

文章编号: 1003-0077(2021)04-0051-07

基于部分标签数据和经验分布的命名实体识别

宋晔璇¹, 陈 钊^{1,2}, 武 刚^{1,2}

(1. 北京林业大学 信息学院, 北京 100083;

2. 国家林业草原林业智能信息处理工程技术研究中心, 北京 100083)

摘 要: 近年来, 基于数据驱动的命名实体识别方法在新闻、生物医疗等领域上取得了很大的成功, 然而许多领域缺少标签, 且人工标注成本高昂。为了降低标注成本, 该文尝试使用含有噪声的部分标签数据进行命名实体识别, 提出了一种基于部分标签数据和经验分布的方法。首先介绍基于部分标签数据的建模方法, 然后引入标签经验分布的假设, 通过将经验分布加入模型, 有效降低了数据中的噪声。最后分别在植物病虫害数据集和优酷视频数据集上进行测试, 结果表明, 该方法优于其他方法。

关键词: 部分标签数据; 经验分布; 命名实体识别

中图分类号: TP391

文献标识码: A

Named Entity Recognition Based on Partially Labelled Data and Empirical Distribution

SONG Yexuan¹, CHEN Zhao^{1,2}, WU Gang^{1,2}

(1. School of Information, Beijing Forestry University, Beijing 100083, China;

2. Engineering Research Center for Forestry-oriented Intelligent Information Processing of National Forestry and Grassland Administration, Beijing 100083, China)

Abstract: In recent years, data-driven named entity recognition(NER) methods have achieved great success in many fields such as news, biomedical and so on. In order to reduce the cost of labeling for a new domain, a NER method based on partially labelled data and empirical distribution is proposed. We describe the modeling method based on partially labelled data, and then introduce the hypothesis of label empirical distribution. By adding the empirical distribution to the model, the noise in the data is effectively reduced. Tested on the datasets of plant diseases & insect pests and Youku video, and the results show that the proposed method is better than other methods.

Keywords: partially labelled data; empirical distribution; named entity recognition

0 引言

命名实体识别(named entity recognition)指抽取文本中的人名、地名、机构名、领域术语等专有名词, 是自然语言处理领域的基础任务。

近年来, 基于深度神经网络的命名实体识别模型, 在新闻^[1]、医学等领域^[2-3]已经展示出了较高的准确性和鲁棒性, 很大原因是这些领域有着大规模的完整标注的语料。然而, 在许多领域, 数据的标注难度大, 标注成本高, 高质量的标注数据难以获得。

针对特定领域缺少标注数据的问题, 研究人员

探索使用更加容易获得的部分标签数据。通过现有资源如目标领域的词典、网页超链接等, 自动标记出数据中的一部分实体, 得到数据中的部分标签。最早对部分标签数据的研究主要基于条件随机场(conditional random field, CRF)^[4], 结构化感知器和结构化支持向量机^[5]。Tsuboi 等^[6]提出了一种基于部分标签数据训练的 CRF 模型, 能够应用于命名实体消歧^[7]、依存分析^[8]、词性标注^[9]和命名实体识别^[10]等任务。但是, 其仍存在的问题是部分标签数据中有一些字符的标签未知, 这些未知的标签会在训练模型时引入噪声^[11]。一种典型的方法是将未标注的位置直接视为非实体标签, 还有一些方

收稿日期: 2020-03-22 定稿日期: 2020-07-25

基金项目: 自然保护区生态监测系统(2018HXKFXX018)

法^[12]假设未标注位置的标签是均匀分布的。但是这两种方法取得的效果并不好,第一种导致大多数实体无法被识别,降低了召回率;第二种虽然能够召回大量实体,但同时也降低了查准率。

迁移学习的方法是从现有资源的角度出发来解决目标领域中没有标注数据或标注数据少的问题^[13]。在自然语言处理领域,迁移学习主要分为基于资源的迁移和基于模型的迁移。基于资源的迁移对额外资源的规模和质量比较敏感,在跨语言迁移任务中^[14-15]比较常见。基于模型的迁移不需要额外资源,通过对模型结构、损失函数或特征表示等进行调整,利用神经网络的自适应性,在两个不同但有关联的任务间传递知识。一些跨领域的序列标注方法^[16-18]通过把一个在成熟源领域数据上训练的命名实体识别模型的参数迁移到另一个有少量标准的目标领域数据集上,从而提高了目标领域的识别效果,验证了不同领域在处理相同任务时具有底层知识相似性,而且这种底层知识具有可迁移性。

本文提出了一种基于部分标签数据和迁移学习的命名实体识别方法。首先扩展了 Tsuboi 等^[6]的方法,使得模型能够联合部分标签数据和标签的经验分布进行训练。然后将新闻领域的知识迁移到本文的测试领域,在测试领域的部分标签数据上训练得到标签的经验分布。最后,将部分标签数据和标签的经验分布输入模型进行训练。

我们分别在植物病虫害和优酷视频数据集上进行了实验,实验结果表明,本文方法能够有效降低部分标签数据的噪声,且取得较好的识别效果。

1 基础模型

中文命名实体识别是一种序列标注问题。由于中文词边界不容易确定,为了避免分词错误影响模型的识别效果,本文选择对序列中的字符进行标注。在含有全部标签的训练数据中,每个字有且仅有一个与之对应的标签。本文使用三标记(BIO),B表示该字符是实体的起始位置,I表示该字符是实体的中间位置或结束位置,O表示该字符为非实体词。如图1所示,在句子“北京是中国的首都”中,“北京”和“中国”为实体词,其他字符为非实体词,该句子的标注序列为:

$$L = \{B, I, O, B, I, O, O, O\}$$

目前,在中文数据集上,识别效果较好的是 Bi-LSTM-CRF 模型^[1],本文使用 Bi-LSTM-CRF 模型作为基础模型,使用以字作为语义单元,字词信息相结合的方式^[19],具体结构如图2所示。

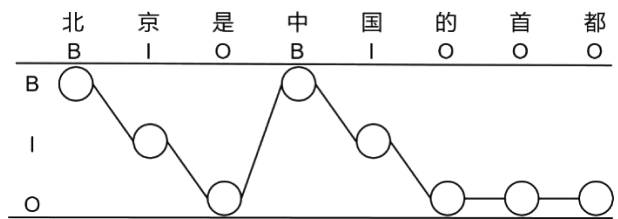


图1 “北京是中国的首都”的标注序列

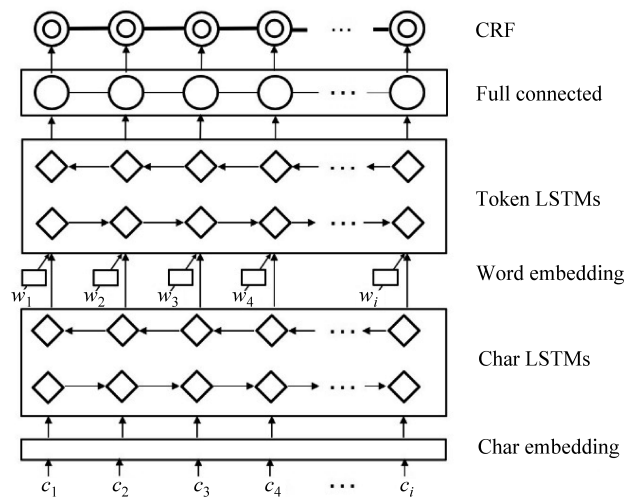


图2 BiLSTM-CRF 网络结构

其中, c_i 表示句子中第 i 个位置的字, w_i 表示该句中该字所在的词,例如,“北”和“京”是所在的词都为“北京”。 c_i 经过字符嵌入层得到字向量 e_i^c , 然后将每个字的字向量输入字符级 LSTM 层, 得到隐层特征 h_i^c 。将 w_i 经过词嵌入层得到的词向量 e_i^w , 将 e_i^c 和 h_i^c 连接得到 e_i 。然后将 e_i 输入标记级 LSTM 层, 能够得到带有全局信息的隐层向量 h_i , 最后经过全连接层计算得到每个位置上的发射得分 A_i , 相关计算方法如式(1)~式(6)所示。

$$e_i^c = C(c_i) \quad (1)$$

$$e_i^w = W(w_i) \quad (2)$$

$$h_i^c = \text{Char_BiLSTM}(h_{i-1}^c, h_{i+1}^c, e_i^c) \quad (3)$$

$$e_i = \text{concat}(h_i^c, e_i^w) \quad (4)$$

$$h_i = \text{Token_BiLSTM}(h_{i-1}, h_{i+1}, e_i) \quad (5)$$

$$A_i = W_i h_i + b_i \quad (6)$$

其中, W_i 和 b_i 是全连接层的权重和偏置。给定完全标注的训练数据 $\{x^j, y^j\}^N$, 其中 N 为样本个数, n 为句子长度。每条样本 x^j 在所有可能的识别结果中只有一条正确的路径 y^j 。根据 Lafferty 等^[4]提出的条件随机场, 模型的损失函数计算如式(7)、式(8)所示。

$$L = - \sum_j^N \log P(y^j | x^j) \quad (7)$$

$$P(y^j | x^j) = \frac{\exp(\text{St}(y^j))}{\exp(\sum_Y \text{St}(Y))} \quad (8)$$

其中, Y 是有可能结果的集合, $\sum_Y \text{St}(Y)$ 是有可能结果得分的总和, $\text{St}(y^j)$ 是正确路径 y^j 的得分。 $\text{St}()$ 被称为得分函数, 计算方法如式(9)所示。

$$\text{St}(y) = \sum_i T_{y_{i-1}, y_i} + \sum_i A_{i, y_i} \quad (9)$$

其中, y_{i-1} 和 y_i 表示第 $i-1$ 个和第 i 个字符的标注结果, T 和 A 分别为转移矩阵和发射矩阵。 T_{y_{i-1}, y_i} 表示从 y_{i-1} 到 y_i 的转移得分, A_{i, y_i} 表示 y_i 的发射得分。

2 基于部分标签数据和经验分布的命名实体识别方法

2.1 基于部分标签数据和标签经验分布的 Partial-CRF

2.1.1 部分标签数据

以图 3 中的句子“北京是中国的首都”为例。在无法获得完整的标注信息的场景下, 只能够确定“中国”是一个地名实体, 由此可以将这两个字符准确地标记为 B, I, 而其他字符的标签是未知的, 可能为 B、I、O 中的任意一个, 这样的数据就是部分标签数据。

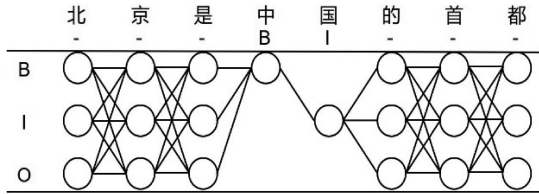


图3 “北京是中国的首都”部分标签数据

在部分标签数据中, 对于每个被观测的句子 x^j , 得到的标注序列不是唯一的, 而是一个由所有满足约束条件的可行序列组成的集合。如句子“北京是中国的首都”在满足“中国”被标记的条件下, 每个字对应的可能标注集合为:

$$P = \{(B, I, O), (B, I, O), (B, I, O), (B), (I), (B, I, O), (B, I, O), (B, I, O)\}$$

设 Y_P 是可行序列的集合, 那么相应的部分标签数据的损失函数就可以粗略计算为:

$$L = - \sum_j \log \sum_{y^j \in Y_P} P(y^j | x^j) \quad (10)$$

2.1.2 标签的经验分布

式(10)在计算损失时, 简单地认为 Y_P 中的每

个标注序列出现的机会是均等的, 也就默认了每个未标注的字符都会被等可能地标记为实体或非实体。然而, 在实践中, 非实体标签往往比实体标签多, 将 Y_P 视为均匀分布会放大噪声数据带来的错误, 导致最终的结果查准率降低。Jie 等^[10]的解决方法是通过自训练的方法预测未标注字符的标签, 从而得到完整标注的训练集。但是, 半监督自训练对于缺少标签或者无标签的样本利用率很低。本文提出了经验分布的假设, 即句子中第 i 个字的标签是 y_i 的概率为 $G_i^{y_i}$, 进而长度为 n 的句子的权重为 $\prod_{i=1}^n G_i^{y_i}$, 设句子“北京是中国的首都”在标签“B, I, O”上的经验分布如图 4 所示, 当标注序列为 {B, I, O, B, I, O, O, O} 时的权重就是“ $0.7 * 0.2 * 0.1 * 1 * 1 * 0.7 * 0.7 * 0.7$ ”。

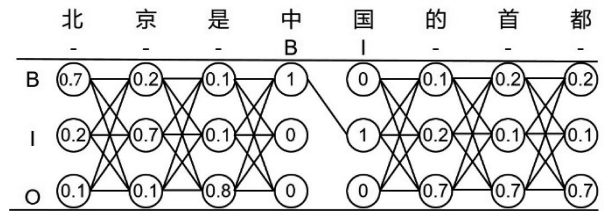


图4 标签的经验分布

在训练模型时, 根据标记序列的权重调整模型参数, 权重越高梯度变化越大, 从而就降低了噪声序列对模型的影响。通过加入标签的经验分布, 使部分标签数据近似于真实数据的分布。进一步提出结合部分标签数据和标签经验分布的 Partial-CRF, 其损失函数表示为式(11)。且当 $\prod_{i=1}^n G_i^{y_i} = 1$ 时, 式(11)与式(10)等价。

$$L = - \sum_j \log \sum_{y^j \in Y_P} (P(y^j | x^j) \prod_{i=1}^n G_i^{y_i}) \quad (11)$$

2.2 基于迁移学习的经验分布的计算方法

本文没有使用统计方法计算概率分布。因为在特定领域缺少标注语料, 很难准确地统计出每个实体出现的频率。此外, 实体类型繁多, 训练数据也无法覆盖所有实体。

本文使用模型迁移的方法计算标签的经验分布。模型迁移的本质是通过迁移模型参数将知识从源领域传输到目标领域。源领域数据一般是来自于公共领域的完整标注数据, 如本文使用了人民日报语料作为源领域的数据。源领域数据包含了丰富的语义和语法信息, 这种底层知识迁移到其他领域同样适用。如连词、助词等在文本中频繁出现的词一

定不是实体,而频繁出现的语义近似的词很有可能是实体。

具体方法是:首先,使用人民日报语料训练源模型,任务目标是预测语料中实体和非实体,并不对实体类别进行区分;然后,将源模型的参数迁移到目标模型,目标模型和迁移模型采用了相同的 Bi-LSTM 结构;最后,将我们的领域数据输入目标模型计算得分,再输入 softmax 层得到概率,就是我们的经验分布。对于已标注的字符,默认其对应标签的概率为 1,其他标签的概率为 0。经验概率 $G_i^{y_i}$ 的计算如式(12)所示,且存在约束式(13)。

$$G_i^{y_i} = \begin{cases} \frac{\exp(A_i)}{\sum \exp(A_i)}, & y_i \text{ is unlabeled} \\ 1, & y_i \text{ is labeled} \end{cases} \quad (12)$$

$$\text{s.t. } \sum_{y_i \in L} G_i^{y_i} = 1 \quad (13)$$

其中, A_i 为全连接层的输出得分, L 为标签集。

2.3 模型训练

本文训练的三个模型都是相同的 Bi-LSTM 结构,分别是用于迁移学习的源模型(source model)、用于计算经验分布的目标模型(target model)和输出最终结果的任务模型(task model)。源模型用 softmax 输出结果,使用交叉熵计算损失。目标模型用 softmax 输出经验分布,然后加入任务模型,共同使用 Partial-CRF 计算模型的损失。源模型和任务模型都使用预训练的词向量和随机初始化的字向量。整个模型框架如图 5 所示。

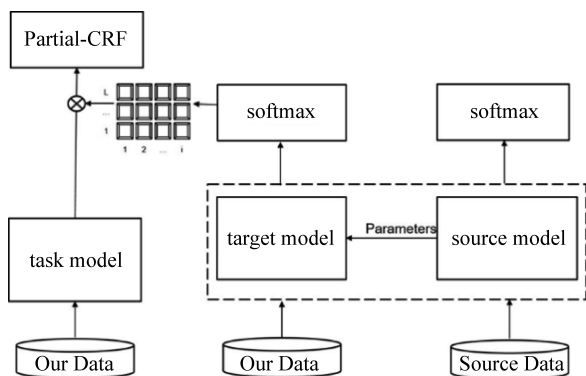


图 5 模型训练框架图

训练顺序如下:

(1) 使用人民日报语料训练源模型,将性能调至最优。然后将参数迁移到目标模型。目标模型的 word embedding 层、char LSTM 层和 token LSTM 层使用源模型的参数,其余层使用随机初始化的参数。

(2) 将我们的领域数据输入目标模型,输出经

验分布;将我们的领域数据输入任务模型,输出 Bi-LSTM 的隐层特征。目标模型和任务模型是同时进行训练的。

(3) 将隐层特征和经验分布共同输入 Partial-CRF 层计算损失,然后进行梯度下降的过程,同时更新目标模型和任务模型的参数,也可以选择固定目标模型的参数。

3 实验与分析

3.1 实验数据

本文使用的含有部分标签的语料来自于植物病虫害领域和视频领域。提到的源领域模型在 1998 年人民日报的标注语料上训练。语料的统计信息如表 1 所示。

表 1 语料统计信息

语料	训练集/句	测试集/句
人民日报	15 573	4 000
优酷视频	8 001	1 001
植物病虫害	14 000	2 000

其中,视频领域的语料来自 Jie 等^[10]发布的公开数据集,他们收集了优酷网页上的标题文本,通过移除语料中的实体标记来获得含有部分标签的语料。我们使用该语料作为训练集,并保留了原本语料中的部分实体标记。考虑到很多领域存在一些专有名词本身就属于实体词或实体词的一部分,有效地利用这些专有名词,同样能够得到标注数据中的部分标签。因此我们还构建了植物病虫害领域的部分标注语料,具体实现方法如下:

语料收集和预处理 我们在《植物保护》《杂草学报》《农药学报》《昆虫学报》中搜集了主题为植物保护和病虫害防治的摘要 5 062 篇,并筛选其中长度为 10 到 100 的句子用于本文语料的构建。然后进行了预处理,包括英文大小写转换和中文简繁体转换等。另外,我们考虑到病虫害领域的名称表示方法不统一,可能使用英文缩写、罗马字符和中文。因此,我们将罗马字符和英文缩写统一转换为中文,例如:

(1) 对农药的单位和剂型进行转换,如将“g a. i./hm²”转换为“克有效成分用量每公顷”,将“SC”转换为“分散剂”;

(2) 对病虫害的品种进行转换,如将“Ⅱ代”转换为“二代”;

(3) 对时间单位进行转换, 如将“y”转换为“年”, 将“M”转换为“月”, 将“d”转换为天。

测试集 人工标注测试集, 标注完成后由专业人员进行校验。另外, 由于林业领域的特点, 有些病虫害实体中包含植物名称, 根据命名实体的标注规则应该对更长的实体进行标记, 如将“水稻卷叶螟”标注为病虫害实体, 就不再对“水稻”单独标注。

训练集 首先收集领域内的词汇构建领域词典, 然后基于词典匹配的方法自动地标记出语料中的部分实体, 从而得到用于训练模型的部分标签数据, 节省了大量的人工成本。本文使用的植物病虫害领域词典包括三类专有名词, 分别来自于中国植物信息网和《农药通用名称 GB 4839-1998》, 词典规模如表 2 所示。

表 2 植物病虫害词典

类型	来源	数量
药剂	《农药通用名称 GB 4839-1998》	804
病虫害	中国林业信息网	1 872
植物	中国林业信息网	3 560

然后使用双向最大匹配算法进行匹配得到部分标签数据, 该算法综合考虑了正向和逆向的匹配结果。

3.2 实验设置

实验中使用随机初始化的字向量和基于 Giga-word 语料库^[19]训练的 100 维的词向量。实验的源模型、目标模型和任务模型采用同样的超参数设置, 训练模型时均使用 Adam 算法更新参数。具体的超参数信息如表 3 所示。

表 3 超参数设置

参数	数值
Batch_Size	20
Embedding_Size	100
Token_LSTMs_Size	128
Char_LSTMs_Size	50
Dropout	0.5
Learning_Rate	0.001

3.3 实验结果及分析

本文采用查准率(Precision, P), 召回率(Recall, R)及 F 值来评估实验结果, 相应的计算如式(14)~式(16)所示。

$$P = \frac{TP}{TP + FP} \times 100\% \quad (14)$$

$$R = \frac{TP}{TP + FN} \times 100\% \quad (15)$$

$$F = \frac{2 \times P \times R}{P + R} \times 100\% \quad (16)$$

TP(true positives)为所有正样本中被正确识别为正样本的数量, FP(false positives)是指负样本被识别成正样本的数量, FN(False Negatives)是指正样本被错误识别为负样本的数量。

为了验证本文提出方法的有效性, 我们设计了五组实验, 具体实验结果如表 4 和表 5 所示。

Dict-Matched: 为了测试实体词典的覆盖程度, 我们在植物领域的数据集上进行了词典匹配的实验。使用领域词典中的实体词对测试集的数据进行匹配, 匹配算法是双向最大匹配;

BiLSTM-CRF: 将部分标签数据中的未被标注的字符标记为“O”, 然后训练基础模型, 最后用该模型在测试集上进行命名实体识别;

Transfer: 用人民日报语料训练源模型, 并将参数迁移到与源模型结构相同的目标模型。用部分标签数据对目标模型进行微调后, 在测试集上进行命名实体识别;

Ours-Frozen: 用本文提出的方法同时训练目标模型和任务模型, 训练过程中保持目标模型的参数不变, 仅优化任务模型的参数, 最后在测试集上进行命名实体识别;

Ours-Finetune: 用本文提出的方法同时训练目标模型和任务模型, 训练过程中使用相同的学习率, 更新两个模型的全部参数, 最后在测试集上进行命名实体识别。

表 4 植物病虫害语料的实验结果

实验	P/%	R/%	F/%
Dict-Matched	88.4	31.7	46.6
BiLSTM-CRF	86.2	50.3	63.5
Transfer	85.6	57.3	68.6
Ours-Frozen	79.6	78.8	79.2
Ours-Finetune	82.4	78.9	80.6

表 5 优酷语料的实验结果

实验	P/%	R/%	F/%
Dict-Matched	—	—	—
BiLSTM-CRF	85.2	48.3	61.7
Transfer	82.6	57.9	68.1
Ours-Frozen	77.4	79.8	78.5
Ours-Finetune	80.3	79.6	79.9

从表 4 的结果可知, Dict-Matched 方法取得了最低的召回率, 这说明尽管领域词典的规模很大, 但仍有很多词覆盖不到, 主要原因有二: 一是由于领域知识更新迭代较快, 出现了很多新词; 二是由于大多数实体词是两个专有名词的组合或嵌套形式, 词典匹配的方法无法识别。在植物病虫害数据集上, BiLSTM-CRF 比 Dict-Matched 方法的召回率提高了 19%, 说明了字向量和词向量相结合的 BiLSTM 能够捕获词边界特征和句子的全局信息。Transfer 方法在两个数据集上的识别效果都有所提升, 说明了不同领域间底层的语义和语法信息是具有共性的, 具有相同目标的不同领域可以通过共享模型参数的方式共享知识。本文提出的方法使查全率和查准率都保持在一个稳定的水平, F 值在两组数据上分别提高了 11% 和 10.8%。我们的方法比迁移学习的方法效果好, 说明经验分布的使用增大了未标记位置被识别的概率, 在一定程度上增大了召回率。对比 Ours-Frozen 方法和 Ours-Finetune 方法, 发现在训练时对目标模型的参数进行微调, 能够提升模型的识别效果。

本文在人民日报语料上预训练的模型查准率为 90.5%, 召回率为 91%, F 值为 90.75%。Ours-Finetune 方法在迁移的过程中损失了将近 10% 的性能, 但仍然比 Transfer 方法的效果要好。说明本文提出的方法也降低了外部语料在部分标签语料上的迁移损失。

我们还在优酷语料和植物病虫害语料上对近年来的方法进行了比较, 实验结果如表 6、表 7 所示。其中, Hedderich 等^[21]在 BiLSTM 模型的基础上添加了噪声适应层, 可以有效地纠正部分标签数据中的错误信息, 但仍需要人工标注少量的数据。Greenberg 等^[12]基于 LSTM 模型和边际条件随机场模型(marginal conditional random field), 把少量部分标签数据和标准数据集一起训练, 取得效果要比单独使用标准数据集更好。Jie 等^[10]同样使用了边际条件随机场模型(Marginal CRF), 他们将部分标签数据分为若干份, 通过自训练迭代模型预测剩

表 6 不同方法在植物病虫害语料上的对比

实验	$P/\%$	$R/\%$	$F/\%$
Hedderich ^[21]	75.4	45.5	56.7
Greenberg ^[12]	20.1	88.3	32.7
Jie ^[10]	75.3	80.4	77.8
Ours	82.4	78.9	80.6

表 7 不同方法在优酷语料上的对比

实验	$P/\%$	$R/\%$	$F/\%$
Hedderich ^[21]	77.3	42.1	54.5
Greenberg ^[12]	17.6	83.7	29.1
Jie ^[10]	78.1	79.6	78.8
Ours	80.3	79.6	79.9

余一份的标签, 经过多次交叉验证得到最终的训练数据。实验证明, 经过自训练得到的数据能够有效提高模型的召回率。

由表 6 和表 7 可以看出, Greenberg^[12]的方法查准率最低, 我们认为边际条件随机场是给模型加入了一个均匀分布, 将未标记的字等可能视为实体或非实体标签, 这样虽然可以最大程度地召回文本中的实体, 但也导致大量的非实体被错判成实体。本文提出方法的 F 值比最好的方法分别提升了 2.8% 和 1.1%, 是因为使用外部知识模拟了标签的经验分布, 在增大了召回率的同时, 也对未标记位置构成约束, 使召回率和查准率达到平衡。

4 结论

命名实体识别作为自然语言处理领域的基础任务, 有着非常重要的作用。目前, 命名实体识别在新闻、生物等领域已经取得了很好的成绩, 但在一些缺乏标注数据的领域, 效果不够理想, 且获得标注数据需要耗费成本。针对这种情况, 本文提出了一种基于部分标签数据的命名实体识别方法。考虑到部分标签数据中含有噪声信息, 会降低模型的识别效果, 本文通过迁移学习计算标签的经验分布, 约束样本在模型训练过程中的权重占比, 从而使部分标签数据更接近于真实数据的分布。在两个数据集上的实验结果表明本文提出的方法是有效的。我们在计算标签的经验分布时仅使用了模型迁移的方法, 所获得的底层知识较少, 在未来的研究中可以尝试更有效的迁移方法。

参考文献

- [1] Lample G, Ballesteros M, Subramanian S, et al. Neural architectures for named entity recognition[J]. arXiv preprint arXiv:1603.01360, 2016.
- [2] Wang Q, Zhou Y, Ruan T, et al. Incorporating dictionaries into deep neural networks for the Chinese clinical named entity recognition[J]. Journal of Bio-medical Informatics, 2019, 92: 103-133.

- [3] Li L, He H, Liu S, et al. Research of word representations on biomedical named entity recognition [J]. Journal of Chinese Computer Systems, 2016, 37(2): 302-307.
- [4] Lafferty J, McCallum A, Pereira F C N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data [C]//Proceedings of the 18th International Conference on Machine Learning. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001: 282-289.
- [5] Tschantz I, Joachims T, Hofmann T, et al. Large margin methods for structured and interdependent output variables [J]. Journal of machine learning research, 2005, 6(9): 1453-1484.
- [6] Tsuboi Y, Kashima H, Mori S, et al. Training conditional random fields using incomplete annotations [C]// Proceedings of the 22nd International Conference on Computational Linguistics. Manchester, UK: CoNLL, 2008: 897-904.
- [7] Hovy D, Hovy E. Exploiting partial annotations with EM training [C]//Proceedings of the NAACL-HLT Workshop on the Induction of Linguistic Structure, 2012: 31-38.
- [8] Flannery D, Miyao Y, Neubig G, et al. A pointwise approach to training dependency parsers from partially annotated corpora [J]. Information and Media Technologies, 2012, 7(4): 1489-1513.
- [9] Mori S, Nakata Y, Neubig G, et al. Pointwise Prediction and Sequence-Based Reranking for Adaptable Part-of-Speech Tagging [C]//Proceedings of the Conference of the Pacific Association for Computational Linguistics. Springer, 2015: 3-17.
- [10] Jie Z, Xie P, Lu W, et al. Better modeling of incomplete annotations for named entity recognition [C]// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019: 729-734.
- [11] Bellare K, McCallum A. Learning extractors from unlabeled text using relevant databases [C]//Proceedings of the 6th international workshop on information integration on the web, 2007.
- [12] Greenberg N, Bansal T, Verga P, et al. Marginal likelihood training of bilstm-crf for biomedical named entity recognition from disjoint label sets [C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018: 2824-2829.
- [13] 庄福振, 罗平, 何清, 等. 迁移学习研究进展 [J]. 软件学报, 2015, 26(1): 26-39.
- [14] Wang M, Manning C D. Cross-lingual projected expectation regularization for weakly supervised learning [J]. Transactions of the Association for Computational Linguistics, 2014, (2): 55-66.
- [15] Zirikly A, Hagiwara M. Cross-lingual transfer of named entity recognizers without parallel corpora [C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, 2015: 390-396.
- [16] Lin B Y, Lu W. Neural adaptation layers for cross-domain named entity recognition [J]. arXiv preprint arXiv:1810.06368, 2018.
- [17] Lee J Y, Dernoncourt F, Szolovits P. Transfer learning for named-entity recognition with neural networks [J]. arXiv preprint arXiv:1705.06273, 2017.
- [18] Yang Z, Salakhutdinov R, Cohen W W. Transfer learning for sequence tagging with hierarchical recurrent networks [J]. arXiv preprint arXiv:1703.06345, 2017.
- [19] 李伟康, 李伟, 吴云芳. 深度学习中汉语字向量和词向量结合方式探究 [J]. 中文信息学报, 2017, 31(06): 140-146.
- [20] Napoles C, Gormley M, Van Durme B. Annotated gigaword [C]//Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction (AKBC-WEKEX), 2012: 95-100.
- [21] Hedderich M A, Klakow D. Training a neural network in a low-resource setting on automatically annotated noisy data [J]. arXiv preprint arXiv: 1807.00745, 2018.



宋晔璇(1995—), 硕士, 主要研究领域为自然语言处理。

E-mail: 1149122136@qq.com



武刚(1962—), 博士, 教授, 主要研究领域为林业信息化。

E-mail: wugang@bjfu.edu.cn



陈钊(1970—), 通信作者, 博士, 副教授, 主要研究领域为智能信息处理、信息系统建模、林业信息化。

E-mail: cz71@bjfu.edu.cn