

文章编号: 1003-0077(2021)05-0063-07

## 基于字符卷积神经网络的生物学变异实体识别方法

宋雅文<sup>1</sup>, 杨志豪<sup>1</sup>, 罗凌<sup>1</sup>, 王磊<sup>2</sup>, 张音<sup>2</sup>, 林鸿飞<sup>1</sup>, 王健<sup>1</sup>

(1. 大连理工大学 计算机科学与技术学院, 辽宁 大连 116024;

2. 军事医学科学院, 北京 100850)

**摘要:** 从海量生物医学文献中挖掘变异信息对生物医学复杂疾病研究具有重要意义。在当前的变异实体识别方法中, 基于条件随机场模型的方法取得了不错效果并成为主流方法, 但存在需要大量特征工程来提升模型性能的缺点。针对此问题, 该文提出一种基于字符卷积神经网络的变异实体识别方法 CharCNN-CNN-CRF。该方法首先利用一个多窗口大小的卷积神经网络获取字符级别的词表示, 然后使用多层卷积神经网络编码上下文信息, 最后通过 CRF 层解码得到整个句子的标签序列。实验结果表明, 该方法仅使用随机初始化的字符向量作为输入就能快速、有效地识别变异实体, 无需复杂的特征工程。同时也在 tmVar 和 MutationFinder 两个数据集上都取得了目前最好的结果( $F$  值分别为 88.34% 和 93.57%)。

**关键词:** 变异实体识别; 卷积神经网络; 条件随机场

**中图分类号:** TP391

**文献标识码:** A

## Biomedical Mutation Entity Recognition Method Based on Character Convolution Neural Network

SONG Yawen<sup>1</sup>, YANG Zhihao<sup>1</sup>, LUO Ling<sup>1</sup>, WANG Lei<sup>2</sup>, ZHANG Yin<sup>2</sup>,

LIN Hongfei<sup>1</sup>, WANG Jian<sup>1</sup>

(1. School of Computer Science and Technology, Dalian University of Technology, Dalian, Liaoning 116024, China;

2. Academy of Military Medical Science, Beijing 100850, China)

**Abstract:** Mining mutation entities from massive biomedical literature is of great significance to the research of complex biomedical diseases. To improve the current solution based conditional random field, this paper proposes a method based on character-level convolutional neural network, i.e. CharCNN-CNN-CRF for short. In this method, we utilize a multi-window convolutional neural network to obtain the character-level word representation. Then we encode the context information with a multi-layer convolutional neural network and obtain the label sequence through the conditional random field layer. The experimental results show that the proposed method achieves state-of-the-art results on both the tmVar and MutationFinder datasets with 88.34% and 93.57% in  $F$ -measure, respectively.

**Keywords:** mutation named entity recognition; convolutional neural network; conditional random field

## 0 引言

随着生物医学研究的发展, 领域专家和学者们为了记录研究成果和促进领域交流, 发表了大量的生物医学文献, 从而促进了生物医学文本挖掘技术的发展。其中在复杂疾病研究领域, 为了分析和解释复杂疾病的序列变异, 从文献中挖掘变异信息成为一项重要的研究内容。

序列变异是指机体内的基因水平的物质(通常是 DNA)发生改变造成的变异。从文献中挖掘变异信息一般指挖掘文本信息中的序列变异, 序列变异的文本提及(text mention)通常称为变异实体, 变异实体识别是变异信息抽取的研究重点之一。本文提到的变异实体主要包含 DNA 变异、蛋白质变异和单核苷酸多态性 SNP 三类, 如图 1 所示, 标注部分 c.356G>A 属于 DNA 变异, rs13267 属于单核苷酸多态性 SNP, C119Y 属于蛋白质变异。

收稿日期: 2019-12-12 定稿日期: 2020-02-02

基金项目: 国家十三五重点研发计划(2016YFC0901900); 国家自然科学基金(61272373, 61340020, 61572102)

The findings of the in vivo confocal microscopy were consistent with those reported in previous reports. Sequencing TACSTD2 revealed a novel homozygous missense mutation c.356G>A, leading to amino acid substitution C119Y in the two affected siblings. The mutation was found to be pathogenic on Sorting Intolerant From Tolerant (SIFT) analysis and was not found in normal controls and unaffected individuals of the family. A synonymous, previously reported, single nucleotide polymorphism (SNP; rs13267) was also seen in all the individuals of the family. Protein modeling studies involving wild-type and mutant protein indicated an exposed cysteine residue in the mutant protein. CONCLUSIONS: A novel TACSTD2 C119Y mutation leading to an amino acid substitution was identified in two affected siblings of a family.

图1 tmVar 数据集变异实体示例

目前变异命名实体识别主要采用基于规则、基于机器学习和基于深度学习的方法。传统基于规则的方法依赖于大量手工设计的正则表达式,其规则的生成也需要领域知识的支撑。如 MutationFinder<sup>[1]</sup>可以识别出文本中的核苷酸或氨基酸变异,该类型变异的命名形式单一固定,仅用简单规则就可达到很好的识别效果。在基于传统机器学习方法中,基于条件随机场(conditional random field, CRF)的方法成为了目前主流。如 VTag<sup>[2]</sup>使用 CRF 方法识别文本中描述的基因组畸变类型、基因组位置和基因组状态变化信息;tmVar<sup>[3]</sup>采用基于 CRF 的综合模型提取蛋白质、DNA 和 RNA 水平描述的广泛序列变异。CRF 模型虽然可以和基于规则的方法相结合,但其性能在很大程度上依赖于复杂的特征工程。

近年来,随着深度学习研究的不断深入,通用新闻领域提出几种相似的应用于命名实体识别任务的神经网络结构并取得了不错的结果<sup>[4-7]</sup>。其中,结合条件随机场的双向长短期记忆网络(BiLSTM-CRF)模型表现较为突出。在生物医学领域,Habibi 等人<sup>[8]</sup>将 BiLSTM-CRF 模型应用到了生物医学语料上(包含药物、疾病、物种、基因蛋白和细胞实体识别五大类语料),其中药物和疾病语料上的实验结果都处于当今先进水平,优于传统的 CRF 基线系统。Zhu 等人<sup>[9]</sup>对比了 BiLSTM-CRF 模型和 CNN-CRF 模型在生物医学领域实体识别上的效果,发现两种模型存在差距,但并不明显。Matos 等人<sup>[10]</sup>将 BiLSTM-CRF 模型应用到变异实体识别任务上,并比较了以字符和单词分别作为模型输入单元的影响,结果表明以字符为输入单元的模型效果更好,但由于没有根据变异实体名的特点进行针对性优化,导致实验结果没有超越基于丰富特征的 CRF 模型。

针对现有方法存在的问题,首先本文提出一种基于字符卷积神经网络的变异实体识别方法。针对变异实体名的特点,本文利用一个多窗口大小的卷积神经网络学习字符级词表示,以充分利用变异实体名的内部字符信息,还能在一定程度上缓解未登录词问题。

其次,本文对比了基于词和基于字符的不同神经网络模型(CNN-CRF 和 BiLSTM-CRF 模型)在变异实体识别任务上的效果。结果表明,与基于词的 BiLSTM-CRF 模型相比,基于字符的 CNN-CRF 模型效果更好、效率更高。最后,本文方法在 MutationFinder<sup>[1]</sup>和 tmVar<sup>[3]</sup>两个语料上均取得了目前的最好结果,模型简单高效且不依赖于复杂的特征工程。

一般实体识别任务被看作序列标注问题,因此本文也采用序列标注最常用的标注机制 IOB (inside, outside, beginning)机制<sup>[11]</sup>。

## 1 变异实体识别模型

本文提出了一种基于字符卷积神经网络的变异实体识别方法,模型 CharCNN-CNN-CRF 整体框架如图 2 所示。模型主要分为基于字符卷积神经网络的词表示层 CharCNN、编码器 CNN 层和解码器 CRF 层三部分。

### 1.1 基于字符卷积神经网络的词表示层

词表示层模型 CharCNN 如图 3 所示。模型将目标词的初始化学向量送入卷积神经网络<sup>[12]</sup>(convolutional neural network, CNN)和最大池化层,抽取词的字符特征,输出目标词的字符级词表示。字符向量选择随机初始化,且在模型的整个网络结构中都是可训的。字符特征抽取过程如式(1)所示。

$$O_i^k = \text{Maxpooling}(\text{ReLU}(W \cdot M_{t,t+k-1} + b)) \quad (1)$$

受 N-Gram 思想启发,本文选择多窗口大小的 CNN 学习字符级词表示,定义窗口大小为  $k$  ( $k \in (1, 2, 3, 4, 5)$ ),每个窗口大小的卷积核数均为 64。经过各自的池化层后,我们尝试了 5 种不同窗口大小的特征合并方式,定义为  $f$ ,分别为最大化(Max)、平均化(Avg)、拼接(Con)和相加(Add)操作。合并过程定义如式(2)所示。

$$O_i = f(O_i^1, O_i^2, O_i^3, O_i^4, O_i^5) \quad (2)$$

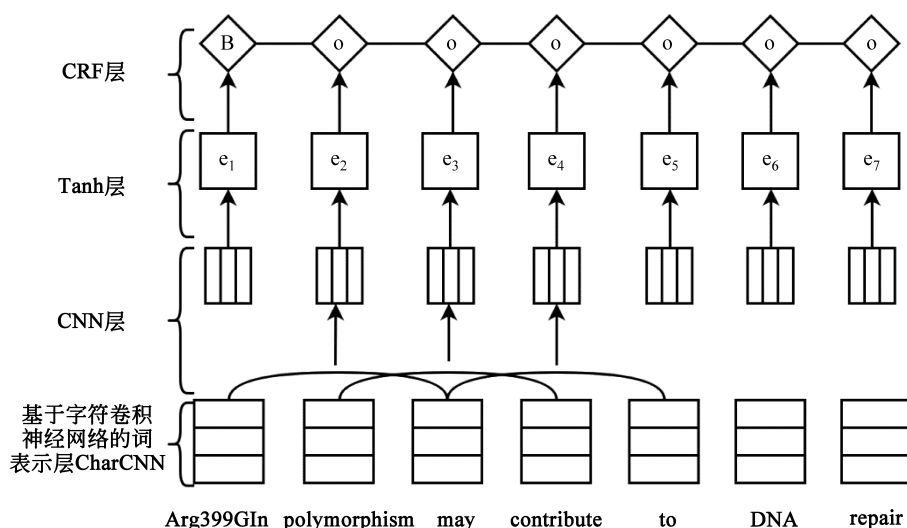


图2 CharCNN-CNN-CRF 模型整体框架

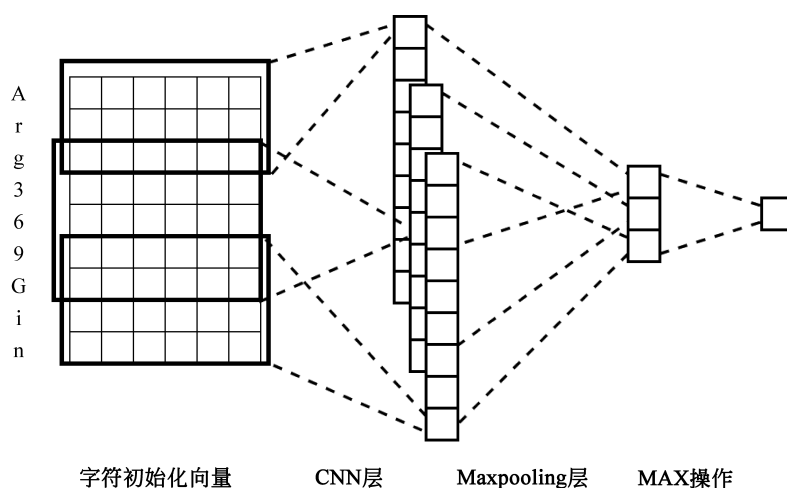


图3 抽取字符特征的卷积神经网络

示例窗口大小 3,4,5

## 1.2 编码器 CNN 层

一般认为 CNN 通过卷积操作提取特定文本粒度下特定窗口大小中的信息,因此更擅长提取重要的局部特征,导致其无法考虑长距离的依赖信息和词序信息,但实践中可以通过栈式叠加多层来缓解。在变异实体识别任务中,变异实体的识别对于长距离信息和词序信息没有太大的依赖性,更加依赖于实体本身的字符组合信息,我们选择如图 2 所示的多层 CNN 的栈式叠加作为文本编码器。

以 CharCNN 抽取的字符级词表示作为输入,文本编码器先用 CNN 层捕获文本中目标词和它相邻词的关联信息,再利用多层 CNN 抽取句子不同语言粒度的局部上下文特征,最后将 CNN 层输出到两层全连接网络,将该词标注作为每一种标签的

最终分值。

## 1.3 解码器 CRF 层

本文将变异实体识别任务当作是序列标注问题。与分类问题相比,序列标注问题中当前的预测标签不仅与当前的输入特征相关,还与之前的预测标签相关,即预测标签序列之间是有强相互依赖关系的。例如,在 IOB 标注机制中, *B* 表示一个实体的开始部分, *I* 表示该实体其余部分, *O* 表示非实体部分。所以在合理的标签序列中 *I* 标签不应该出现在 *O* 标签之后。为了考虑预测标签之间的依赖性,本文使用 CRF 模型来进行解码,从训练数据中获得约束性的规则,保证预测的标签是合法的。

CRF 模型是一种对数线性模型。给定一个句子,CRF 模型将为句子中每一个词预测一个标签。

定义输入句子的单词序列  $x = (x_1, \dots, x_n)$ , 预测标签序列  $y = (y_1, \dots, y_n)$ ,  $Y(x)$  是所有可能的标注序列集合,  $P_{i,y_i}$  是 CNN 编码器输出的第  $i$  个词标注为标签  $y_i$  的预测分值。定义一个标签转移矩阵  $T_{y_i, y_{i+1}}$ , 其中  $T_{i,j}$  表示从标签  $i$  转化为标签  $j$  的分数。模型最终的得分函数  $s(x, y)$  如式(3)所示。

$$s(x, y) = \sum_{i=0}^n T_{y_i, y_{i+1}} + \sum_{i=1}^n P_{i, y_i} \quad (3)$$

使用归一化指数函数计算生成给定输入序列  $x$  的标签  $y$  的条件概率分布是  $P(y|x)$ , 如式(4)所示。然后通过最大化对数似然概率来训练模型参数, 如式(5)所示, 使用 Adam 优化器来优化参数, 选择得分最高的标签序列作为标签预测序列。最后通过维特比算法<sup>[13]</sup>寻找最佳标签序列, 如式(6)所示。

$$P(\mathbf{Y} | \mathbf{X}) = \frac{\exp(s(\mathbf{X}, \mathbf{Y}))}{\sum_{\bar{y} \in Y(\mathbf{X})} \exp(s(x, \bar{y}))} \quad (4)$$

$$\ln(P(\mathbf{Y} | \mathbf{X})) = s(\mathbf{X}, \mathbf{Y}) - \ln \sum_{\bar{y} \in Y(\mathbf{X})} e^{s(x, \bar{y})} \quad (5)$$

$$y^* = \text{Argmax}(s(\mathbf{X}, \bar{y})), \bar{y} \in Y(\mathbf{X}) \quad (6)$$

其中,  $Y(\mathbf{X})$  表示所有可能的标签序列集合,  $y^*$  是得到的最佳标签序列。

## 2 实验分析

### 2.1 实验设置

我们选用 MutationFinder<sup>[1]</sup> 数据集和 tmVar<sup>[3]</sup> 数据集作为实验数据集。tmVar 数据集中标注的变异实体主要包含 DNA 变异、蛋白质变异和单核苷酸多态性 SNP 三类。MutationFinder 数据集中标注的变异实体均为核苷酸或氨基酸变异(属于蛋白质变异类别)。数据集统计信息如表 1 所示。模型的超参数设置如表 2 所示。此外, 从两个数据集的训练集中分别抽取 10% 的样本作为开发集, 用于超参数的选择, 并通过在开发集上的早停策略<sup>[14]</sup>选择模型训练迭代次数。

表 1 数据集统计信息

数据集		句子数量	实体数量
tmVar	训练集	3 648	967
	测试集	1 762	464
MutationFinder	训练集	2 628	264
	测试集	4 642	485

训练 Word2Vec 的数据包括 tmVar 数据集文

摘、MutationFinder 数据集文摘和从 PubMed 上爬取的一部分生物学文摘, 共计 2GB。所有文摘在训练之前用 nltk 工具包<sup>[15]</sup>进行标准分词。

表 2 模型的超参数设置

模块	每层卷积核数量
CharCNN 窗口大小	{1, 2, 3, 4, 5}
CharCNN 卷积核数量	64
CNN 编码器窗口大小	3
CNN 编码器卷积核数量	256
字符查找表规模	87

变异实体识别中, 当实体左右边界和实体类别均正确时为预测正确, 否则均为预测不正确。我们使用查准率 (precision,  $P$ )、召回率 (recall,  $R$ ) 和  $F$  值 ( $F$ -score) 对模型在测试集上的预测结果进行评价。

### 2.2 字符级词表示实验

为了探索适合变异实体的词表示方法, 基于 CNN-CRF 模型, 本文设计了一组词表示实验。实验比较了抽取字符特征的 CNN 和 BiLSTM 模型, 对比了 CNN 层中合并字符特征的不同方式, 如表 3 所示。结果显示窗口大小固定为 3 的 CNN 层模型的  $F$  值是 86.91%, 比 BiLSTM 的  $F$  值 85.01% 高了近两个百分点。可能原因是绝大多数变异实体名中的字符组成遵循一定的行业标准, 有一定的结构特性(图 4), 所以其实体名对字符组成这种局部信息有依赖性。虽然 CNN 和 BiLSTM 都能学习上下文信息, 但 BiLSTM 的优势主要是捕获长距离依赖信息, 而文本中每个词的长度有限, 词内部字符之间的上下文联系紧密, 不至于依赖词外部长距离的字符信息, 故 BiLSTM 的优势在词表示方面无法体现, 因此词表示模型选用卷积神经网络效果更好, 且更轻便。

表 3 字符级词表示实验结果

特征	特征合并	$P/\%$	$R/\%$	$F/\%$
CharCNN(窗口大小 {1, 2, 3, 4, 5})	Max	86.47	90.28	<b>88.34</b>
	Avg	85.09	89.81	87.39
	Con	84.18	88.66	86.36
	Add	85.03	90.74	87.79
CharCNN(窗口大小 {1, 2, 3})	Max	83.95	88.42	86.13
CharCNN(窗口大小为 3)	无	85.45	88.42	<b>86.91</b>
CharLSTM	无	81.66	88.65	85.01



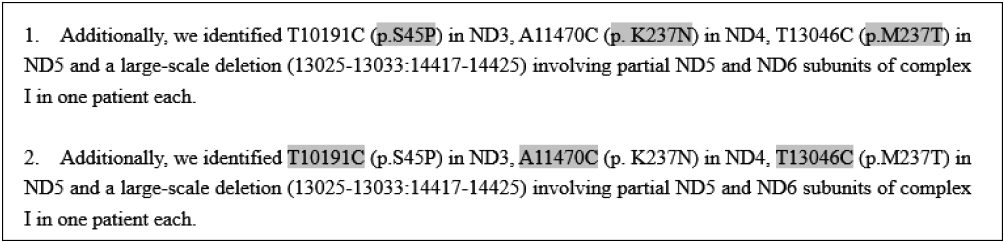


图 4 样本 1 灰色高亮为蛋白质变异实体,样本 2 灰色高亮为 DNA 变异实体

进一步探索如何更好地抽取字符特征,受 N-gram 思想的启发,输入先各自通过窗口大小分别为 1,2,3,4,5 的卷积神经网络,再经过各自的最大池化层,最后以不同的方式合并这些特征,生成字符级别的词表示。我们将抽取字符级词表示的模型命名为 CharCNN。

模型 CharCNN 的  $F$  值为 88.34%, $F$  值和窗口大小仅为 3 的模型相比高约 1.4%。这说明 CharCNN 获得了比常见卷积层窗口大小固定的模型更好的词表示,原因可能是窗口大小不同的卷积层能捕捉到多层次、多角度的字符信息,词表示中包含了更加丰富、对预测更有用的字符特征。

合并方式中最大化操作(Max)的结果  $F$  值为

88.34%,优于其他三种方式。相加操作仅次于最大化操作, $F$  值为 87.79%,而最常用的拼接操作  $F$  值最低,是 86.36%。可能的原因是 Max 合并可以保留对实体识别最关键、最显著和更有利的字符特征信息,为预测阶段奠定更好的基础。

此外实验结果中召回率均高于准确率,通过可视化测试集预测结果发现,预测结果中存在一定量的样本,如图 5 所示,样本中某些词被模型错误预测为实体,其词结构与数据集中 DNA 变异和蛋白质变异实体的结构极其相近,导致了模型的误判。因此,我们认为这种现象的存在是降低模型预测准确率的原因之一。

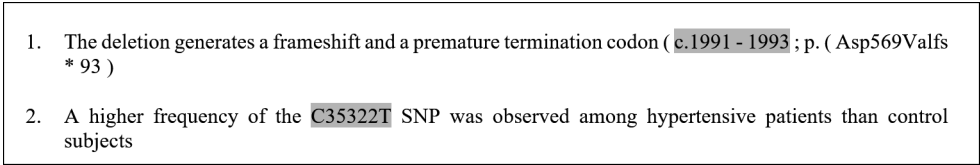


图 5 错误预测示例

2.3 字符和词特征在不同神经网络上的对比实验

常见的命名实体识别任务多以词特征为主体特征,而将字符特征作为辅助特征,以便更好地刻画单词的属性。对于神经网络模型 BiLSTM,由于其能学习到句子长距离依赖信息的优势而在命名实体识别任务中得到广泛的应用。

为了探索不同特征和不同神经网络对模型性能的影响,我们在 tmVar 数据集上设计了一组对比实验,比较了基于字符特征和基于字符特征与词特征组合的 CNN-CRF 模型。对比了基于字符特征的 CNN-CRF 模型和 BiLSTM-CRF 模型,实验结果如表 4 所示。

表 4 tmVar 数据集上字符特征和词特征在不同神经网络上的对比实验结果

模型	特征	$P/\%$	$R/\%$	$F/\%$
CNN-CRF	CharCNN	86.47	90.28	88.34
BiLSTM-CRF	CharCNN	82.31	90.50	86.21

续表

模型	特征	$P/\%$	$R/\%$	$F/\%$
CNN-CRF	Word2Vec	61.97	61.11	61.53
BiLSTM-CRF	Word2Vec	71.05	63.65	67.15
CNN-CRF	CharCNN+Word2Vec	82.21	90.97	86.37

对比 CNN-CRF 模型和 BiLSTM-CRF 模型,当特征仅有字符特征时,模型编码器选用 CNN 的  $F$  值是 88.34%,选用 BiLSTM 的  $F$  值是 86.21%,两者相差近两个百分点,且 CNN-CRF 模型的结果较高。可能的原因是当判断目标词是否属于某个变异实体的一部分时,词序信息或者该词与句子中其他词的长距离依赖信息的联系不够大,更多地是依赖于变异实体名中的局部上下文、字符的组合信息和单词的前缀、后缀等信息。因此当选用字符特征时,CNN 既能捕获词的局部上下文信息,又能从词表示中获取字符信息,而且效率相对较高,导致 BiLSTM 能捕获长距离依赖信息的优势无法体现。

基于词特征的模型中 BiLSTM-CRF 模型的  $F$

值比 CNN-CRF 模型高,可能的原因是当无法利用字符信息时,模型只能通过上下文内容进行预测。此时,BiLSTM 和 CNN 相比更有优势。

当仅用词特征时,CNN-CRF 模型和 BiLSTM-CRF 模型的  $F$  值与各自基于字符特征的模型相比均大幅下降,分别是 61.53% 和 67.15%。可能的原因是基于词特征的模型仅仅包含词级别的信息,无法获取词的字符信息,因此实体识别过程只能依赖于仅有的词信息。这类模型需要事先选择一个词汇量固定的词汇库,不属于词汇库的词汇被映射为未知词。在生物医学文本中,变异实体分词结果与普通词汇相比词频偏低,因而当用 Word2Vec 训练词嵌入时,这类词会被过滤而成为输入文本中的未登录词(out of vocabulary, OOV)。Labeau 等人<sup>[16]</sup>提出了一种基于字符或者词组件方式构建词表示方法,相较于传统语言模型词嵌入初始化的词表示可以有效解决未登录词问题,其结论与本文实验结果一致。

综合各方面因素,我们选择卷积神经网络作为模型的编码器。本文提出的卷积神经网络抽取字符级别词表示模型虽然属于词级别模型,但具有不受词汇表限制、能充分利用字符信息等优点,能缓解基于词特征模型存在的未登录词问题。此外,构建字符特征需要的词汇库很小,计算成本低。

## 2.4 主流方法对比实验

本文提出的模型是 CharCNN-CNN-CRF,在 MutationFinder 数据集上的对比方法包括基于规则的 MutationFinder 方法、基于 CRF 模型的 tmVar 方法。在 tmVar 数据集上的对比方法有 tmVar 方法、Matos<sup>[10]</sup>提出的基于深度学习的方法和基于规

则的方法。基于规则的方法复现了 tmVar 提供的正则表达式,实验结果如表 5 所示。

表 5 主流方法对比实验结果

方法	数据集	$P/\%$	$R/\%$	$F/\%$
MutationFinder <sup>[1]</sup>	MutationFinder	98.40	81.90	89.40
tmVar <sup>[3]</sup>		91.38	91.40	91.39
CharCNN-CNN-CRF		93.09	93.06	<b>93.57</b>
tmVar <sup>[3]</sup>	tmVar	<b>92.01</b>	<b>83.72</b>	<b>87.67</b>
Pedro Matos <sup>[10]</sup>		<b>88.1</b>	<b>86.6</b>	<b>87.4</b>
CharCNN-CNN-CRF		<b>86.47</b>	<b>90.28</b>	<b>88.34</b>
规则方法		<b>75.09</b>	<b>42.24</b>	<b>54.06</b>

对于 MutationFinder 数据集,本文方法 CharCNN-CNN-CRF 的  $F$  值达到了 93.57%,相比于 MutationFinder 方法有所提高,可能原因是该数据集中的变异实体形式单一,模型中的字符级词表示层能充分学习变异实体名的字符特征,生成更具优势的词表示,从而获得较好的实体识别能力。

对于 tmVar 数据集,CharCNN-CNN-CRF 在没有后处理的情况下,与 tmVar 的基于综合 CRF 模型  $F$  值 87.67% 相比有所提升。原因可能是语料中除遵循 HGVS 命名规则的结构化变异实体名外,不可避免地存在一些结构信息不明显、需要上下文信息才能识别出来的变异实体,如图 6 所示。本文方法不仅能通过字符表示层学习实体名的字符等内部细节信息,还能通过编码器编码文本内容,互补地学习到变异实体名外部的上下文信息,从而从多方面提高模型的实体识别的能力。

RESULTS : The patient 's GR gene had a heterozygotic mutation ( **G - - > A** ) at nucleotide position **2141** ( exon 8 ) , which resulted in substitution of arginine by glutamine at amino acid position 714 in the ligand - binding domain ( LBD ) of the GR alpha.

图 6 不遵循 HGVS 标准的正样本示例

Matos<sup>[10]</sup>等人在 tmVar 语料上比较了词级和字符级 BiLSTM-CRF 模型,其实验结果显示字符级模型的效果更好, $F$  值为 87.4%,不过仍低于本文实验结果。可能的原因是字符级模型是预测文本序列中每一个字符的标签,没有考虑词的边界和词的上下文等词级别的细节信息,所以即使最终结果优于预测文本序列中每一个词标签的词级模型,但仍有局限性。

CharCNN-CNN-CRF 虽然也属于词级模型,但 CharCNN 层的输出是字符级的词表示,文本编码器通过栈式叠加多层 CNN 能够学到词的上下文信

息,因此该模型既包含词的字符信息,又包含必要的词信息,为标签预测过程提供更多的关键信息,不仅可以识别出结构化变异实体名,还能识别出不遵循标准的变异实体名。

CharCNN-CNN-CRF 模型在两个不同语料上均取得最好的结果,表明该模型具有较好的泛化能力。

## 3 结论

本文提出了一种用于生物医学变异命名实体识

别的基于字符卷积神经网络的 CNN-CRF 模型,实验结果表明,该模型仅使用字符特征和简单的 CNN-CRF 模型就能在 tmVar 和 MutationFinder 语料上取得最好的结果。我们还进一步发现:

(1) 在变异命名实体识别任务上,单独的字符特征比常见实体识别模型中单独的词特征或拼接词特征与字符特征的效果更好,还能缓解未登录词问题。

(2) 基于 N-gram 思想的字符级卷积神经网络通过编码实体内字符信息,抽取对变异命名实体识别更有利的字符特征,从而获得更好的词表示。

(3) 使用 CNN 作为模型的文本编码器比 BiLSTM 效率更高。

(4) CRF 层的使用能极大地降低预测中非法标签序列出现的概率。

目前的方法仅用字符级的词表示,虽包含一定的上下文信息,但未能找到适用于变异实体的直接结合词特征的方法,因此不能充分挖掘和利用上下文词信息,所以如何在词级别上提高实验性能是未来研究的重点。

## 参考文献

- [1] Caporaso J G, Baumgartner W A, Randolph D A, et al. MutationFinder: A high-performance system for extracting point mutation mentions from text [J]. Bioinformatics, 2007, 23(14):1862-1865.
- [2] McDonald R T, Winters R S, Mandel M, et al. VTag: An entity tagger for recognizing acquired genomic variations in cancer literature[J]. Bioinformatics, 2004, 20(17):3249-3251.
- [3] Wei C H, Harris B R, Kao H Y, et al. tmVar: A text mining approach for extracting sequence variants in biomedical literature[J]. Bioinformatics, 2013, 29(11):1433-1439.
- [4] Collobert R, Weston J, Bottou, Léon, et al. Natural language processing (almost) from scratch[J]. Journal of Machine Learning Research, 2011, 12(1):2493-2537.
- [5] Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF models for sequence tagging[J]. arXiv preprint, arXiv: 1508.01991, 2015.
- [6] Lample G, Ballesteros M, Subramanian S, et al. Neural architectures for named entity recognition[J]. arXiv preprint arXiv: 1603.0136/ov3, 2016..
- [7] Ma X, Hovy E. End-to-end sequence labeling via bidirectional LSTM-CNNs-CRF[J]. arXiv preprint arXiv: 1603.0135 4v5, 2016.
- [8] Habibi M, Weber L, Neves M, et al. Deep learning with word embeddings improves biomedical named entity recognition[J]. Bioinformatics, 2017, 33(14):i37-i48.
- [9] Qile Z, Xiaolin L, Ana C, et al. GRAM-CNN: A deep learning approach with local context for named entity recognition in biomedical text [J]. Bioinformatics, 2017,34(9): 1547-1554.
- [10] Matos P, Matos S. Recognition of genetic mutations in text using deep learning[C]//Proceedings of the 1st International Conference on Data Science, E-learning and Information Systems (DATA' 18), 2018, 24:1-4.
- [11] Sang E F T K, Veenstra J. Representing text chunks [C]//Proceedings of the Ninth Conference of the European Chapter of the Association for Computational Linguistics, 1999: 173-179.
- [12] Kim Y. Convolutional neural networks for sentence classification[J]. arXiv preprint arXiv: 1408.5882 v2.
- [13] Viterbi A. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm[J]. IEEE Trans.informat.theory, 1967, 13(2):260-269.
- [14] Prechelt L. Automatic early stopping using cross validation: Quantifying the criteria [J]. Neural Networks, 1998, 11(4):761-767.
- [15] Bird S, Loper E. NLTK: The natural language toolkit[C]//Processings of the Acl Interactive Poster and Demonstration Sessions, 2004: 214-217.
- [16] Labeau M, Allauzen A. Character and subword-based word representation for neural language modeling prediction[C]//Proceedings of the 1st Workshop on Subword and Character Level Models in NLP, 2017: 1-13.



宋雅文(1995—),硕士研究生,主要研究领域为基于深度学习的变异信息抽取技术。  
E-mail: sywdlut@mail.dlut.edu.cn



罗凌(1988—),博士研究生,主要研究领域为基于深度学习的文本挖掘技术。  
E-mail: lingluo@mail.dlut.edu.cn



杨志豪(1973—),通信作者,博士,教授,主要研究领域为自然语言处理、文本挖掘、机器学习、知识图谱构建及其应用。  
E-mail: yangzh@dlut.edu.cn