

基于全局节点和多片段的格栅命名实体识别

郭江涛, 彭甫榕

(山西大学 大数据科学与产业研究院, 太原 030006)

摘要: 现有命名实体识别模型对标注数据量要求较高, 基于主动学习的命名实体识别模型需要人工分词造成标注代价大。针对上述问题, 提出一种结合全局节点和多片段的格栅命名实体识别模型。将Transformer的全连接结构替换为全局节点和多片段结构, 每个节点仅与构造的上下文向量进行注意力计算, 全局和片段节点分别获取全局和局部信息, 从而降低对标注数据的需求量。对Flat-Lattice结构进行改进, 解决现有主动学习策略需要分词的问题, 从而在保证模型性能的情况下降低数据标注代价。在MSRA、OntoNotes 5.0、Weibo、PeopleDaily这4个数据集上的实验结果表明, 与FLAT模型相比, 所提模型达到对应F1阈值所需的标注数据量分别降低了39.90%、2.17%、34.60%和35.67%。

关键词: 全局节点; 多片段; 格栅; 命名实体识别; 主动学习; Transformer结构

开放科学(资源服务)标志码(OSID):



中文引用格式: 郭江涛, 彭甫榕. 基于全局节点和多片段的格栅命名实体识别[J]. 计算机工程, 2023, 49(12): 96-102.

英文引用格式: GUO J T, PENG F R. Lattice named entity recognition based on global nodes and multi-fragments[J]. Computer Engineering, 2023, 49(12): 96-102.

Lattice Named Entity Recognition Based on Global Nodes and Multi-fragments

GUO Jiangtao, PENG Furong

(Institute of Big Data Science and Industry, Shanxi University, Taiyuan 030006, China)

[Abstract] A lattice Named Entity Recognition (NER) model that combines global nodes and multiple fragments is proposed to address the high demand for annotation data in existing NER models and the high cost of annotation incurred by manual word segmentation in active learning-based NER models. The fully connected structure of the Transformer is replaced with a global node and multi-fragment structure, where each node performs only attention calculations with the constructed context vector. The global and fragmented nodes obtain global and local information separately, thereby reducing the demand for annotated data. The Flat-Lattice structure is improved to solve the problem of word segmentation in existing active learning strategies, thereby reducing the cost of data annotation and ensuring model performance. The experimental results on four datasets, namely MSRA, OntoNotes 5.0, Weibo, and PeopleDaily, show that compared with the FLAT model, the proposed model reduces the amount of annotated data required to reach the corresponding F1 threshold by 39.90%, 2.17%, 34.60%, and 35.67%, respectively.

[Key words] global node; multi-fragment; lattice; Named Entity Recognition (NER); active learning; Transformer structure

DOI: 10.19678/j.issn.1000-3428.0066449

0 概述

命名实体识别(Named Entity Recognition, NER)的目标是标记非结构化文本中的实体词, 并分为对应的类别, 是自然语言处理(Natural Language Processing, NLP)领域的一项重要研究内容。由于语言的特殊性, 中文命名实体识别相比英文更加复杂, 在命名实体识别研究工作中备受关注。

命名实体识别任务的发展初期主要关注人名

(PER)、地名(LOC)、机构名(ORG)等3种专有名词的识别, 随着任务的不断完善, 专有名词的范围被进一步扩大, 逐步发展至对特定领域的特定类别实体进行识别。在生物医学领域, 命名实体识别任务更关注领域专有名词的识别, 如蛋白质、DNA等, 这也就意味着每个特定领域都需要标注新的数据集, 但数据标注的代价是相当大的。研究^[1]表明: 对无标签数据进行准确的标注需要该领域中大量的专家参与, 并且标注数据花费的时间是无标签数据的10倍

基金项目: 国家自然科学基金面上项目(62276162)。

作者简介: 郭江涛(1996—), 男, 硕士, 主研方向为自然语言处理; 彭甫榕(通信作者), 副教授、博士。

收稿日期: 2022-12-05 修回日期: 2023-02-02 E-mail: pengfr@sxu.edu.cn.

以上。

近年来,深度学习在命名实体识别中取得了相当不错的成绩,在标注丰富领域(如新闻)的数据集上表现较好,例如 FLAT^[2]和 SoftLexicon^[3]模型在 MSRA 和 PeopleDaily 上的 F1 值已经超过 0.95。然而现有深度模型相对复杂,需要大量标记数据对模型参数进行迭代训练。在这种情况下,设计一套面向标注稀缺领域的命名实体识别模型具有重要的现实意义和应用价值。

主动学习模仿了生物学习行为的内部过程,将过去学习到的经验推广到新的数据机制。通过对样本进行不确定性排序,可以让专家(人工)集中精力标注模型最不确定的样本。这有助于模型更快地学习,使专家无需标注对模型帮助较小的数据,从而极大地减少所需标注的数据量。

现有基于主动学习的命名实体识别大多基于统计机器学习,并侧重于设计不同的主动学习策略,例如:HUANG 等^[4]基于条件随机场(CRF)^[5]模型设计一种融合 K 均值聚类和信息熵的主动学习策略。基于主动学习的命名实体识别在深度学习方面的研究相对较少,如何将深度学习技术与已有主动学习策略相结合仍是一个挑战。

理论而言,当数据量很小时,深度学习算法会表现不佳,但随着大规模预训练模型的发展,该问题会得到很大改善。SHEN 等^[6]将深度学习技术用于基于主动学习的命名实体识别过程,设计一种 CNN-CNN-LSTM 模型,采用 CNN 作为字编码器和词编码器,使用 LSTM 作为解码器,但 CNN 词编码器需要提前将文本分词,使用自动分词工具容易引入分词错误,而人工分词标注成本更高。Lattice LSTM^[7]提出一种格栅思想,将字符所有可能的词都输入模型,由模型判断最佳序列。FLAT^[2]模型设计了 Flat-Lattice 结构,使得模型可并行计算,有效提高了模型训练速度。

本文提出一种基于全局节点和多片段的格栅命名实体识别模型。将 FLAT 模型中基于全连接注意力机制的 Transformer^[8]改为融合全局节点和片段节点的注意力机制,减少结构复杂度,从而降低对标注数据的需求量。基于全局节点的结构有助于获取全局语义信息,确定词汇边界,片段节点用于捕获局部的语义。同时,对 Flat-Lattice 结构进行改进,可以较好地添加词汇信息而不需要分词。

1 相关工作

1.1 嵌入表示

在自然语言处理中,通过嵌入层将文本处理为可由计算机计算的数值数据。早期的嵌入层使用独热码^[9](One-hot)把语料转化为向量表示,但这种方式得到的特征是高维离散稀疏的向量。之后,出现了 Word2Vec^[10]、GloVe^[11]等静态词向量,静态词向量能够考虑单词的上下文相关词,形成词向量的固定

表征,可以更好地解决词性孤立且不连贯的问题,但存在一词多义现象^[12]。随着预训练模型的发展,Elmo^[13]、GPT^[14]、BERT^[15]等动态词向量技术相继出现,动态词向量技术能够考虑词之间的相关性,还能解决词的多义性问题,从而有效提升最终效果。

1.2 命名实体识别

目前,命名实体识别主要分为基于统计机器学习和基于深度学习。基于统计机器学习的模型依赖于领域专家知识量和人工总结特征,但泛化能力高,可解释性强。基于深度学习的模型使用深度学习自动发现隐藏特征,常结合统计机器学习模型以提升可解释性。近年来,基于深度学习的 NER 模型成为主流,并取得了较好的结果。

Bi-LSTM^[16]是基于深度学习的 NER 模型中比较经典的模型,但 Bi-LSTM 在处理中文 NER 时面临分词问题,基于词编码器的模型^[17]容易引入分词错误,基于字编码器的模型^[18-19]无需分词,但在处理词汇边界问题时精度不高。

Lattice LSTM^[7]对一系列输入字符以及与词典匹配的所有潜在词进行编码,显式地利用了单词和单词序列信息,不存在切分错误,因此本文借鉴这种保留所有潜在词的 Lattice 结构。

SoftLexicon^[3]把字典信息编码到向量表示中来规避复杂的模型结构,提升运算速度,但仍基于 LSTM 串行结构。TENER^[20]通过带位置感知的注意力机制来改进 Transformer Encoder 结构,同时捕捉单词的位置和方向信息,建模词级别和字符级别的上下文信息。FLAT 将 Lattice 结构转换为由跨度组成的平面结构,凭借 Transformer 和融入词汇信息的位置编码,克服了 LSTM 串行结构的缺点,具有出色的并行化能力,但 FLAT 基于全连接的 Transformer,每个跨度之间都要进行注意力计算,这使得有较多的训练参数,导致需要更多的标注数据来训练这些参数。

LGN^[21]设计一种具有全局语义的基于词典的图神经网络,其中词典知识用于连接字符以捕获局部组成,而全局中继节点可以捕获全局语义和长距离依赖。Star Transformer^[22]是一种 Transformer 的轻量级替代方案,将全连接结构替换为星形拓扑,其中每两个非相邻节点通过共享中继节点连接。因此,复杂性从二次型降低到线性型,同时保留捕获局部组成和长期依赖的能力。本文借鉴该思路,对 FLAT 模型进行改进,减少不必要的注意力连接和参数训练,从而降低所需标注数据量。

1.3 命名实体识别中的主动学习策略

大规模获取命名实体识别标注数据的成本很高,主动学习可通过特定策略选择要注释的示例来改善这一问题,以更少的注释获得更高的性能。目前,有关命名实体识别中主动学习研究较少,基于主动学习的 CNN-CNN-LSTM^[6]模型使用 CNN 作为字和词编码器,LSTM 作为解码器,但 CNN 词编码器需要提前将文本分词,使用自动分词工具容易引入分

词错误,人工分词标注成本更高,并且长距离依赖于文本处理具有重要意义,而CNN在获取长距离依赖方面相对欠缺。

综上,当前命名实体识别模型对标注数据的要求较高,基于主动学习的命名实体识别模型还需分词,相对落后。因此,本文综合考虑已有命名实体识别模型和主动学习策略之间的互补性,提出一种基于全局节点和多片段的格栅命名实体识别模型,通过全局节点和多片段结构减少模型对标注数据的需求,利用Flat-Lattice结构解决现有主动学习策略需

要分词的问题,从而在保证模型性能的前提下降低数据标注代价。

2 基于全局节点和多片段的格栅命名实体识别模型

为了降低标注成本,建立一种基于全局节点和多片段的格栅命名实体识别模型,模型结构如图1所示(彩色效果见《计算机工程》官网HTML版)。该模型主要包括嵌入层、全局节点和多片段模块、改进的平面格栅模块、多头注意力层和残差网络模块、CRF输出层等5个部分。

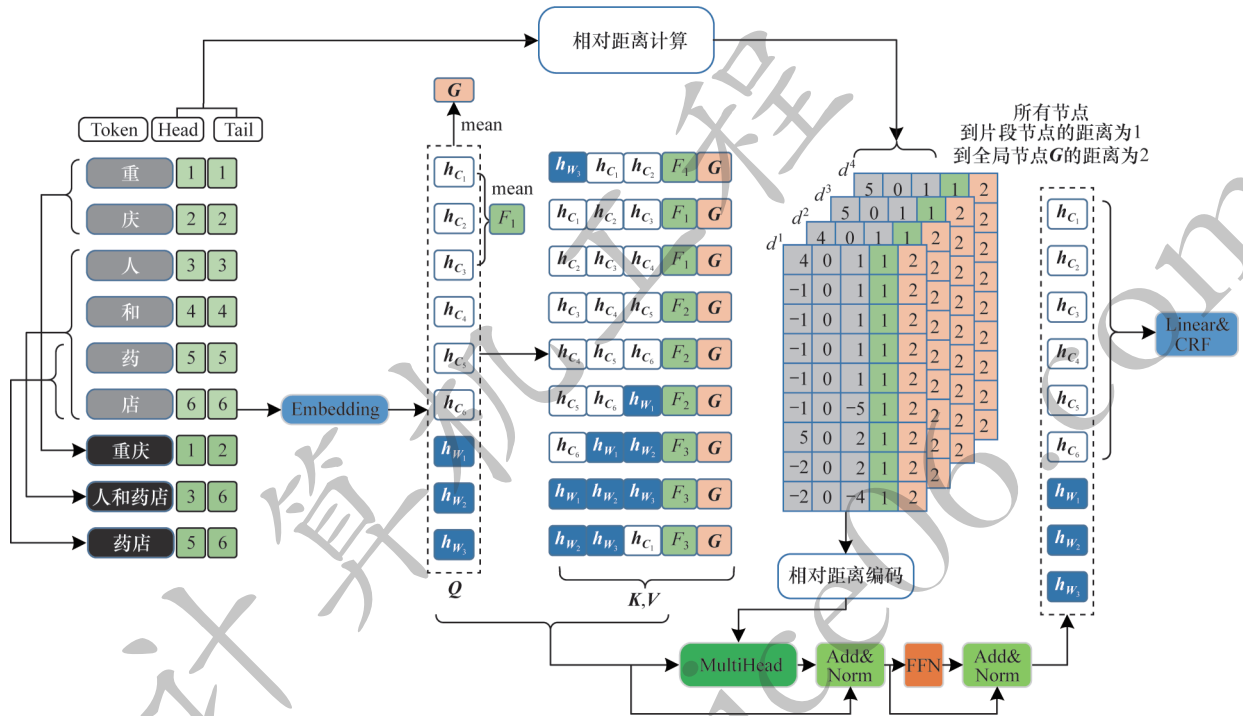


图1 基于全局节点和多片段的格栅命名实体识别模型结构

Fig.1 Structure of lattice named entity recognition model based on global nodes and multi-fragments

2.1 嵌入层

由预处理得到整个句子的字和词后,生成一个跨度序列,一个跨度包括一个标记、一个头部和一个尾部。跨度是一个字或一个词。头和尾表示原始序列中标记的第一个和最后一个字的位置索引,表示标记在格中的位置。对于字而言,头部和尾部是相同的。所有字按照原本的句子顺序位于整个序列前,词按照头部的顺序排在字序列后。将整个序列输入由预训练语言模型BERT组成的嵌入层得到部分跨度的向量表示。

2.2 全局节点和多片段模块

在FLAT模型中,每个字都与整个句子直接进行注意力计算,这使得注意力连接的复杂度为 $O(n^2)$,其中 n 为单个句子长度,需要训练的参数越多,需要标注的数据也越多。为了降低注意力连接的复杂度,借鉴Star-Transformer中的连接方式,但删除了其中用于多层叠加的 e_i 节点,其中 i 代表编码器的层数,因为在命名实体识别任务中仅需一层Transformer编码器。考虑仅靠一个全局节点可能不

能捕获所有非邻居节点间的依赖,设计一种融合全局和局部节点的方案,为每个跨度向量 h_x 构造上下文矩阵 C_x ,上下文矩阵 C_x 的计算方法如式(1)所示:

$$C_x = \text{cat}(h_{x-1}, h_x, h_{x+1}, F_a, G) \quad (1)$$

其中: x 表示跨度在整个序列中的位置,将整个序列视为循环序列,当 $x=0$ 时 $x-1$ 代表跨度序列的最后一个跨度;每个句子平均分为 A 个片段,每个片段拥有一个片段节点 F_a , F_a 的初始值为片段内节点的均值;在全局节点 G 的作用下,每两个不相邻的节点都是两跳邻居,并通过与全局节点 G 的注意力计算获取长距离依赖。在这种结构下,每个字仅与对应上下文矩阵 C_x 进行注意力计算,复杂度降为 $O(n)$ 。

2.3 改进的平面格栅模块

在中文命名实体识别中,词汇信息能够帮助确定实体边界,平面格栅结构可以很好地支持并行计算并添加词汇信息,根据当前模型结构对其进行改进。

相对距离计算方式如式(2)~式(5)所示:

$$d_{xy}^1 = \text{head}[y] - \text{head}[x] \quad (2)$$

$$d_{xy}^2 = \text{head}[y] - \text{tail}[x] \quad (3)$$

$$d_{xy}^3 = \text{tail}[y] - \text{head}[x] \quad (4)$$

$$d_{xy}^4 = \text{tail}[y] - \text{tail}[x] \quad (5)$$

其中: x 代表当前跨度在句子中的下标; y 代表上下文矩阵 C_x 中任一向量 C_{x_y} 在句子中的下标; d_{xy}^1 代表 h_x 的头部和 C_{x_y} 的头部之间的距离; d_{xy}^2 、 d_{xy}^3 和 d_{xy}^4 分别代表头与尾、尾与头、尾与尾之间的距离。在特殊情况下:当 C_{x_y} 代表片段节点 F_a 时, d_{xy}^1 、 d_{xy}^2 、 d_{xy}^3 和 d_{xy}^4 都设定为1,表示所有跨度到全局节点的距离为1;当 C_{x_y} 代表全局节点 G 时, d_{xy}^1 、 d_{xy}^2 、 d_{xy}^3 和 d_{xy}^4 都设定为2,表示所有跨度到全局节点的距离为2。

相对距离编码计算方式与Transformer相同,具体如下:

$$p_d^{(2k)} = \sin(d/10\,000^{2k/d_{\text{model}}}) \quad (6)$$

$$p_d^{(2k+1)} = \cos(d/10\,000^{2k/d_{\text{model}}}) \quad (7)$$

其中: d 为 d_{xy}^1 、 d_{xy}^2 、 d_{xy}^3 或 d_{xy}^4 ; k 表示位置编码的维度索引。

跨度的最终相对距离编码是4个距离的简单非线性变换,具体如下:

$$R_{xy} = \text{ReLU}(W_r(P_{d_{xy}^1} \oplus P_{d_{xy}^2} \oplus P_{d_{xy}^3} \oplus P_{d_{xy}^4})) \quad (8)$$

2.4 多头注意力层和残差网络模块

通过多头注意力层和残差网络将相对距离编码和跨度特征嵌入进行融合计算。Transformer通过 M 个注意力头分别对序列进行自注意力计算,然后将 M 注意力头的结果拼接起来。

将自注意力计算改为将序列中每个跨度向量 h_x 与对应上下文矩阵 C_x 进行注意力计算。为了简单起见,忽略多头注意力层的头部索引,每个头的计算公式如下:

$$\text{Attention}(A, V) = \text{Softmax}(A)V \quad (9)$$

$$A_{xy} = W_q^T h_x^T C_{x_y} W_{k,h} + W_q^T h_x^T R_{xy} W_{k,R} + u^T C_{x_y} W_{k,h} + v^T R_{xy} W_{k,R} \quad (10)$$

$$V = C_{x_y} W_v \quad (11)$$

其中: W_q 、 $W_{k,h}$ 、 $W_{k,R} \in \mathbb{R}^{d_{\text{model}} \times d_{\text{head}}}$ 和 u 、 $v \in \mathbb{R}^{d_{\text{head}}}$ 是可学习的参数。

多头注意力层输出作为残差网络的输入,残差网络与Transformer编码器相同,残差网络输出后舍

去词的嵌入表示,仅将字符表示嵌入CRF^[5]输出层。

2.5 CRF输出层

注意力机制能够对特征进行融合,得到每个跨度的最大概率分布和标签,但标签之间相互独立,无法解决相邻标签之间的合理性问题。CRF层通过在标签之间添加转移矩阵分数排除部分不合理标签序列,从而更容易获得最佳预测序列。

3 实验与结果分析

3.1 数据集

在MSRA^[23]、OntoNotes 5.0^[24]、PeopleDaily、Weibo^[25]等4个中文NER数据集上评估所提模型:

1) MSRA数据集是由微软亚洲研究院发布的一个新闻领域NER数据集,实体类型主要包括LOC、ORG、PER。中文分词在训练集中可用,但在测试集中不可用。

2) OntoNotes 5.0是OntoNotes数据集的最后一个版本,由BBN科技公司、科罗拉多大学、宾夕法尼亚大学和南加州大学信息科学研究所合作构建。该语料库由英语、汉语和阿拉伯语3种语言的文本组成,本文仅使用其中的中文数据集。通过脚本将其处理为命名实体识别专用的BMES格式后,包括数量、日期、事件、语言、法律、位置、金钱、组织、百分比、人名、产品、时间、艺术品等18种粗粒度的实体类型。

3) PeopleDaily:由《人民日报》语料标注而成,并标注了LOC、ORG、PER 3种命名实体,不包含人工标注分词。

4) Weibo数据集:由来自新浪微博的注释NER信息组成。数据集包含PER、LOC、ORG、行政区名(GPE)等粗粒度实体类型,且每种实体都包括特指和泛指两种细粒度实体类型。

采用BMES标记格式,B代表实体的首部,M代表实体中部,E代表实体尾部,O代表非实体。由于BERT输入的最大长度限制为512,为了充分利用标注数据,对数据集进行分段处理,当句子长度大于200时将句子以标点符号为分隔符将句子分为多个片段,标点划分到之后片段的句首。根据数据集规模,按照不同比例划分为训练集、验证集和测试集。处理后的数据集规模信息如表1所示。

表1 数据集规模信息

Table 1 Dataset scale information

指标	MSRA			OntoNotes 5.0			Weibo			PeopleDaily		
	训练集	验证集	测试集	训练集	验证集	测试集	训练集	验证集	测试集	训练集	验证集	测试集
平均句长	48	47	49	32	28	32	54	53	54	46	47	47
最大句长	200	199	200	200	195	197	175	145	146	200	200	200
句子总数/句	40 711	4 557	3 489	36 529	6 093	4 477	1 351	271	271	20 939	2 332	4 657
字总数/字	1 954 433	216 961	172 601	1 197 481	173 288	147 368	73 778	14 509	14 842	979 180	109 870	219 197

3.2 评价指标

采用F1值作为模型精度评价指标,F1值的计算方法如式(12)~式(14)所示:

$$P = \frac{T_p}{T_p + F_p} \quad (12)$$

$$R = \frac{T_p}{T_p + F_N} \quad (13)$$

$$F_1 = \frac{2P \times R}{P + R} = \frac{2T_p}{2T_p + F_p + F_N} \quad (14)$$

其中: P 代表精确率,即预测正确的实体数占有所有预测为实体的数量的比例; R 代表召回率,即预测正确的实体数占有实际为实体的数量的比例; T_p 、 F_p 、 F_N 分别代表预测正确的实体数、预测为实体但实际不为实体的实体数、实际为实体但预测错误的实体数。

本文的目标是在保持精度的情况下降低数据标注代价,因此还需定义一种标注代价方面的评价指标。在命名实体识别中,数据标注代价与标注字数被认为正相关,且在主动学习过程中存在波动,因此设定在主动学习的多轮循环中添加标注数据,当3轮训练结果达到设定对应F1阈值时,标注数据集的总字数更少则表现更好。

3.3 对比实验

3.3.1 实验设置

所有实验都在主动学习框架下完成。将已有训练集的2%初始化为标注数据集,其他放入待标注数据集(即不使用标注,仅使用文本评估可信度),实验中的主动学习策略统一设定为最低可信度策略。每个主动学习循环添加2%训练集的标注数据,每个主动学习循环通过主动学习策略从无标注数据集中选择训练集句子总数2%的句子转到标注数据集(模拟现实中的人工标注)。模型仅使用标注数据集进行训练,默认训练100个迭代(epoch),且25次迭代没有提升精度就结束训练。主动学习算法的伪代码如下:

```
labeled_data, unlabeled_data=get_dataset(train_dataset,
choice_number)//将训练集按比例随机划分为标注数据集和
//未标注数据集,测试集和训练集不变
for i in range(0,stop_echo):
    for epoch in range(1,n_epochs+1):
        model_train()//使用标注数据集训练模型
        model_test()//使用测试集测试模型
        scores = eval_unlabeled_data()//评估未标注数据
        ids =strategy.select_ids(scores)//主动学习根据参数得
//到句子id
        move_to_annotated_dataset(ids)//根据句子id将未标注
//集中对应句子移动到标注数据集,模拟人工标注
```

3.3.2 对比模型

对比模型具体如下:

- 1)基于主动学习的CNN-CNN-LSTM^[6]模型。
- 2)在主动学习框架中复现的FLAT^[2]模型。
- 3)Star_Trans模型:由Star Transformer^[22]和平面格栅结构组合而来,在上下文向量 $C_x = \text{cat}(h_{x-1}, h_x, h_{x+1}, e_i, G)$ 中保留Star Transformer中用于多层叠加的 e_i 节点。

3.3.3 对比实验结果分析

按照实验设置分别对4个中文NER数据集进行实验,各数据集的实验结果如图2所示。由图2可以看出,所提模型(简称为Star_Frag)在MSRA数据集上明显优于其他模型,且两种基于全局节点的模型也都优于FLAT模型;在OntoNotes 5.0数据集上,两种基于全局节点的模型相差不多,略优于FLAT模型,明显优于CNN-CNN-LSTM模型;在Weibo数据集上,Star_Frag明显优于其他模型,且两种基于全局节点的模型也都优于FLAT模型;在PeopleDaily数据集上,3种模型相差不多,在标注字数小于200 000时,FLAT模型相对较差。

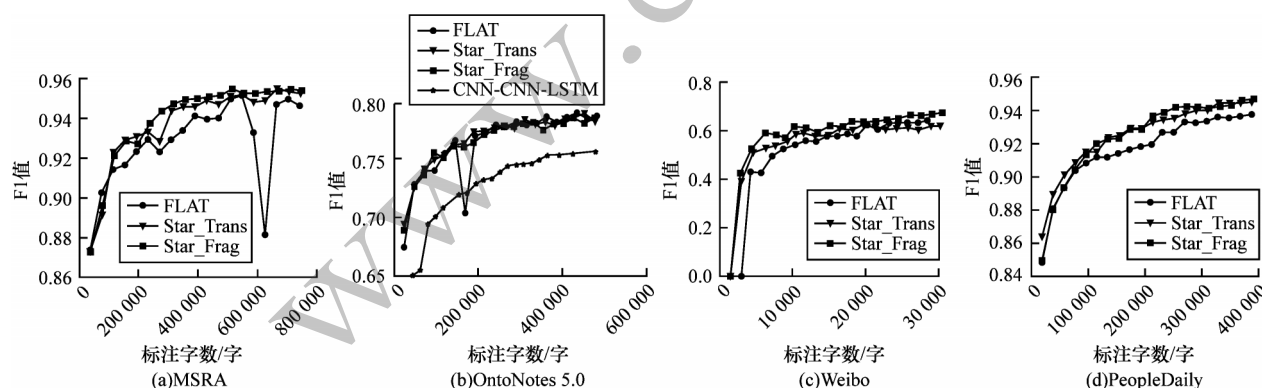


图2 对比实验结果

Fig.2 Comparative experimental results

通过图2分析比较直观,但不够精确。由于数据标注代价与标注字数正相关,并且不同数据集在主动学习过程中存在波动,取模型3次达到F1阈值所需的标注字数为评价指标。F1阈值的取值以使用20%标注数据时模型的平均结果为基准,结果如表2

所示。根据表2中的实验结果可得,Star_Frag模型在4个数据集上所需标注代价更低,相较于FLAT模型,达到对应F1阈值所需的标注数据量分别降低了39.90%、2.17%、34.60%和35.67%,取得了最好的结果。

表 2 3 次达到 F1 阈值所需的标注字数				
Table 2 Required number of annotated words to reach				
F1 threshold three times				
	单位:字			
数据集	MSRA (F1 阈值为 0.945)	OntoNotes 5.0 (F1 阈值为 0.775)	Weibo (F1 阈值为 0.600)	PeopleDaily (F1 阈值为 0.920)
FLAT	663 258	291 448	23 216	269 859
Star_Trans	429 337	286 714	19 688	174 890
Star_Frag	398 617	285 115	15 183	173 589

3.4 消融实验

消融实验的实验设置与对比实验相同。

1)删除片段节点(Star_withoutFa)。删除片段节点 F_a 后的上下文矩阵 C_x 计算公式如下:

$$C_x = \text{cat}(h_{x-1}, h_x, h_{x+1}, G)$$

在相对距离编码计算时,设定当 P 代表 G 时, $d_{xy}^1, d_{xy}^2, d_{xy}^3$ 和 d_{xy}^4 都设定为 1,表示所有跨度到全局节

点的距离为 1。

2)删除全局节点(Star_withoutG)。删除全局节点 G 后的上下文矩阵 C_x 计算公式如下:

$$C_x = \text{cat}(h_{x-1}, h_x, h_{x+1}, F_a)$$

在相对距离编码计算时,设定当 P 代表 F_a 时, $d_{xy}^1, d_{xy}^2, d_{xy}^3$ 和 d_{xy}^4 都设定为 1,表示所有跨度到片段节点的距离为 1。

消融实验结果如图 3 所示,由图 3 可以看出:在 MSRA 数据集上,删除全局节点的模型表现最差,其次是删除片段节点的模型,Star_Frag 模型表现最好;在 OntoNotes 5.0 数据集上,3 种模型相差不大,无法明显区分;在 Weibo 数据集上,删除片段节点的模型表现最差,其次是删除全局节点的模型,Star_Frag 模型表现最好;在 PeopleDaily 数据集上,3 种模型在 200 000 字之前相差不大,在 200 000 字之后 Star_Frag 模型相对更好。

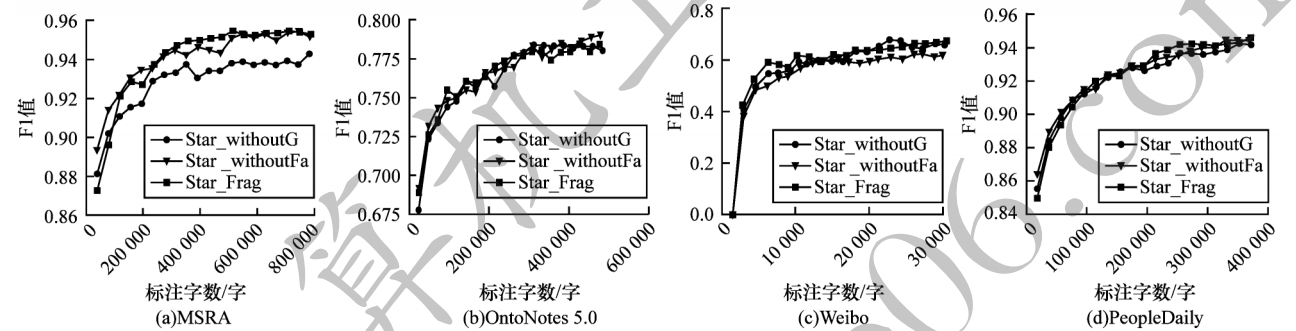


图 3 消融实验结果

Fig.3 Ablation experimental results

3.5 参数分析实验

Transformer 和 Star-Transformer 的编码器都设置了多层叠加结构,但在 FLAT 模型中仅使用 1 层。对此进行参数分析实验,为了节约训练时间,参数分析实验仅进行 10 个主动学习循环,每次添加

100 句句子。实验结果如图 4 所示,由图 4 可以看出,实验结果没有随着编码器叠加层数的增加而变好,在叠加 3 层时 F1 值甚至会出现明显的下降趋势,因此多层叠加编码器在当前结构中不会带来性能提升。

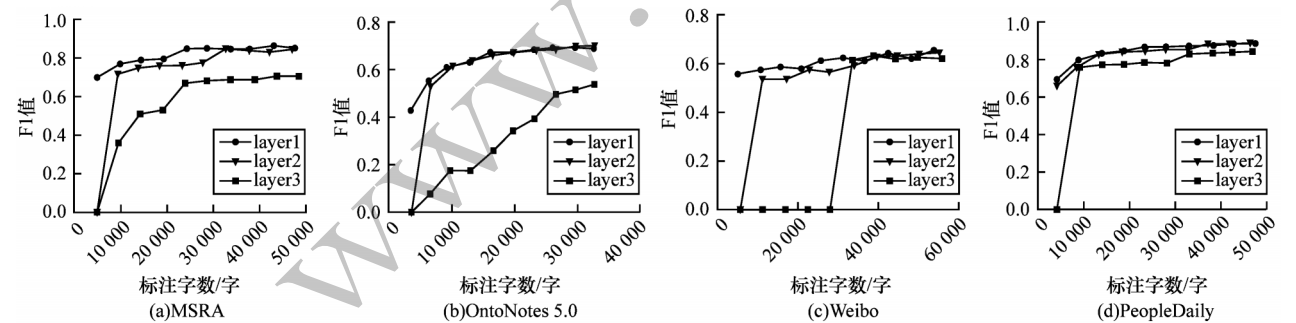


图 4 参数分析实验结果

Fig.4 Parameter analysis experimental results

4 结束语

本文提出一种融合全局节点和片段节点的格栅命名实体识别模型,通过全局节点和片段节点的结构有效减少了标注数据的需求量,对 Flat-Lattice 结构的改进可有效添加词汇信息,避免了人工分词。

对比实验结果表明,所提模型达到对应 F1 阈值所需的标注数据量相比于已有命名实体识别模型更少,有效降低了标注代价。后续将对主动学习策略进行改进,利用 CRF 层的转移矩阵和输出分数评价样本质量,进一步降低标注代价。

参考文献

- [1] 孔玲玲. 面向少量标注数据的中文命名实体识别技术研究[D]. 杭州:浙江大学, 2019.
KONG L L. Research on Chinese named entity recognition technology for a small amount of labeled data [D]. Hangzhou: Zhejiang University, 2019. (in Chinese)
- [2] 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. Stroudsburg, USA: Association for Computational Linguistics, 2020: 1-9.
- [3] PENG M L, MA R T, ZHANG Q, et al. Simplify the usage of Lexicon in Chinese NER [EB/OL]. [2022-11-12]. <https://arxiv.org/abs/1908.05969>.
- [4] HUANG H, WANG H Y, JIN D W. A low-cost named entity recognition research based on active learning [EB/OL]. [2022-11-12]. <https://www.hindawi.com/journals/sp/2018/1890683/>.
- [5] LAFFERTY J, MCCALLUM A, PEREIRA F. Conditional random fields: probabilistic models for segmenting and labeling sequence data [EB/OL]. [2022-11-12]. <https://readpaper.cn/paper/2147880316>.
- [6] SHEN Y Y, YUN H, LIPTON Z C, et al. Deep active learning for named entity recognition [EB/OL]. [2022-11-12]. <https://arxiv.org/abs/1707.05928>.
- [7] ZHANG Y E, YANG J E. Chinese NER using Lattice LSTM [C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, USA: Association for Computational Linguistics, 2018: 1-12.
- [8] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. New York, USA: ACM Press, 2017: 6000-6010.
- [9] 梁杰, 陈嘉豪, 张雪芹, 等. 基于独热编码和卷积神经网络的异常检测[J]. 清华大学学报(自然科学版), 2019, 59(7): 523-529.
LIANG J, CHEN J H, ZHANG X Q, et al. One-hot encoding and convolutional neural network based anomaly detection [J]. Journal of Tsinghua University (Science and Technology), 2019, 59(7): 523-529. (in Chinese)
- [10] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space [EB/OL]. [2022-11-12]. <https://arxiv.org/abs/1301.3781>.
- [11] PENNINGTON J, SOCHER R, MANNING C. GloVe: global vectors for word representation [C]//Proceedings of 2014 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, USA: Association for Computational Linguistics, 2014: 1532-1543.
- [12] 杨长沛, 廖列法. 基于门控空洞卷积特征融合的中文命名实体识别[J]. 计算机工程, 2023, 49(8): 85-95.
YANG C P, LIAO L F. Chinese named entity recognition based on dilated gated convolution feature fusion [J]. Computer Engineering, 2023, 49(8): 85-95. (in Chinese)
- [13] PETERS M E, NEUMANN M, IYYER M, et al. Deep contextualized word representations [EB/OL]. [2022-11-12]. <https://arxiv.org/abs/1802.05365>.
- [14] RADFORD A, NARASIMHAN K. Improving language understanding by generative pre-training [EB/OL]. [2022-11-12]. <https://blog.csdn.net/dawnyang/article/details/121318545>.
- [15] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional Transformers for language understanding [C]//Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language. Minneapolis, USA: Association for Computational Linguistics, 2019: 4171-4186.
- [16] HUANG Z H, XU W, YU K. Bidirectional LSTM-CRF models for sequence tagging [EB/OL]. [2022-11-12]. <https://arxiv.org/abs/1508.01991>.
- [17] PENG N Y, DREDZE M. Improving named entity recognition for Chinese social media with word segmentation representation learning [EB/OL]. [2022-11-12]. https://blog.csdn.net/YW_Vine/article/details/106481816.
- [18] 张应成, 杨洋, 蒋瑞, 等. 基于 BiLSTM-CRF 的商情实体识别模型 [J]. 计算机工程, 2019, 45(5): 308-314.
ZHANG Y C, YANG Y, JIANG R, et al. Commercial intelligence entity recognition model based on BiLSTM-CRF [J]. Computer Engineering, 2019, 45(5): 308-314. (in Chinese)
- [19] 顾亦然, 霍建霖, 杨海根, 等. 基于 BERT 的电机领域中文命名实体识别方法 [J]. 计算机工程, 2021, 47(8): 78-83, 92.
GU Y R, HUO J L, YANG H G, et al. BERT-based Chinese named entity recognition method in motor field [J]. Computer Engineering, 2021, 47(8): 78-83, 92. (in Chinese)
- [20] YAN H, DENG B C, LI X N, et al. TENER: adapting Transformer encoder for named entity recognition [EB/OL]. [2022-11-12]. <https://arxiv.org/abs/1911.04474>.
- [21] GUI T, ZOU Y C, ZHANG Q, et al. A lexicon-based graph neural network for Chinese NER [C]//Proceedings of 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Stroudsburg, USA: Association for Computational Linguistics, 2019: 1-7.
- [22] WU C H, WU F Z, HUANG Y F. DA-Transformer: distance-aware Transformer [C]//Proceedings of 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, USA: Association for Computational Linguistics, 2021: 1-11.
- [23] LEVOW G A. The third international Chinese language processing bakeoff: word segmentation and named entity recognition [EB/OL]. [2022-11-12]. <https://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=012D90A8EE6687762C2CFF89AA6BBC1B?doi=10.1.1.211.3353&rep=rep1&type=pdf>.
- [24] WEISCHDEL R, PALMER M, MARCUS M, et al. OntoNotes release 5.0 [EB/OL]. [2022-11-12]. <https://catalog.ldc.upenn.edu/LDC2013T19>.
- [25] PENG N Y, DREDZE M. Named entity recognition for Chinese social media with jointly trained embeddings [C]//Proceedings of 2015 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, USA: Association for Computational Linguistics, 2015: 1-8.

编辑 陆燕菲