids. IEEE, 2012: 143-149.

[15] Guo H,Zhu H,Guo Z,et al. Domain adaptation with latent semantic association for named entity recognition[C]//Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Association for Computational Linguistics,2009: 281-289.

[16] Nallapati R, Surdeanu M, Manning C. Blind domain transfer for named entity recognition using generative latent topic models[C]//Proceedings of the NIPS 2010 Workshop on Transfer Learning Via Rich Generative Models,2010: 281-289.

[17] Tian T,Dinarelli M,Tellier I,et al. Domain adaptation for named entity recognition using CRFs[C]//Proceedings of the Tenth International Conference on Language Resources and Evaluation. European Language Resources Association (ELRA),2016.

[18] Arnold A,Nallapati R,Cohen W W. Exploiting feature hierarchy for transfer learning in named entity recognition[C]//Proceedings of ACL-08: HLT. Association for Computational Linguistics, 2008: 245-253.

[19] Yang Z, Salakhutdinov R, Cohen W W. Transfer learning for sequence tagging with hierarchical recurrent networks[J]. arXiv preprint arXiv:1703.06345, 2017.

[20] Dong X,Chowdhury S,Qian L,et al. Transfer bi-directional LSTM RNN for named entity recognition in Chinese electronic medical records[C]//Proceedings of e-Health Networking, Applications and Services (Healthcom),2017 IEEE 19th International Conference on. IEEE,2017:1-4.

[21] 吴友政. 汉语问答系统关键技术研究[D]. 北京：中国科学院自动化研究所博士学位论文，2006.
[22] 郭国荣. 世界人名翻译大辞典[J]. 北京：中国对外翻译出版公司，新华社译名室，1993.
[23] Okazaki N. Crfsuite：a fast implementation of conditional random fields (crfs)[EB/OL]. [2015-03-24]. http://www.chokkan.org/software/crfsuite.

宋希良(1990－)，硕士，主要研究领域为信息抽取和知识图谱。
E-mail：xiliangsong@126.com

韩先培(1984－)，博士，研究员，主要研究领域为信息抽取、知识库构建及自然语言处理。
E-mail：hanxianpei@qq.com

孙乐(1971－)，博士，研究员，主要研究领域为信息检索与自然语言处理。
E-mail：lesunle@163.com

(上接第 71 页)

郭磊(1992—)，硕士研究生，主要研究领域为话题演化分析技术。
E-mail：gl940121@icloud.com

李弼程(1970—)，通信作者，博士，教授，博士生导师，主要研究领域为文本分析与理解、语音处理与识别、图像/视频处理与识别、信息融合。
E-mail：lbclm@163.com

赵军磊 (1992—)，硕士研究生，主要研究领域为信息抽取。
E-mail：747285253@qq.com