



文章编号:1001-9081(2019)05-1288-05

DOI:10.11772/j.issn.1001-9081.2018102155

基于深度神经网络的法语命名实体识别模型

严 红¹, 陈兴蜀^{1,2}, 王文贤^{3*}, 王海舟³, 殷明勇³

(1. 四川大学 计算机学院, 成都 610065; 2. 四川大学 网络空间安全学院, 成都 610065;
3. 四川大学 网络空间安全研究院, 成都 610065)
(* 通信作者电子邮箱 catean@scu.edu.cn)

摘要:现有法语命名实体识别(NER)研究中,机器学习模型多使用词的字符形态特征,多语言通用命名实体模型使用字词嵌入代表的语义特征,都没有综合考虑语义、字符形态和语法特征。针对上述不足,设计了一种基于深度神经网络的法语命名实体识别模型 CGC-fr。首先从文本中提取单词的词嵌入、字符嵌入和语法特征向量;然后由卷积神经网络(CNN)从单词的字符嵌入序列中提取单词的字符特征;最后通过双向门控循环神经网络(BiGRU)和条件随机场(CRF)分类器根据词嵌入、字符特征和语法特征向量识别出法语文本中的命名实体。实验中,CGC-fr 在测试集的 F1 值能够达到 82.16%,相对于机器学习模型 NERC-fr、多语言通用的神经网络模型 LSTM-CRF 和 Char attention 模型,分别提升了 5.67、1.79 和 1.06 个百分点。实验结果表明,融合三种特征的 CGC-fr 模型比其他模型更具有优势。

关键词:命名实体识别;法语;深度神经网络;自然语言处理;序列标注

中图分类号:TP391.1 **文献标志码:**A

Recognition model for French named entities based on deep neural network

YAN Hong¹, CHEN Xingshu^{1,2}, WANG Wenxian^{3*}, WANG Haizhou³, YIN Mingyong³

(1. College of Computer Science, Sichuan University, Chengdu Sichuan 610065, China;
2. College of Cybersecurity, Sichuan University, Chengdu Sichuan 610065, China;
3. Cybersecurity Research Institute, Sichuan University, Chengdu Sichuan 610065, China)

Abstract: In the existing French Named Entity Recognition (NER) research, the machine learning models mostly use the character morphological features of words, and the multilingual generic named entity models use the semantic features represented by word embedding, both without taking into account the semantic, character morphological and grammatical features comprehensively. Aiming at this shortcoming, a deep neural network based model CGC-fr was designed to recognize French named entity. Firstly, word embedding, character embedding and grammar feature vector were extracted from the text. Then, character feature was extracted from the character embedding sequence of words by using Convolution Neural Network (CNN). Finally, Bi-directional Gated Recurrent Unit Network (BiGRU) and Conditional Random Field (CRF) were used to label named entities in French text according to word embedding, character feature and grammar feature vector. In the experiments, F1 value of CGC-fr model can reach 82.16% in the test set, which is 5.67 percentage points, 1.79 percentage points and 1.06 percentage points higher than that of NERC-fr, LSTM (Long Short-Term Memory network)-CRF and Char attention models respectively. The experimental results show that CGC-fr model with three features is more advantageous than the others.

Key words: Named Entity Recognition (NER); French; neural network; Natural Language Processing (NLP); sequence labeling

0 引言

命名实体识别(Named Entity Recognition, NER)是指从文本中识别出特定类型事务名称或者符号的过程^[1]。它提取出更具有意义的人名、组织名、地名等,使得后续的自然语言处理任务能根据命名实体进一步获取需要的信息。随着全球化发展,各国之间信息交换日益频繁。相对于中文,外语信息更能影响其他国家对中国的看法,多语言舆情分析应运而

生。法语在非英语的语种中影响力相对较大,其文本是多语种舆情分析中重要目标之一。法语 NER 作为法语文本分析的基础任务,重要性不可忽视。

专门针对法语 NER 进行的研究较少,早期研究主要是基于规则和词典的方法^[2],后来,通常将人工选择的特征输入到机器学习模型来识别出文本中存在的命名实体^[3-7]。Azpeitia 等^[3]提出了 NERC-fr 模型,模型采用最大熵方法来识别法语命名实体,用到的特征包括词后缀、字符窗口、邻近词、

收稿日期:2018-10-26;修回日期:2018-12-26;录用日期:2018-12-26。 基金项目:国家自然科学基金资助项目(61802270);国家“双创”示范基地之变革性技术国际研发转化平台项目(C700011);四川省重点研发项目(2018G20100)。

作者简介:严红(1994—),女,四川广元人,硕士研究生,主要研究方向:命名实体识别、舆情分析; 陈兴蜀(1968—),女,四川自贡人,教授,博士,主要研究方向:网络安全、云计算、大数据安全; 王文贤(1978—),男,福建晋江人,讲师,博士,主要研究方向:网络安全、云计算、大数据安全; 王海舟(1986—),男,四川南充人,讲师,博士,主要研究方向:网络安全、P2P; 殷明勇(1983—),男,陕西汉中人,博士研究生,主要研究方向:舆情分析。



词前缀、单词长度和首字母是否大写等。该方法取得了不错的结果,但可以看出用到的特征多为单词的形态结构特征而非语义特征,缺乏语义特征可能限制了模型的识别准确率。

近几年深度神经网络在自然语言处理领域取得了很好的效果:Hammerton^[8]将长短时记忆网络(Long Short-Term Memory network, LSTM)用于英语 NER; Rei 等^[9]提出了多语言通用的 Char attention 模型,利用 Attention 机制融合词嵌入和字符嵌入,将其作为特征输入到双向长短时记忆网络(Bi-directional Long Short-Term Memory network, BiLSTM)中,得到序列标注产生的命名实体; Lample 等^[10]提出 BiLSTM 后接条件随机场(Conditional Random Field, CRF) 的 LSTM-CRF 模型,它也是多语言通用的,使用了字词嵌入作为特征来识别英语的命名实体,但 LSTM-CRF 模型应用在法语上,和英语差距较大,这个问题可能是因为没有用到该语言的语法特征,毕竟法语语法的复杂程度大幅超过英语。

为了在抽取过程中兼顾语义、字符形态和语法特征,更为准确地抽取法语的命名实体,本文设计了模型 CGC-fr。该模型使用词嵌入表示文本中单词的语义特征,使用卷积神经网络(Convolutional Neural Network, CNN) 提取字符嵌入蕴含的单词字符形态特征以及预先提取的法语语法特征,拼接后输入到双向门控循环网络(Gated Recurrent Unit Neural Network, GRU) 和条件随机场结合的复合网络中,来识别出法语命名实体。CGC-fr 充分利用了这些特征,通过实验证明了每种特征的贡献度,并与其它模型进行比较证明了融合三种特征的 CGC-fr 模型更具有优势。除此之外,本文贡献了一个法语的数据集,包含 1 005 篇文章,29 016 个实体,增加了法语命名实体识别的数据集,使得后续可以有更多的研究不被数据集的问题困扰。

1 法语命名实体识别模型 CGC-fr

本文设计的面向法语命名实体识别的神经网络模型 CGC-fr,其结构如图 1 所示,主要分为三层:文本特征层、上下文特征层和 CRF 层。接下来将从模型的输入开始,由底向上的方式详细介绍它的结构。

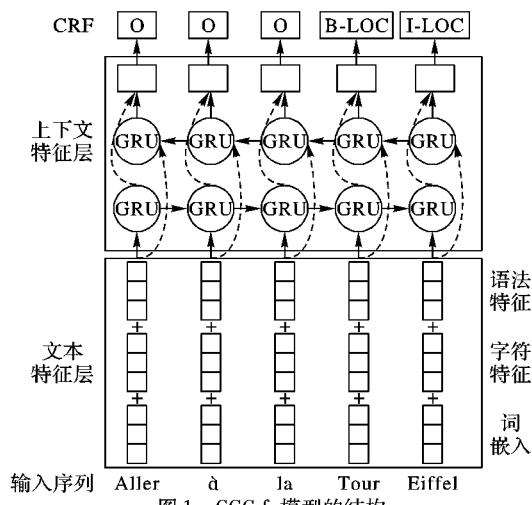


Fig. 1 Structure of CGC-fr model

1.1 文本特征层

在法语 NER 的第一步,本文对经过 tokenize(相当于中文

分词步骤)后的单词序列提取特征。文本特征层的作用在于提取法语文本中的多个特征,将句子表示成包含法语单词字符形态、语义和语法信息的特征序列。

文本特征层作为模型的第一层,和输入层密切相关。输入为一个句子,由 N 个单词 $[w_1, w_2, \dots, w_N]$ 组成。文本特征层把其中每个法语单词转换成一个特征向量 $r = [r^{\text{word}}; r^{\text{char}}; r^{\text{sem}}]$ 。 r^{word} 表示词嵌入,代表单词的语义特征; r^{char} 表示由该词的所有字符嵌入提取得到的特征,代表单词的形态结构信息,比如说词根词缀信息等; r^{sem} 则代表法语语法特征。 r 由三者拼接而成。

1.1.1 语义特征

词嵌入 r^{word} 的表示法和独热编码(one-hot)其实只是相差一个词嵌入矩阵 \mathbf{W}^{word} ,但也正是这个矩阵的存在导致词嵌入比 one-hot 表示法蕴含更多的语义信息。一个词的词嵌入 r^{word} 计算方式为式(1):

$$r_i^{\text{word}} = \mathbf{W}^{\text{word}} \times v_i \quad (1)$$

矩阵 \mathbf{W}^{word} 代表大小为 $|V|$ 的词汇表中所有词嵌入,每列 $\mathbf{W}_i^{\text{word}}$ 代表词汇表中的第 i 个词的词嵌入, v_i 是一个大小为 $|V|$ 的向量,除了词 w_i 所在索引 i 为 1 其余都为 0,也就是一个 one-hot 向量。将词嵌入表示为模型的参数,即可在训练模型时得到。训练的输入输出是词的上下文,所以生成的词嵌入代表词在该语料中的语义信息^[11-12]。模型中直接加载外部已经训练好的词嵌入,已经训练好的词嵌入相较于训练时生成的词嵌入会使得模型的效果更好^[10]。因为已经训练好的词向量,通常由大规模的语料训练生成,出现在各类领域或情景的文本中,更能代表一个词的语义信息。

1.1.2 形态结构特征

字符嵌入和词嵌入的定义类似,给定一个法语单词 w ,这个词的字符经过分割后可表示为字符嵌入序列 $C = [c_1, c_2, \dots, c_M]$ 。词的字符序列不仅难以表达出词的形态特征,还增加了模型的计算复杂度。卷积神经网络(CNN)采用了局部连接和权值共享技术,对局部的特征非常敏感,在提取字符特征方面富有成效^[13],所以本文使用 CNN 提取一个单词的字符嵌入序列所蕴含的字符形态特征。图 2 展示了 CNN 从一个给定法语单词(Bonjour)中提取字符特征的过程,其中〈PAD〉表示填充字符,为了使所有单词长度对齐,方便训练模型。

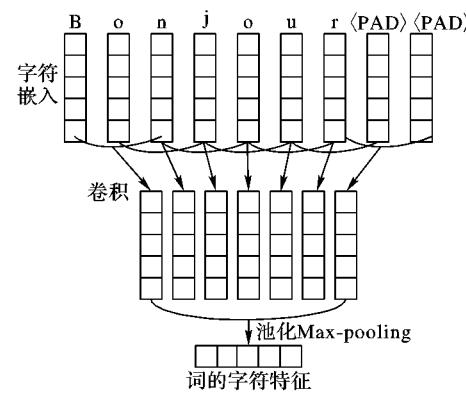


图 2 CNN 提取单词字符特征的过程

Fig. 2 Process of extracting character features of a word by CNN

本文将每个单词表示为包含 M 个字符的字符嵌入序列,作为 CNN 的输入。定义 F 个卷积核,每个卷积核以 k^{char} 大小的窗口在字符嵌入序列上以步长为 1 滑动(选择步长为 1 是



为了不漏过每个可能的词根词缀信息),每次滑动得到一个字符嵌入的子序列,根据式(2)计算得到:

$$\mathbf{p}_i = (\mathbf{c}_i, \mathbf{c}_{i+1}, \dots, \mathbf{c}_{i+k-\text{char}-1})^T \quad (2)$$

再通过池化 Max-pooling 得到全局字符特征 \mathbf{r}^{char} ,这个特征第 j 位元素的计算方式如式(3):

$$r_j^{\text{char}} = \max_{0 < i < (M-k+\text{char})} (\mathbf{W}^p \mathbf{p}_i + \mathbf{b}^p) \quad (3)$$

其中 \mathbf{W}^p 为所有卷积核的权重。

1.1.3 语法规特征

通过观察发现,法语和英语语法上有很多差异。最直观的一点是句子中有非常多的“de”“la”这类词,它们看似和核心单词不沾边,却在法语的命名实体中经常出现。虽然英语中也有冠词例如“the”“an”等,但法语里的冠词都更为复杂,不仅分为定冠词、不定冠词和部分冠词,定冠词还能细分阴性、阳性、复数形式。其位置也不似英语中冠词常出现在命名实体之前,它们经常出现在命名实体的中间,作为命名实体的一部分而存在。在法语中,句子里单词序列的词性和英语中是不同的。作为语法的一部分,法语单词的词性,也有助于从语法的角度来丰富文本的特征,使得后续的过程中能通过具有丰富含义的特征更有效地提取命名实体。因此本文将每个单词的词性特征表示为 one-hot 形式的向量 \mathbf{r}^{sem} ,代表法语单词的语法特征。

词嵌入 \mathbf{r}^{word} 、字符特征 \mathbf{r}^{char} 、语法特征 \mathbf{r}^{sem} 串联起来得到最终的词表示 \mathbf{r} 。文本特征层将一个法语句子表示为特征序列 $\mathbf{S} = [\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_N]$ 。

1.2 上下文特征层

文本的上下文信息往往是双向,当前词语不仅与之前的序列有关还与之后序列有关,比如说猜测一句话中缺少的一个词语,可以从左到右推测,也可从右到左推测,甚至两者结合来看。本文用特征 $\mathbf{S} = [\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_N]$ 代表句子 $[w_1, w_2, \dots, w_N]$ 本身时,通常希望能综合全文上下文信息来识别句子中的每个命名实体。目前的循环神经网络就可以达成这个目标。最开始循环神经网络(Recurrent Neural Network, RNN)被期待能具有记忆功能,保持前文的信息,传递给后面的单元使用,然而它实际表现效果并不好,会遇到梯度消失问题。为了解决 RNN 梯度消失问题而提出的门控循环网络(Gated Recurrent Unit network, GRU)^[14],在传递前文信息的情况下,包含更少的参数,训练更快。BiGRU 则比 GRU 更强大,由正向 GRU 和逆向 GRU 组成,接受上文或者下文传来的信息,综合考虑当前和上下文信息得到输出。其结构在图 1 中可以看到。

文本特征层通过 BiGRU 提取法语文本序列中的上下文信息。BiGRU 中 t 时刻的输入为 \mathbf{r}_t ,输出为 \mathbf{a}_t 。正向 GRU 网络输出的结果为 $\mathbf{g}_t^{\text{left}}$,反向 GRU 网络输出结果为 $\mathbf{g}_t^{\text{right}}$,都由式(4)~(8)计算得来。

$$\mathbf{g}_t = \sigma(\mathbf{W}^g \mathbf{x}_t + \mathbf{U}^g \mathbf{h}_t + \mathbf{b}^g) \quad (4)$$

$$\mathbf{z}_t = \sigma(\mathbf{W}^z \mathbf{x}_t + \mathbf{U}^z \mathbf{h}_{t-1}) \quad (5)$$

$$\mathbf{u}_t = \sigma(\mathbf{W}^u \mathbf{x}_t + \mathbf{U}^u \mathbf{h}_{t-1}) \quad (6)$$

$$\tilde{\mathbf{h}}_t = \tanh(\mathbf{W}^h \mathbf{x}_t + \mathbf{u}_t \mathbf{U}^h \mathbf{h}_{t-1}) \quad (7)$$

$$\mathbf{h}_t = (1 - \mathbf{z}_t) * \tilde{\mathbf{h}}_t + \mathbf{z}_t * \mathbf{h}_{t-1} \quad (8)$$

二者串联起来得到 $\mathbf{a}_t = [\mathbf{g}_t^{\text{left}}; \mathbf{g}_t^{\text{right}}]$, 双向 GRU 网络的输出序列则为 $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_N]$ 。再通过一个线性层,压缩特征向量的维度,得到句子的上下文特征,这也是上下文特征

层的输出 $\mathbf{L} = [l_1, l_2, \dots, l_N]$ 。

1.3 CRF 层

本文把法语命名实体识别看作一个序列标注问题。序列中的每个词都有对应的命名实体的标签,模型只需要预测每个词的标签。命名实体往往是一个词组,因此命名实体的标签不仅标识着类别还标识该词在命名实体中的位置信息。本文使用 BIO2^[15]的方式来表示词的实体类型和在实体中的位置信息。比如说文本序列为“*Aller à la Tour Eiffel*”(去巴菲尔铁塔),其中“*Tour Eiffel*”为地名 LOC 实体,整个句子的实体 BIO2 标签序列为“O, O, O, B-LOC, I-LOC”(O 代表非实体的标签,B-前缀代表实体的第一个词,I-前缀表示实体非头部的词)。在标签序列中,I-LOC 标签后肯定不可后接 I-ORG,所以引入条件随机场(CRF)来学习标签序列间的关系,CRF 能有效地捕获序列内部之间的联系,尤其是序列中前后邻近元素词的关系。定义输入特征序列为 $\mathbf{L} = [l_1, l_2, \dots, l_N]$ 且实际标签序列为 $\mathbf{Y} = [y_1, y_2, \dots, y_N]$ 情况下的条件概率由式(9)计算得到:

$$P(\mathbf{Y} | \mathbf{L}) = \frac{\prod_{i=1}^N \psi_i(y_{i-1}, y_i, \mathbf{L})}{\sum_{y' \in f(L)} \prod_{i=1}^N \psi_i(y_{i-1}', y_i', \mathbf{L})} \quad (9)$$

其中: $\psi_i(y_{i-1}, y_i, \mathbf{L})$ 表示 CRF 的势函数, y_i 表示实际标签序列 \mathbf{Y} 中第 i 个标签, $y' \in f(L)$ 表示预测的标签。训练 CRF 时,用最大似然估计使得条件概率最大化,得到概率最大的标签序列,就预测得到了句子的命名实体标签序列。

2 CSRI_FR_NER 数据集构建

由于缺乏法语命名实体识别数据集,本文通过人工标注的方式构建了一个法语命名实体识别数据集 CSRI_FR_NER (<http://csri.scu.edu.cn/news/308>)。

CSRI_FR_NER 的语料从三大法国新闻网站法国今日报 *Aujourd'hui en France*、法国解放报 *Libération*、法国人道报 *L'Humanité* 上采集而来,我们从其中随机挑选出不同类别共 1005 篇的新闻报道,交由 5 位法语专业的标注者人工标注。数据集包含三类命名实体:人名、地名和组织名。标注规范参照 CONLL2003,工具为 brat。数据集中的文章类别和标注的实体数目如表 1 所示,共 11 种类别的新闻文章,实体的总数达到 29016 个。

表 1 CSRI_FR_NER 内文章类型数目和命名实体类型数目

Tab. 1 The number of article types and the number of named entity types in CSRI_FR_NER

新闻类型	文章数	地名数	组织数	人名数	实体总数
时事	100	1076	696	663	2435
财经	100	758	1112	1514	3384
未来	100	1084	1799	819	3702
科技	3	4	29	8	41
国际	100	1541	1131	940	3612
世界	100	1230	1059	912	3201
自然	100	1316	420	332	2068
政治	99	673	1492	1672	3837
科学	105	629	545	645	1819
生态	98	616	1293	755	2664
交通	100	1086	975	192	2253
合计	1005	10013	10551	8452	29016



3 实验和分析

将实验数据集 CSRI_FR_NER 随机划分为三个部分, 作为训练集、验证集和测试集, 分别有 800、103 和 102 篇文章。其中, 训练集用于训练模型, 验证集防止模型过拟合, 测试集评估模型的效果和泛化能力。

CGC-fr 模型中已训练的词嵌入由 gensim 工具的 skip-gram 模型训练得到, 语料从已采集的新闻报道中随机挑选 55 000 篇。实验的评估度量方式为准确率(Precision, P)、召回率(Recall, R)、F1 值, 对三类命名实体人名、地名、组织名进行评估:

$$P = \frac{\text{该类别预测正确的标签数目}}{\text{该类别预测的总标签数目}}$$

$$R = \frac{\text{该类别预测正确的标签数目}}{\text{数据集中该类别实际标签数目}}$$

$$F1 = \frac{2PR}{P + R}$$

经过进行多次训练, 得到使模型在测试集上表现最优的超参数组合: 词嵌入维度 100, 字符嵌入维度 25, CNN 卷积核数 25, CNN 窗口大小 3, GRU 隐含层维度 100, GRU Dropout 概率 0.2。训练使用 Adam 优化算法, 迭代 50 轮。

本文设计了 2 组实验, 第 1 组实验比较各个特征的贡献度, 第 2 组实验将 CGC-fr 模型与 NERC-fr 模型^[3]、Char attention 模型^[9]和 LSTM-CRF 模型^[10]进行比较。

3.1 模型中各个特征的贡献度

在本组实验中, CGC-fr 模型分为以下几种情况:

1) 只缺少词嵌入的 CGC-fr 模型, 测试词嵌入作为特征对模型的贡献度。

2) 只缺少字符特征的 CGC-fr 模型, 测试字符特征对模型的贡献度。

3) 只缺少语法特征的 CGC-fr 模型, 测试语法特征对模型的贡献度。

将三种情况下的模型和包含所有特征的 CGC-fr 模型进行对比, 表 2 为对比的实验结果。

表 2 CGC-fr 模型特征贡献度的比较

Tab. 2 Contribution comparison of each feature in CGC-fr model

模型	准确率/%	召回率/%	F1 值/%
缺少词嵌入的 CGC-fr 模型	69.86	69.37	69.62
缺少字符特征的 CGC-fr 模型	82.74	72.87	77.49
缺少语法特征的 CGC-fr 模型	81.55	80.82	81.19
CGC-fr 模型	82.23	82.09	82.16

从表 2 中的结果可以看出, 情况 1) 缺乏词嵌入的模型 F1 值下降最多, 达到了 12.54 个百分点, 说明对于整个模型来说, 词嵌入所代表的语义特征最重要。情况 2) 缺乏字符特征也会使模型的 F1 值下降, 为 4.67 个百分点, 虽然准确率仍然很高, 但是召回率下降 9.22 个百分点, 说明缺乏字符特征的情况下, 识别到正确实体数目与测试集中真实实体数目的占比并不高, 有较多识别错误的实体。字符特征对于模型的贡献度排名第二。情况 3) 缺乏语法特征的模型 F1 值下降 0.97 个百分点, 说明语法特征也有一定贡献度, 但相比前两者的小

很多。最后, 包含三个特征的情况下, CGC-fr 模型的效果达到最好, 验证了这三个分别包含语义、字符形态和语法信息的特征有效性。

3.2 与其他两种模型对比

本文选择了文献[3]中的最大熵模型 NERC-fr、文献[9]中的 Char attention 模型和文献[10]中的 LSTM-CRF 模型进行对比。NERC-fr 模型是使用了 8 个特征的最大熵模型。Char attention 模型使用 Attention 机制融合词嵌入和字符嵌入的 BiLSTM 网络进行识别。LSTM-CRF 模型使用字词嵌入作为特征的 BiLSTM-CRF 识别命名实体。根据文献[3]公开的模型参数设置和源码训练 NERC-fr 模型。Char attention 模型与 LSTM-CRF 模型则在超参数方面和本文保持一致, 使用相同的已训练词嵌入。实验结果如表 3 所示。

表 3 三种模型在测试集上的评估结果比较

Tab. 3 Comparison of evaluation results of three models on test set

方法	实体类别	准确率/%	召回率/%	F1 值/%
NERC-fr ^[3]	人名	85.41	70.88	76.49
	地名	79.11	76.15	77.66
	组织名	80.14	70.00	74.73
	平均	81.29	72.23	76.49
LSTM-CRF ^[10]	人名	90.67	84.20	87.32
	地名	76.12	83.40	79.59
	组织名	80.68	70.13	75.03
	平均	82.08	78.74	80.37
Char attention ^[9]	人名	90.21	84.51	87.26
	地名	79.09	83.60	81.29
	组织名	78.56	72.99	75.67
	平均	82.28	79.96	81.10
CGC-fr	人名	87.64	88.59	88.12
	地名	81.21	84.61	82.87
	组织名	78.50	74.28	76.33
	平均	82.23	82.09	82.16

由实验结果可以发现, CGC-fr 模型效果优于 NERC-fr 模型, F1 值在人名上提高 11.63 个百分点, 在地名上提高 5.21 个百分点, 在组织名上提高 1.6 个百分点, 平均提高 5.67 个百分点。NERC-fr 模型在三个实体类别上的准确率和本文模型差距很小, 甚至组织名准确率高 1.64 个百分点, 但所有类型召回率普遍比本文模型要低。说明 CGC-fr 模型识别正确的实体数目对数据集中实际实体总数的占比更高, 这是它识别效果 F1 值整体效果优于 NERC-fr 模型的原因。NERC-fr 模型从文本中提取 8 个特征, 而且大多是单词的形态特征, 相对于 CGC-fr 模型少了语义方面的特征, 可能是造成其效果不如 CGC-fr 模型的原因。

CGC-fr 模型效果稍优于 Char attention 模型, 平均 F1 值高 1.06 个百分点, Char attention 模型比 CGC-fr 模型少了语法特征, 并且两者的字符特征提取方式也不同。Attention 机制融合字词嵌入比 LSTM-CRF 表现稍好, 说明字符特征的提取方式对命名实体的识别效果有一定影响。

CGC-fr 模型效果稍优于 LSTM-CRF 模型, 平均 F1 值高 1.79 个百分点, 两者较为不同的地方在于 CGC-fr 模型的字符特征的提取方式和语法特征, 说明这两种特征都对法语命名



实体具有一定的影响力。

而三个模型的共同之处在于人名都普遍高于另外两种类型的命名实体,推测原因可能在于人名相对来说比较短,规律性要强一些。而组织名识别率都低一些,推测组织名包含缩写、全称、简称等形式,长度不一,变化性较强。

4 结语

本文设计了用于法语命名实体识别的深度神经网络CGC-fr模型,并构建了一个法语命名实体识别数据集。CGC-fr模型将法语文本中单词的词嵌入作为语义特征,从单词对应的字符嵌入序列提取单词的形态结构特征,结合语法特征完成对命名实体的识别。这增加了传统统计机器学习方法中特征的多样性,丰富了特征的内涵,也避免了多语言通用方法对法语语法的忽视。实验对比模型中各个特征的贡献度,验证了它们的有效性;还将CGC-fr模型与最大熵模型NERC-fr、多语言通用模型Char attention和LSTM-CRF对比。实验结果表明,CGC-fr模型相对三者的F1值都有提高,验证了融合三种特征的本文模型在法语命名实体识别上的有效性,进一步提高了法语命名实体的识别率。

然而,本文模型也存在着不足,在法语文本中组织名的识别率相比其余两种命名实体类型差距较大,模型对形式存在较大变化的命名实体类型的识别效果不是很好;其次,相对于英语较高的命名实体识别准确率,法语命名实体识别还有较大的提升空间。

参考文献 (References)

- [1] NADEAU D, SEKINE S. A survey of named entity recognition and classification[J]. *Lingvisticae Investigationes*, 2007, 30(1): 3 – 26.
- [2] WOLINSKI F, VICHOT F, DILLET B. Automatic processing of proper names in texts[C]// Proceedings of the 7th Conference on European Chapter of the Association for Computational Linguistics. San Francisco, CA: Morgan Kaufmann Publishers, 1995: 23 – 30.
- [3] AZPEITIA A, CUDADROS M, GAINES S, et al. NERC-fr: supervised named entity recognition for French[C]// TSD 2014: Proceedings of the 2014 International Conference on Text, Speech and Dialogue. Berlin: Springer, 2014: 158 – 165.
- [4] POIBEAU T. The multilingual named entity recognition framework [C]// Proceedings of the 10th Conference on European Chapter of the Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2003: 155 – 158.
- [5] PETASIS G, VICHOT F, WOLINSKI F, et al. Using machine learning to maintain rule-based named-entity recognition and classification systems[C]// Proceedings of the 39th Annual Meeting on Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2001: 426 – 433.
- [6] WU D, NGAI G, CARPUAT M. A stacked, voted, stacked model for named entity recognition[C]// Proceedings of the 7th Conference on Natural Language Learning at HLT. Stroudsburg, PA: Association for Computational Linguistics, 2003: 200 – 203.
- [7] NOTHMAN J, RINGLAND N, RADFORD W, et al. Learning multilingual named entity recognition from Wikipedia[J]. *Artificial Intelligence*, 2013, 194: 151 – 175.
- [8] HAMMERTON J. Named entity recognition with long short-term memory[C]// Proceedings of the 7th Conference on Natural Language Learning at HLT. Stroudsburg, PA: Association for Computational Linguistics, 2003: 172 – 175.
- [9] REI M, CRICHTON G, PYYSALO S. Attending to characters in neural sequence labeling models[J/OL]. arXiv Preprint, 2016, 2016: arXiv: 1611. 04361[2016-11-14]. <https://arxiv.org/abs/1611.04361>.
- [10] 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. Stroudsburg, PA: Association for Computational Linguistics, 2016: 260 – 270.
- [11] LE Q, MIKOLOV T. Distributed representations of sentences and documents[C]// Proceedings of the 31st International Conference on Machine Learning. New York: JMLR. org, 2014: 1188 – 1196.
- [12] PENNINGTON J, SOCHER R, MANNING C. Glove: global vectors for word representation[C]// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA: Association for Computational Linguistics, 2014: 1532 – 1543.
- [13] SANTOS C D, ZADROZNY B. Learning character-level representations for part-of-speech tagging[C]// Proceedings of the 31st International Conference on Machine Learning. New York: JMLR. org, 2014: 1818 – 1826.
- [14] CHO K, van MERRIENBOER B, GULCEHRE C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[C]// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA: Association for Computational Linguistics, 2014: 1724 – 1734.
- [15] SANG E F, VEENSTRA J. Representing text chunks[C]// Proceedings of the 9th Conference on European Chapter of the Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 1999: 173 – 179.

This work is partially supported by National Natural Science Foundation of China (61802270), the Transformative Technology International R&D and Transformation Platform of the National “Double Creation” Demonstration Base (C700011), the Sichuan Key Research and Development Project (2018G20100).

YAN Hong, born in 1994, M. S. candidate. Her research interests include named entity recognition, public opinion analysis.

CHEN Xingshu, born in 1968, Ph. D., professor. Her research interests include network security, cloud computing, big data security.

WANG Wenxian, born in 1978, Ph. D., lecturer. His research interests include network security, cloud computing, big data security.

WANG Haizhou, born in 1986, Ph. D., lecturer. His research interests include network security, P2P.

YIN Mingyong, born in 1983, Ph. D. candidate. His research interest includes public opinion analysis.