计算机科学
COMPUTER SCIENCE

基于边界定位与纠偏的中文命名实体提取规则研究

刘盼, 郭延明, 雷军, 老明瑞, 李国辉

引用本文

刘盼, 郭延明, 雷军, 老明瑞, 李国辉基于边界定位与纠偏的中文命名实体提取规则研究[J]. 计算机科学, 2023, 50(3): 276-281.

LIU Pan, GUO Yanming, LEI Jun, LAO Mingrui, LI Guohui. Study on Chinese Named Entity Extraction Rules Based on Boundary Location and Correction [J]. Computer Science, 2023, 50(3): 276-281.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.220200020

基于边界定位与纠偏的中文命名实体提取规则研究

刘　盼[1]　郭延明[1]　雷　军[1]　老明瑞[2]　李国辉[1]
1 国防科技大学系统工程学院　长沙 410000
2 莱顿大学 LIACS 媒体实验室　莱顿 2333CA
(liupan09@nudt.edu.cn)

摘　要　相对于英文天然由单词组成而言,中文由于没有分词符,汉字之间的组词更灵活,在命名实体识别时,其边界更加难以确定。当前的主流方法将命名实体识别任务转化为序列标注任务,文中采用 BIOES 标注方案,针对预测的标签序列进行研究。通过单独比较实体头部标签 B 或尾部标签 E,计算实体边界准确率,结果表明提高边界准确率能够进一步提升实体识别准确率;对具有连续标签的实体边界进行拓展和重定位,采用实体最后一个字符的类型标签对实体类型进行纠偏,利用分词信息对标签不完整的实体进行填充;最后,提出增加边界标记的 BIO$^+$ES 标注方案,用于区分实体边界的非实体字符,以进一步提升中文命名实体识别的性能。

关键词:中文命名实体识别;标注方案;实体提取
中图法分类号　TP391

Study on Chinese Named Entity Extraction Rules Based on Boundary Location and Correction

LIU Pan[1],GUO Yanming[1],LEI Jun[1],LAO Mingrui[2] and LI Guohui[1]
1 College of Systems Engineering,National University of Defense Technology,Changsha 410000,China
2 LIACS Media Lab,Leiden University,Leiden 2333CA,The Netherlands

Abstract　Compared with English text which is naturally composed of words,Chinese text has no word delimiters,so the combination of Chinese characters is more flexible,and it's more difficult to determine the entity boundaries in Chinese named entity recognition(NER). Current mainstream methods transform the NER task into a sequence labeling task. This paper studies the predicted label sequence under the BIOES tag scheme and calculates the entity boundary accuracy by separately considering the entity head label B or tail label E,which shows that increasing the boundary accuracy can further improve the accuracy of entity recognition. We expand the boundaries of entities with continuous labels,use the label type of the last character of the entity to correct the entity type,and use the word segmentation information to fill in the entity with incomplete labels. Finally,this paper proposes a BIO$^+$ES labeling scheme that adds boundary labels to distinguish non-entity characters at entity boundaries and further improves the performance of Chinese NER.

Keywords　Chinese named entity recognition,Tag scheme,Entity extraction

1　引言

命名实体识别(Named Entity Recognition,NER),指从文本中识别出具有特定意义的实体的位置和类型。它不仅仅是独立的信息抽取任务,在信息检索、自动文本概要、问答任务、机器翻译以及知识图谱等自然语言处理应用中也发挥着重要作用。图 1 以"人民日报"数据集中的一句语料为例,命名实体识别的目标就是从"新华社记者刘前刚摄"这个句子中识别出"新华社"这个组织实体(Organization,ORG)和"刘前刚"这个人名实体(Person,PER)。

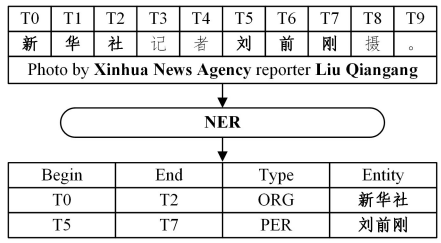

图 1　命名实体识别任务举例

Fig. 1　Example of named entity recognition

在英语中,单词与单词之间都存在空格来标识边界,每个

到稿日期:2022-02-01　返修日期:2022-05-13
基金项目:国家自然科学基金(61806218,71673293);湖南省自然科学基金(2019JJ50722)
This work was supported by the National Natural Science Foundation of China(61806218,71673293) and Natural Science Foundation of Hunan Province,China(2019JJ50722).
通信作者:郭延明(guoyanming@nudt.edu.cn)

单词都具有完整的词义；而中文是以汉字为基本单元，没有明确的词分隔符，模糊的词汇边界会造成大量的边界歧义。有的汉语字符有自己的完整含义，更多的汉语字符则需要与其他字符组成一个词语，而汉字组词的灵活性就增加了中文命名实体边界识别的难度。

当前主流的基于深度学习的中文 NER 模型将 NER 任务转换为序列标注任务，通过特定标注方案对训练集进行标注后送入 NER 模型进行训练。训练后的模型预测需要识别文本中每一个字符的标签，再根据标注方案从预测的标签序列中提取命名实体。

本文针对 Weibo[1] 数据集，采用 BIOES[2] 标注方案，对预测的标签序列进行研究，计算实体边界预测的准确率，通过改进的提取规则来分别提高 NER 的准确率和召回率，进而达到提高 F1 分数的目的。本文的主要贡献如下：

(1)通过单独考虑实体的头部标签 B 或尾部标签 E，即通过判断实体第一个字符或者最后一个字符的位置和类型是否正确，来计算实体边界的准确率。

(2)对具有连续标签的实体边界进行拓展，以提取较长的被嵌套实体，同时采用实体最后一个字符的类型标签对实体类型进行纠偏，以提高实体识别的准确率。

(3)利用分词信息对标签不完整的实体进行填充，找到其缺失的边界，实现不完整标签实体的提取，进一步提升实体识别的召回率。

(4)提出增加边界标记的 BIO$^+$ES 标注方案来增加边界标记，以区分实体附近的汉字与其他非实体汉字，从而更好地利用实体边界信息，进一步提升 NER 的整体性能。

2　相关工作

受益于深度神经网络模型的非线性建模特性，深度学习已被用于命名实体识别系统，性能提升明显。它无须进行复杂特征工程和具备丰富的领域知识，便可以从输入数据中自动地学习错综复杂的隐藏特征表示，扩展性好。近年来，基于深度学习的方法已经全面超越了传统的基于规则的方法和基于统计的命名实体识别方法。

基于深度学习的命名实体识别任务被转换为序列标注任务。标注方案是序列标注任务中一种标记字符序列的方法，通过它可以唯一确定句子中实体的类型和位置。标注方案在训练前将训练集标记为带有标签的字符序列，随后投入 NER 模型进行训练。预测时将未标注的文本投入经过训练的 NER 模型，由模型预测文本的标签序列，然后根据标注方案从预测的标签序列中提取出文本中的实体。

基于标注方案的作用，本文总结了构建标注方案的两个原则：

(1)每个标记的标签由标注方案唯一确定，这意味着标注方案生成的标签序列对于每个句子都是唯一的。

(2)实体可以根据标注方案通过标签序列唯一提取，这意味着在确定每个实体的边界和类型时没有歧义。

Ramshaw 等[3] 将分块任务视为序列标记任务，该方法是将分块结构编码到每个单词的标签中。NP 块是不重叠的、非递归的名词短语。他们将分块任务作为一个标记任务：NP 块内部的单词被标记为 I，NP 块外部的单词被标记为 O，并且一个特殊标签 B 用于紧跟另一个 NP 块之后的 NP 块中的第一个单词。

Ratnaparkhi[4] 的标记方法从左边开始，为每个(词，词性标签)对分配了一个“块”标签，即开始标签、中间标签或其他标签。所有词块起始词都分配一个开始标签(类似于 B 标签)，而词块中的其余词与一个中间标签(类似于 I 标签)配对。这消除了标签歧义，并且无论它们出现在什么上下文中，相同的名词短语都会收到相同的标签序列。

Sang 等[5] 分别将 Ramshaw 和 Ratnaparkhi 的标注方案命名为 IOB1 和 IOB2，并引入了两个新的标注方案，即 IOE1 和 IOE2。在 IOE1 中，紧接另一个 NP 块之前的 NP 块中的最后一个词被分配了 E 标签。而在 IOE2 中，所有 NP 块的最后一个词都被分配一个 E 标签。他们的实验结果表明，IOB1 的表现最好，但与其他标注方案的结果差异不显著。

为了解决日语命名实体识别任务，Uchimoto 等[2] 将每种类型分为 4 个子标签，分别代表一个命名实体的开头、中间和结尾，或者一个命名实体由单个语素组成，这种方案被称为 BIOES 或 BILOU 标注方案。

事实证明，标注方案会影响命名实体识别的性能。Ratinov 等[6] 在两个数据集上对比了 BIO 和 BILOU 标注方案，得出的结论是标注方案的选择对系统性能有很大影响，并且 BILOU 标注方案明显优于 BIO 标注方案。他们认为，BILOU 标注方案只需要少量增加要学习的参数数量就可以学习到具有更好表达力的模型。Tkachenko 等[7] 也得到了类似的结果，他们的实验说明在爱沙尼亚语的命名实体识别方面，BILOU 优于 BIO。Malik 等[8] 使用乌尔都语作为案例进行研究，并为包含后置词的主宾动词语言提出了一种标注方案 Begin Inside Last-2(BIL2)。他们使用隐马尔可夫模型和条件随机场比较了标注方案 IO，BIO2，BILOU 和 BIL2 获得的 F1 分数。他们的结果表明，BIL2 标注方案在乌尔都语中的表现优于其他 3 种标注方案。Reimers 等[9] 比较了 IOB，BIO，BIOES 标注方案，并表明 IOB 方案在命名实体识别任务中的表现比 BIO 和 BIOES 方案差。Yang 等[10] 研究了序列标记模型在多个任务中的应用，并通过命名实体识别数据集的对照实验表明，BIOES 优于 BIO。

这些研究表明，在大多数情况下，BIOES(BILOU)往往比其他标注方案表现得更好，也表明具有更多样化标签的标注方案可能表现更好。本文立足 BIOES 标注方案进行研究，分析提取实体时的边界误差，不再严格按照标注方案进行实体提取，而是通过改进实体提取规则来纠正实体边界偏差，并提取出标签不完整的实体，从而提高实体的提取效果。

3　实体边界误差分析与提取规则改进

3.1　实体边界误差计算与分析

在 BIOES 标注方案中，B(Begin)代表实体的第一个标记，I(Inside)代表实体的内部标记，E(End)表示实体的最后一个标记。如果一个实体仅由一个字符组成，则用 S(单 Singleton)表示。如果字符不属于任何实体(非实体 token)，则用 O(Outside)表示。

基于 BIOES 标注方案，如果只考虑实体的头和尾，即

只要一个实体的第一个字符或者最后一个字符位置预测正确且实体类型正确，则将其提取出来作为候选实体的头部或者尾部，那么就会提取出比头部和尾部同时正确时更多的实体。表 1 列出了从 Weibo 数据集中的一个句子中提取实体的例子，句子中有两个真实标签标注的实体："龙门石窟(LOC)"和"潜溪寺(ORG)"。而根据预测标签，通过 BIOES 标注方案的规则只能提取出"潜溪寺(LOC)"一个实体，通过头部标签 *B* 则能提取"龙(LOC)"和"潜(LOC)"，也就意味着只考虑头部比同时考虑头部和尾部可以多提取出一个候选的实体头部"龙"(龙门石窟的头部)。

表 1 根据头部、尾部或两者一起提取的实体

Table 1 Extract entities based on head, tail or both of them

字符	龙	门	石	窟	潜	溪	寺
真实标签	B-LOC	I-LOC	I-LOC	E-LOC	B-ORG	I-ORG	E-ORG
预测标签	B-LOC	I-LOC	I-LOC	I-LOC	B-LOC	I-LOC	E-LOC
提取规则	头部和尾部(BIOES)				头部(B)	尾部(E)	
真实实体	龙门石窟(LOC) 潜溪寺(ORG)				龙(LOC) 潜(ORG)	窟(LOC) 寺(ORG)	
预测实体	潜溪寺(LOC)				龙(LOC) 潜(LOC)	寺(LOC)	

本文将只考虑单边边界(头部或尾部)时得到的候选实体的准确率、召回率和 F1 值，分别记为 head-P, head-R, head-F1 和 tail-P, tail-R, tail-F1，将同时考虑头部和尾部正常提取的实体准确率、召回率和 F1 值标记为 exact-P, exact-R, exact-F1。由于只考虑单边边界放松了对实体另一侧边界的要求，因此 head-F1 和 tail-F1 的分数会高于 exact-F1。例如，如果实体的第一个字符被正确标记，但最后一个字符被命名实体识别系统标记错误，则在计算 head-F1 时会被正确识别，而在计算 exact-F1 时则会忽略掉。

表 2 列出了一个 NER 模型在 WEIBO 数据集上的头部、尾部和精确指标。对于表 2 中的实体类型，GPE 代表地缘政治实体，LOC 代表地点，ORG 代表组织，PER 代表人名。无论是从各个类型实体的结果还是总体表现来看，由于仅考虑实体的头部或尾部时提取的实体变多，在准确率基本不降的情况下，可以大大提高召回率，从而使得 head-F1 和 tail-F1 的分数高于 exact-F1。

表 2 NER 模型在 WEIBO 数据集上的表现

Table 2 Performances of NER model on WEIBO dataset

类型	头部(B)			尾部(E)			头部和尾部(BIOES)		
	head-P	*head-R*	*head-F*$_1$	*head-P*	*head-R*	*head-F*$_1$	*head-P*	*head-R*	*head-F*$_1$
GPE	61.76	77.78	68.85	63.89	85.19	73.02	57.58	70.37	63.33
LOC	75.00	50.00	60.00	85.71	50.00	63.16	100.00	41.67	58.82
ORG	59.09	50.00	54.17	57.50	44.23	50.00	55.56	28.85	37.97
PER	79.87	82.55	81.19	82.24	83.89	83.06	81.00	75.84	78.34
ALL	75.89	76.86	**76.37**	78.04	77.63	**77.84**	77.03	68.12	**72.31**

因此，本文后面的几项工作就是在现有的基础上，在由标签提取实体时，通过合理确定候选实体的边界，对错误的实体边界进行纠正，并对不完整的标签进行实体提取，一定程度地提高了实体边界的准确率以及实体识别效果。

3.2 实体边界和类别的偏差纠正

嵌套实体指一个实体包含在另一个实体中，如"外婆家"和"外婆"都可以视作一个实体，而"外婆"这个 PER(人名)实体就被嵌套在"外婆家"这个 LOC(地名)实体中。在普通非嵌套的实体识别中，通常只标注较长的被嵌套的实体，而目前在根据序列标注任务结果提取实体时，直接根据标注方案对符合标注方案规则的实体进行提取，这样就容易忽略一些容易被嵌套的较长的实体，从而导致实体边界确定错误。

表 3 列出了 Weibo 数据集上一些实体的预测标签、提取结果和真实结果。可以看到，这些实体在预测时被部分提取，导致实体提取错误。

表 3 部分实体的预测标签及提取结果举例

Table 3 Examples of predicted labels and extraction results of some entities

字符	台	南	市		外	婆	家		加	拿	大
预测标签	B-PER	E-GPE	E-GPE		B-PER	E-PER	E-LOC		B-GPE	B-GPE	E-GPE
提取结果	台南(GPE)				外婆(PER)					拿大(GPE)	
真实标注	台南市(GPE)				外婆家(LOC)				加拿大(GPE)		
字符	湖	南	广	播	电	视	台	广	告	中	心
预测标签	B-GPE	E-GPE	I-ORG	I-ORG	I-ORG	I-ORG	E-ORG	I-ORG	I-ORG	I-ORG	E-ORG
提取结果	湖南(GPE)										
真实标注	湖南广播电视台广告中心(ORG)										

3.2.1 实体边界偏差纠正

本文通过实体周围的标签，重新确定这类实体的边界，提高边界准确率。对于每一个实体尾部 *E*，如果其后一个字符的标签是 *I* 或者 *E*，则将实体尾部延长至下一个 *E* 标签所在的位置；如果其后是 *O*, *S* 或者 *B*，则 *E* 是实体的真实尾部。对于每一个实体头部 *B*，如果其前一个字符的标签是 *I* 或者 *B*，则将实体头部向前延伸至上一个 *B* 标签所在的位置；如果其前面是 *O*, *S* 或者 *E*，则 *B* 是实体的真实头部。

以表 3 中"湖南广播电视台广告中心"的几个字的预测

标签为例，“湖南”两个字的标签已经构成一个实体，按照BIOES标注方案就只能提取出“湖南”这个实体，而后面的标签作废。而采用本文方法时，由于“南”字后面的“广”字标签是 I，因此将“湖南”这个实体的尾部延长到下一个 E，即“台”字所在的位置，构成“湖南广播电视台”这个实体。而“台”字后面的“广”字的标签还是 I，因此将这个实体的尾部继续延长到下一个 E，即“心”字所在位置，从而提取出更完整的“湖南广播电视台广告中心”实体。

3.2.2　实体类别偏差纠正

通过观察，本文认为汉语中实体后面的字往往更能揭示实体的类型，如表3中的“外婆家”中“外”和“婆”两个字的标签类型为PER，“家”字的标签类型为LOC，“外婆”是PER（人名）实体，而“外婆家”却是LOC（地点）实体。因此，区别于英文等其他语言普遍以第一个字符类型作为实体的类型的做法，我们使用实体最后一个字符的标签类型作为实体的类型，这样能够有效地纠正部分实体的类别偏差。

3.3　基于中文分词的不完整标签实体提取

在BIOES标注方案下，除了单字符组成的实体是用 S 标识，其他长度不低于2的实体都是以 B 开头、E 结尾的标签表示。而在实际预测中，存在许多不完整标签，即有的 B 标签后面没有与之对应的 E 标签，有的 E 标签前面没有对应的 B 标签，这些标签未构成完整的 B-E 闭环，精确匹配时不能提取其中的实体，而导致这些标签作废。但当我们只考虑实体头部（B）或尾部（E）时，这些实体的边界是会被检测出来的，而表2中显示head-F1和tail-F1比exact-F1更高，这说明如果能够提取这些边界标签不完整的实体，并正确识别其边界，exact-F1就可以向head-F1和tail-F1逼近，逼近程度取决于边界准确程度。

为了找到缺失的实体边界，本文利用中文分词（Chinese Word Segmentation，CWS）信息，采用最近的分词结果填补缺失的标签，从而确定实体的缺失边界。

表4中，以Weibo数据集中的两句语料为例，利用分词信息，将只有实体尾部标签 E 而没有头部标签的实体，取其前面第一个非实体字符所属的分词结果作为实体的头部，从而找到其缺失的边界，实现对标签不完整的实体的提取，如表4中的“酒店”前第一个非实体字符“级”字所在的分词结果为“四星级”，因此用其作为实体头部，提取出实体“四星级酒店”；“校”字前面第一个非实体字符“驾”字所在的分词结果为“驾校”，则可以提取出实体“驾校”。

表4　用分词结果提取不完整标签实体示例

Table 4　Extract incomplete-label entities with Chinese word segmentation results

字符	圣	诞	节	四	星	**级**	酒	店	自	助
预测标签	O	O	O	O	O	**O**	I-ORG	E-ORG	O	O
分词结果		圣诞节		四星级			酒店		自助	
真实标注	四星级酒店(ORG)									
字符	…	，	**驾**	校	太	给	力	。		
预测标签		O	**O**	E-ORG	O	O	O	O		
分词结果		，	驾校		太	给	力	。		
真实标注	**驾校(ORG)**									

3.4　增加边界标记的BIO$^+$ES标注方案

现有的标注方案都只考虑实体的内部字符，而忽略了实体外字符的信息，对于非实体的字符一律都是统一标注为“O”。这种方式忽略了很重要的信息，即靠近实体的非实体字符能够更好地指示实体的边界位置，但其与其他非实体字符一起被统一标注了，而这种边界信息正是中文NER所需要的。

自从Uchimoto等[2]于2009年提出BIOES标注方案以后，人们对标注方案便少有关注，基本都是采用现有的标注方案。而BIOES由于其标签的多样性获得了比其他标注方案更好的效果，因此本文认为增加标签的多样性能够带来模型效果的提升。同时，由于靠近实体的字符更可能表示实体的边界，为了将这些字符与其他字符区别对待，本文提出了一种改进的标注方案BIO$^+$ES，对于每一个实体，在其实体类型前增加边界标记“O－”和“O＋”，用于标记实体前后的字符，而不是对所有实体外的字符都使用“O”进行标记，这样就可以区分实体附近的汉字与其他非实体汉字的区别。表5列出了以Weibo数据集的一句语料为例，BIO$^+$ES和BIOES标注方案的区别。

表5　BIO$^+$ES和BIOES标注方案的区别

Table 5　Difference between BIO$^+$ES and BIOES labelling schemes

	誓	要	去	台	湾	饮	喜	酒
BIOES	O	O	O	B-GPE	E-GPE	O	O	O
BIO$^+$ES	O	O	O-GPE	B-GPE	E-GPE	O＋GPE	O	O

BIO$^+$ES只对实体外的字符标记进行了改进，在提取实体时的规则与BIOES的规则完全一样，同时增加标签，不需要对命名实体识别模型本身进行改动，这是一种简单有效的改进。

4　实验

4.1　实验设置

本文使用社交领域的公开NER数据集，即Weibo数据集。按照Liu等[11]总结的基于深度学习的中文NER模型的基本框架，中文NER模型由汉字的字符表示、上下文编码器以及标签解码器构成。

字符表示是中文命名实体识别中的研究重点。汉字的字符表示包括字符嵌入和汉字的其他有效表示，当前像BERT[12]这样的大规模预训练语言模型显著增强了命名实体识别的性能，产生了最先进的性能。ERNIE[13]是百度于2019年提出的，旨在学习通过知识屏蔽策略增强的中文语言表示。受BERT掩码策略的启发，ERNIE引入了短语掩码和命名实体掩码，并预测了整个掩码短语或命名实体。与BERT相比，整合短语信息和命名实体信息，使模型能够获得更好的语言表示。实验结果[13]表明，ERNIE在自然语言推理、语义相似性、命名实体识别、情感分析和问答这5个中文自然语言处理任务上取得了最先进的结果。为了验证ERNIE的效果，比较不同预训练语言模型在中文命名实体识别上

的性能，本文分别使用 BERT 和 ERNIE 进行实验。

上下文编码器使用循环神经网络等模型来捕获标记之间的上下文相关性。如果引入字符的其他外部表示（如词性、词根、笔划等外部特征），由于这些外部特征本身没有上下文信息，NER 模型需要上下文编码器来捕获字符的上下文关系。我们选择最常用的双向 LSTM[14]（BILSTM）作为上下文编码器，并通过实验研究仅使用预训练语言模型的情况下 BILSTM 的作用。

标签解码器将上下文表示作为输入，并生成与输入序列对应的标签序列。使用条件随机场（CRF）[15]作为标签解码器，可以在训练过程中逐渐学习到这种标签间的依赖，捕获标签依赖性并约束输出标签序列的顺序，从而避免一些语法错误。例如，在 BIOES 标注方案中，B-PER 后面只能跟 I-PER，E-PER 后面只能跟 O，否则会增大模型的损失，因此 CRF 被用作大多数中文命名实体识别研究中的标签解码器。

为了便于比较，本文开展了多个对比实验，并将各个 NER 模型配置如下：

模型 1　ERNIE＋CRF

模型 2　ERNIE＋BILSTM＋CRF

模型 3　BERT＋CRF

模型 4　BERT＋BILSTM＋CRF

本文将采用实体边界和类别纠正的模型，用“＋Cor”（意指纠正，Correct）表示；同时，采用 jieba[1)] 分词工具获取中文分词信息，并将采用分词结果提取不完整标签实体的模型用“＋CWS”表示。

实验中使用的优化器是 Adam，预训练语言模型只进行微调，学习率设置为 3×10^{-5}，模型其他部分学习率设置为 1×10^{-3}，dropout 参数设置为 0.1。同时，为了避免模型过拟合，本文取非整数轮次（Epoch）的模型结果进行验证（即以固定数量 batch 的间隔对模型进行验证，这个间隔数量非一轮 batch 的整数倍），并保存最佳结果。训练过程中，模型性能会随着随机初始化参数而变化，对于每个模型，本文分别训练 20 个随机种子参数下的表现，取最好的结果。

4.2 实验结果及分析

表 6 列出了本文 4 个模型在 BIOES 和 BIO^+ES 标注方案下，改进实体提取规则后的结果。本文分别从标注方案、预训练语言模型、上下文编码器和实体提取规则 4 个方面讨论实验结果。

表 6　本文模型改进实体提取规则后的 F1 分数

Table 6　F1 scores of the proposed model after improving entity extraction rules

（单位：%）

模型	基础模型	＋Cor	＋CWS	＋Cor＋CWS
模型 1(BIOES)	72.20	73.21	72.97	73.83
模型 1(BIO^+ES)	**73.44**	**75.22**	73.54	**75.00**
模型 2(BIOES)	73.21	73.92	73.71	73.73
模型 2(BIO^+ES)	73.37	74.36	**74.31**	74.82
模型 3(BIOES)	72.47	74.06	73.22	74.07
模型 3(BIO^+ES)	72.68	73.61	74.03	73.84
模型 4(BIOES)	72.34	72.36	72.64	72.68
模型 4(BIO^+ES)	72.34	73.37	72.26	72.98

（1）标注方案。本文将 4 个模型下两种标注方案中 F1 分数较高的画下划线，可以看出，多数情况下，BIO^+ES 优于 BIOES 标注方案，且在不改进实体提取规则的基础模型上，BIO^+ES 全部优于 BIOES 标注方案。在 BIO^+ES 方案中，当使用维特比算法寻找最佳路径时，更可能出现在实体边界的字符得到“O－”或“O＋”标签的观察概率更高，从而能够强化实体边界信息。因为具有更多不同标签，且重视利用实体边界信息，所以本文提出的 BIO^+ES 标注方案是一种更具表现力的标注方案，在中文 NER 模型中表现得更好。

（2）预训练语言模型。模型 1 和模型 3 使用的是 ERNIE，模型 2 和模型 4 使用的是 BERT。模型 1（ERNIE＋CRF）与模型 3（BERT＋CRF）相比，模型 3 在 BIOES 标注方案下优于模型 1，模型 1 在 BIO^+ES 标注方案下优于模型 3，且取得了所有模型中最优的结果。模型 2（ERNIE＋BILSTM＋CRF）与模型 4（BERT＋BILSTM＋CRF）相比，无论是否改进实体提取规则，模型 2 在两种标注方式下都优于模型 4。这说明 ERNIE 在中文 NER 中的表现比 BERT 更好，本文认为 ERNIE 在多源中文语料库上进行预训练，比 BERT 更适用于中文自然语言处理任务。

（3）上下文编码器。模型 1 和模型 3 没有使用上下文编码器，而模型 2 和模型 4 使用 BILSTM 作为上下文编码器。可以看出模型 3 在同标注方案优于模型 4，模型 1 也在大多数情况下优于模型 2。因此添加 BILSTM 这类上下文编码器不能给 NER 带来稳定的提升，本文认为预训练的语言模型可以捕捉汉字的大部分含义，添加上下文编码器在训练时会削弱预训练的编码信息，在不引入其他外部字符表示（如偏旁部首、笔画）的前提下，可以舍弃上下文解码器，直接将标签解码器与预训练语言模型连接。

（4）实体提取规则。采用实体边界和类别纠正（以下简称“＋Cor”）或者用分词结果提取不完整标签实体（以下简称“＋CWS”）都能带来比基础模型更大的改进。在多数情况下，“＋Cor”带来的改进比“＋CWS”更大，而两者结合会比单独的一个改进更大，但存在少数单独使用“＋Cor”或者“＋CWS”比两者共同使用时改进更大的情况，这是因为“＋Cor”或“＋CWS”能为模型带来的改进并不均衡，同时使用两者的情况下 F1 值可能会被改进效果较小的方法拉低，但不会比单独使用“＋Cor”或“＋CWS”都低。

表 7 列出了本文模型和当前最先进的（State of the Art，SOTA）模型在 Weibo 数据集上的 F1 分数。可以看出，本文模型在这个数据集上取得了优于当前 SOTA 模型的结果，使 F1 分数由其他 SOTA 方法最高的 71.81%提高到 75.22%。

表 7　本文模型和 SOTA 模型在 Weibo 上的比较

Table 7　Comparison of the proposed model and SOTA models on Weibo dataset

MODELS	*F1*/%
BERT＋Glyph＋LSTM＋CRF[16]	71.81
BERT＋GLYCE＋Transformer＋CRF[17]	67.60
BERT＋FLAT＋CRF[18]	68.55
SofetLexicon＋BERT＋LSTM＋CRF[19]	70.50
模型 1(BIO^+ES)＋Cor	**75.22**
模型 2(BIO^+ES)＋Cor＋CWS	74.82
模型 3(BIOES)＋Cor＋CWS	74.07
模型 4(BIO^+ES)＋Cor	73.37

1) https://pypi.org/project/jieba/

这些结果表明，使用改进的实体提取规则能够突破当前瓶颈，进一步提升转换为序列标注任务的 NER 任务的性能。

结束语　通过对中文特点的观察来判断实体边界，对中文命名实体识别具有重要意义，因此本文改进现有的实体提取规则。基于 BIOES 标注方案，本文使用实体头部和尾部标签分别计算实体头部和尾部的准确率，分析判断实体边界误差，通过提高实体边界准确率来提高实体识别效果；针对较长被嵌套的实体提取不全的情况，对具有连续标签的实体边界进行拓展，重新定位较远的边界；通过对汉字重心偏后的观察，采用实体最后一个字符的类型标签对实体类型进行纠偏；对于未构成实体的不完整的单边界标签，利用分词信息找到其另一边界，实现对标签不完整的实体的提取；针对现有标注方案未有效利用实体附近字符的边界信息的问题，提出增加边界标记的 $BIO^{+}ES$ 标注方案，用于区分实体边界的非实体字符。

实验中，本文还检验了不同的预训练语言模型和上下文编码器的作用。实验结果表明，改进的标签提取规则和标注方案，能够有效提升 NER 模型在 Weibo 数据集上的性能；ERNIE 在中文数据集上的表现优于 BERT；在不使用外部字符表示的前提下，直接将标签解码器与预训练语言模型连接比在中间加入上下文编码器的效果更好。

下一步，我们考虑加入外部字符表示，提高字符表示效果，同时在更多的中文数据集上验证本文方法的有效性。

参 考 文 献

[1] PENG N, 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. 2015: 548-554.

[2] UCHIMOTO K, MA Q, MURATA M, et al. Named entity extraction based on a maximum entropy model and transformation rules[C]// Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics. 2000: 326-335.

[3] RAMSHAW L A, MARCUS M P. Text chunking using transformation-based learning[M]// Natural Language Processing Using Very Large Corpora. Springer, Dordrecht, 1999: 157-176.

[4] RATNAPARKHI A. Maximum entropy models for natural language ambiguity resolution[D]. Philadelphia: University of Pennsylvania, 1998.

[5] VEENSTRA J, SANG E F T K. Representing Text Chunks [C]// Proceedings of the NinthConference of the European Chapter of the Association for Computational Linguistics (EACL'99). Association for Computational Linguistics, 1999: 173-179.

[6] RATINOV L, ROTH D. Design challenges and misconceptions in named entity recognition[C]// Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009). 2009: 147-155.

[7] TKACHENKO A, PETMANSON T, LAUR S. Named entity recognition in estonian[C]// Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing. 2013: 78-83.

[8] MALIK M K, SARWAR S M. Named entity recognition system for postpositional languages: urdu as a case study[J]. International Journal of Advanced Computer Science and Applications, 2016, 7(10): 141-147.

[9] REIMERS N, GUREVYCH I. Optimal Hyperparameters for Deep LSTM-Networks for Sequence Labeling Tasks[J]. arXiv: 1707.06799, 2017.

[10] YANG J, LIANG S, ZHANG Y. Design Challenges and Misconceptions in Neural Sequence Labeling[C]// Proceedings of the 27th International Conference on Computational Linguistics. 2018: 3879-3889.

[11] LIU P, GUO Y, WANG F, et al. Chinese named entity recognition: The state of the art[J]. Neurocomputing, 2022, 473: 37-53.

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

[13] SUN Y, WANG S, LI Y, et al. ERNIE: Enhanced Representation through Knowledge Integration[J]. arXiv: 1904.09223, 2019.

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

[15] LAFFERTY J D, MCCALLUM A, PEREIRA F C N. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data[C]// ICML. 2001.

[16] SEHANOBISH A, SONG C H. Using Chinese Glyphs for Named Entity Recognition[J]. arXiv: 1909.09922, 2019.

[17] MENG Y, WU W, WANG F, et al. Glyce: glyph-vectors for chinese character representations[C]// Proceedings of the 33rd International Conference on Neural Information Processing Systems. 2019: 2746-2757.

[18] LI X, YAN H, QIU X, 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.

[19] MA R, PENG M, ZHANG Q, et al. Simplify the Usage of Lexicon in Chinese NER[C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 5951-5960.

LIU Pan, born in 1990, postgraduate. His main research interests include natural language processing, computer vision and deep learning.

GUO Yanming, born in 1989, Ph.D, associate professor. His main research interests include computer vision, natural language processing and deep learning.

（责任编辑：喻藜）