

基于注意力机制的命名实体识别模型研究

——以军事文本为例

单义栋¹ 王衡军¹ 黄 河² 闫 倩³

(解放军信息工程大学三院 郑州 450001)¹ (61660 部队 北京 100000)²
(山东省军区 济南 250000)³

摘 要 针对双向长短时记忆网络模型提取特征不充分的特点,将字向量和词向量同时作为双向长短时记忆网络的输入,并利用注意力机制分别提取两者对当前输出有用的特征,用维特比算法约束最终输出的标签序列,构建一种新的命名实体识别模型。实验结果表明,在军事文本的命名实体识别中,该模型取得了较优的识别率。

关键词 注意力机制,字向量,词向量

中图法分类号 TP391 **文献标识码** A

Study on Named Entity Recognition Model Based on Attention Mechanism

——Taking Military Text as Example

SHAN Yi-dong¹ WANG Heng-jun¹ HUANG He² YAN Qian³

(The Third Institute, PLA Information Engineering University, Zhengzhou 450001, China)¹
(61660 Army, Beijing 100000, China)² (Shandong Military District, Jinan 250000, China)³

Abstract Due to the insufficiency of extracting features by bi-directional long-short term memory network model, the character vector and the word vector are used as the input and the attention mechanism is used to extract the features that are useful for the current output. In this paper, a new named entity recognition model was constructed by constraining the final output tag sequence with the Viterbi algorithm. The experimental results show that the model has achieved a better recognition rate in the identification of military texts.

Keywords Attention mechanism, Character vector, Word vector

1 引言

近年来,自然语言处理技术飞速发展,在日常生活的各个领域得到了广泛应用。尤其是在军事领域,美国等发达国家不惜投入大量人力、物力研发自然语言处理系统,并将其应用于军事装备及军事行动中。在未来作战中,战场形势瞬息万变,战机稍纵即逝,对于交战双方而言,在最短时间内准确处理大量作战命令等军事文本信息,将是赢得战争主动的关键环节。命名实体识别是自然语言理解词性分析的关键技术之一,通过自动识别敌对双方文本中人名、地名、组织机构等实体信息,可为后续语义分析奠定基础,提升情报侦察、指挥决策、组织实施等行动的效率。

目前,基于统计的学习方法广泛应用于命名实体识别研究。基于统计的学习方法主要包括支持向量机(Support Vector Machine, SVM)、隐马尔可夫模型(Hidden Markov Model, HMM)、条件随机场(Conditional Random Fields, CRF)等,它们需要人工提取大量特征,包括单词特征、上下文特征、语义特征等,由于这些特征是针对特定问题,因此可以取得很好的命名实体识别效果。俞鸿魁等^[1]提出一种基于层叠隐马尔可夫模型的命名实体识别模型,它先利用底层隐马尔可夫模型识别简单的人名、地名,而后利用高层隐马尔可夫

模型识别嵌套的人名、地名。胡文博等^[2]提出基于多层条件随机场的命名实体识别模型,其原理与俞鸿魁等提出的模型原理类似,利用条件随机场模型取代了隐马尔可夫模型,两种模型都取得了较好的命名实体识别效果。然而这种统计的方法也存在很大的缺陷,命名实体识别时需要大量人工设计的特征,代价很大,同时由于不同的任务选择的人工特征不尽相同,导致其通用性不佳,泛化能力差。

近年来,基于神经网络的深度学习方法逐渐得到重视,相比于统计的方法,它无需人工干预,可以自动获取特征,在命名实体识别任务中取得了很好的效果。Passos 等^[3]从如何改进词向量方面入手,提出了词汇注入的方法来扩展 Skip-Gram 模型,实验表明,这种方法在命名实体识别方面取得了较好的结果。Chiu 等^[4]提出将双向长短时记忆网络(Bi-directional Long-Short Term Memory network, BLSTM)和卷积神经网络(Convolutional Neural Network, CNN)相结合,应用于命名实体识别,与 Collobert 等^[5]提出的模型相比,该模型利用双向长短时记忆模型打破固定窗口大小的限制,同时利用 CNN 获取诸如前缀和后缀的明确的字符级特征。这两项改进极大地提高了命名实体识别的效果。文献^[6]则提出利用双向长短时记忆网络模型来进行命名实体识别,同时以基于上下文的词向量和基于字的词向量作为模型输入,提高了

单义栋(1988—),男,硕士,主要研究方向为自然语言处理;王衡军(1973—),博士,副教授,主要研究方向为神经网络、机器学习, E-mail: 2793730105@qq.com(通信作者);黄 河(1982—),硕士,主要研究方向为神经网络;闫 倩(1985—),主要研究方向为机器学习。

命名实体识别的精确率和 F 值。文献[7]以双向长短时记忆网络模型为基线系统,利用切片段整体分配标记取代字符分配标记,取得了很好的命名实体识别效果。然而,在命名实体识别过程中,深度学习的方法并没有考虑到不同的词语会对命名实体识别结果产生不同程度的影响,因此识别准确率有待进一步提高。

近几年,注意力机制被广泛应用于图像识别^[8]、自然语言理解^[9-11]等领域,它模仿人类的注意力机制,将有限的资源集中于需要重点关注的信息,减少了对无用信息的关注。在命名实体识别任务中,注意力机制则是根据各个词语对目标词语的重要程度,分配各词语的权重,在一定程度上弥补了深度学习的不足。

针对军事文本的命名实体识别,本文主要做了以下工作:
(1)构建军事语料库,准确标注命名实体,尤其是军事类的命名实体。

(2)将注意力机制与深度学习模型相结合,构建命名实体识别模型。

(3)在模型中构建字向量特征提取部分和词向量特征提取部分,使军事语料中的字向量特征与对应的词向量特征相结合,改善了军事类命名实体错误拆分和识别错误的问题,提高了命名实体识别率。

2 命名实体识别模型

本文构建的命名实体识别模型如图 1 所示。该模型包括输入层、隐层、输出层和条件随机场层,其中,隐层由 BLSTM 和注意力机制两个子层构成。输入层和隐层分别由字向量特征提取部分和词向量特征提取部分并列组成。利用模型隐层提取特征时,句子字向量特征与词向量特征的提取同步进行。融合了注意力机制的命名实体识别模型,可以确保提取到与当前字向量对应的词向量特征,最后将生成的字特征向量与对应的词特征向量相结合,共同作为命名实体识别模型输出层的输入。

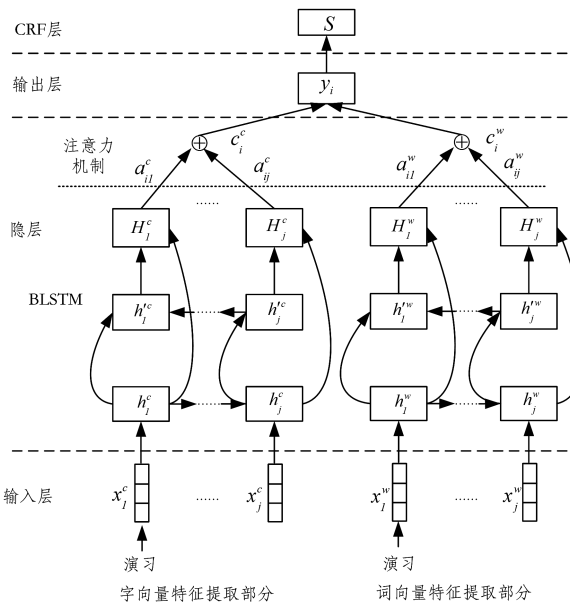


图 1 命名实体识别模型图

2.1 字向量与词向量相结合的输入层

在以往的中文命名实体识别任务中,多采用序列标注模

型,输入一般为单个汉字,但单个汉字包含的信息有限,在利用神经网络模型提取特征时,只能提取与该字相关的特征信息,无法有效提取其所在词代表的语义特征。而词语含有丰富的语义信息,通过特征提取,可以获得此处字所代表的具体含义,对命名实体识别任务有益。文献[12]将字向量分别与所对应的词向量相结合,重新组成新的向量作为输入,实验结果显示与字向量单独作为输入相比,该方法取得了更优的结果,说明字向量与词向量相结合的方法可以提高命名实体的识别效果。

借鉴文献[12]的思想,但并没有在输入层将字向量与词向量相结合,而是在隐层充分提取字向量与词向量特征后,将生成的字特征向量与词特征向量相结合,以改善之前只提取字向量导致的特征不充足的问题,使不同词语中的同一个字特征更加准确,提升命名实体识别效果。

目前部分常用词汇与军事词汇往往相似或相近,识别过程中很容易出现错误。例如:将“成都高新区”和“济南军区”作为训练语料的词汇,“成都高新区”标注为一个地名,而“济南军区”标注为一个军事类的组织机构名。当识别“成都军区”时,如果仅以单字作为输入,由于字向量包含的信息有限,在识别时很可能错误地将其识别为成都市的某个区。而如果把字与词共同作为输入时,“成都军区”作为一个整体,生成的词向量可以对字向量作语义上的补充,识别过程中,其可能与“济南军区”的语义更接近,将其识别为地名时可能因不符合其语义特征而将其正确标注,从而减少错误标注的概率,改善识别错误的情况。

对应字向量特征提取部分和词向量特征提取部分,输入层分别输入一个句子的字拆分序列和词拆分序列,如句子“演习联合指挥所由空军某基地开设”拆分为 15 个单字序列和“演习”“联合指挥所”“由”“空军”“某”“基地”“开设”7 个词组成的词序列,句子的词序列由前续分词模型完成。

训练和测试模型时,如果当前样本中单字和词组的数量小于所设固定长度,则采用填充机制在样本的单字序列和词组序列末尾填充至固定的输入长度。

语料的字序列和词序列通过查找向量表分别得到对应的向量序列,作为模型输入。字向量序列表示为 $(x_1^c, \dots, x_i^c, \dots, x_j^c)$,对应的词向量序列表示为 $(x_1^w, \dots, x_i^w, \dots, x_j^w)$ 。

2.2 结合 BLSTM 和注意力机制的隐层

隐层是整个命名实体识别模型中最重要的部分,它用于提取输入层向量的特征,分为字特征向量和词特征向量两部分,分别用于获取字特征向量 c_i^c 及词特征向量 c_i^w ,每部分由双向长短时记忆网络与注意力机制共同组成。

2.2.1 双向长短时记忆网络

标准的 LSTM 模型只能利用上文信息来推测当前结果,而在序列标注任务中,下文信息对当前结果的推测也有一定的作用。在标准 LSTM 上增加一个后向的 LSTM 构建 BLSTM 模型,前向 LSTM 层和后向 LSTM 层连接同一层,为每个字提供完整的上下文信息,可以将所有信息综合考虑,取得较好的输出结果。

将输入的字向量序列 $(x_1^c, \dots, x_i^c, \dots, x_j^c)$ 分别输入字特征向量部分的前向 LSTM 和后向 LSTM 中,其计算过程如下。

将字向量序列中某一个向量 x_i^c 输入图 2 所示的前向 LSTM 单元中。

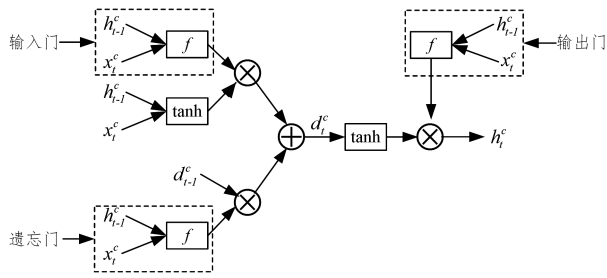


图2 LSTM单元内部结构

x_i^c 和上一个 LSTM 单元的输出 h_{i-1}^c 共同作为输入来计算遗忘门,公式表示如下:

$$a = f(w_0 \cdot (h_{i-1}^c, x_i^c) + b_0) \quad (1)$$

遗忘门得到的结果 a 与记忆单元前一刻的输出 d_{i-1}^c 相乘,可以去除记忆单元需要遗忘的信息,得到需要保留的信息,而经 sigmoid 层计算得到的输入门 i 与 tanh 层得到的结果 d_i^c 相乘,可以获得要更新的信息,将两者相加得到新的记忆单元的信息 d_i^c 。公式表示如式(2)~式(4)所示。

$$i = f(w_1 \cdot (h_{i-1}^c, x_i^c) + b_1) \quad (2)$$

$$d_i^c = \tanh(w_2 \cdot (h_{i-1}^c, x_i^c) + b_2) \quad (3)$$

$$d_i^c = a \cdot d_{i-1}^c + i \cdot d_i^c \quad (4)$$

在得到新的记忆单元信息 d_i^c ,将其经 tanh 层计算,将其与 sigmoid 层的输出门 o 相乘得到输出门输出的信息 h_i^c , h_i^c 即是字向量 x_i^c 输入到前向 LSTM 单元的输出结果。公式表示如式(5)、式(6)所示。

$$o = f(w_3 \cdot (h_{i-1}^c, x_i^c) + b_3) \quad (5)$$

$$h_i^c = o \cdot \tanh(d_i^c) \quad (6)$$

根据以上公式最终得到前向 LSTM 输出向量序列为 $(h_1^c, \dots, h_i^c, \dots, h_j^c)$ 。公式中 w_0, w_1, w_2, w_3 代表权重, b_0, b_1, b_2, b_3 代表偏置项。在字向量序列 $(x_1^c, \dots, x_i^c, \dots, x_j^c)$ 依次输入到前向 LSTM 单元的同时,字向量序列 $(x_j^c, \dots, x_i^c, \dots, x_1^c)$ 也依次输入到后向 LSTM 单元中,其计算过程与前向 LSTM 计算过程相似,得到后向 LSTM 输出向量序列为 $(h_j^c, \dots, h_i^c, \dots, h_1^c)$ 。将同一位置前向 LSTM 的单元的输出与后向 LSTM 输出拼接后得到 BLSTM 最终的输出向量序列 $(H_1^c, \dots, H_i^c, \dots, H_j^c)$,其中, t 时刻 BLSTM 的输出定义为 $H_t^c = [h_t^c, h_t^c]$ 。

同时,为提取相应的词向量特征,将输入层相应的词向量序列 $(x_1^w, \dots, x_i^w, \dots, x_j^w)$ 输入到词特征向量部分的 BLSTM 模型中,得到 BLSTM 模型的输出向量序列为 $(H_1^w, \dots, H_i^w, \dots, H_j^w)$ 。

2.2.2 注意力机制

双向长短期记忆网络在计算过程中将所有上下文信息都考虑其中,虽然全面,但却无法凸显上下文中关键信息的重要作用,而注意力机制则着重考虑上下文中部分信息与当前信息的相关程度。在命名实体识别中,并不是每个字词都对正确识别出实体有相同的贡献,例如:“这篇文章的笔墨集中在直招士官李翰林身上”,“这篇文章”等字对识别出人名“李翰林”意义不大,但“士官”对正确识别更为重要,因此利用注意力机制可以让命名实体识别模型更专注于寻找与当前输出更有关联的输入信息,减少对冗余信息的关注,通过对输入的不

同部分赋予不同的权重,优化资源配置,提高输出质量。

为了使得到的特征向量更加准确,利用注意力机制为 BLSTM 隐层的输出分配不同的权重,新的输出向量则是由各特征向量与对应权重的乘积相加后获得。注意力机制如图3所示。

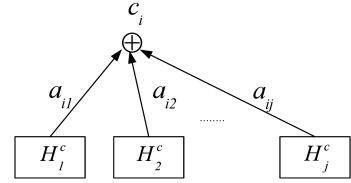


图3 注意力机制

字特征向量计算公式如(7)所示:

$$c_i^c = \sum_{j=1}^n a_{ij}^c H_j^c \quad (7)$$

其中, c_i^c 表示利用注意力机制输出新的字特征向量,它是由 BLSTM 模型输出的各特征向量 H_j^c 与对应权重 a_{ij}^c 的乘积和计算得到。本模型采用式(8)和式(9)计算 a_{ij}^c 。其中 e_{ij}^c 由前一时刻字特征向量 c_{i-1}^c 与 BLSTM 模型输出的特征向量 h_j^c 共同计算得到, v_c, w_c, m_c 分别代表权重。

$$a_{ij}^c = \frac{\exp(e_{ij}^c)}{\sum_{k=1}^n \exp(e_{ik}^c)} \quad (8)$$

$$e_{ij}^c = v_c \tanh(w_c c_{i-1}^c + m_c H_j^c) \quad (9)$$

与提取字特征向量类似,利用注意力机制得到词特征向量 c_i^w 的计算公式如下:

$$c_i^w = \sum_{j=1}^l a_{ij}^w H_j^w \quad (10)$$

$$a_{ij}^w = \frac{\exp(e_{ij}^w)}{\sum_{k=1}^l \exp(e_{ik}^w)} \quad (11)$$

$$e_{ij}^w = v_w \tanh(w_w c_{i-1}^w + m_w H_j^w) \quad (12)$$

2.3 归一化输出层

将隐层计算得到的字特征向量 c_i^c 与词特征向量 c_i^w 共同作为输出层的输入,并通过 softmax 函数进行归一化处理,计算后得到一个 m 维的向量(m 为标签类别的数量),每一个维度对应一个标签的概率值。计算公式如式(13)所示:

$$y_i = \text{softmax}(w_4 c_i^c + w_5 c_i^w + b) \quad (13)$$

其中, w_4, w_5 代表权重, b 代表偏置项。

2.4 CRF 层

传统的命名实体识别方法通过输出层可以计算得到各时刻概率最高的标签序列,这种方法将各标签看成相互独立、毫无联系的个体。事实上,标签之间存在一定的约束关系,而条件随机场模型可以很好地考虑这种约束关系。

模型将输出层的结果输入 CRF 层,并引入转移矩阵 T 作为 CRF 层的参数,利用动态规划算法,求出最优的标签序列。最终标签预测结果由输出层输出 y_i 和转移矩阵 T 共同决定。假设 l_i 代表 i 时刻的标签, T_{ij} 表示从第 i 个标签转移到第 j 个标签的概率,则最后输出的标签序列 S 为:

$$S = \sum_{i=1}^n T_{l_{i-1}, l_i} + \sum_{i=1}^n y_{i, l_i} \quad (14)$$

2.5 损失函数及模型优化

采用交叉熵作为损失函数,计算公式如式(15)、式(16)所示:

$$E_t(Y_t, y_t) = -Y_t \log y_t$$

(15)

$$E(Y, y) = \sum_t E_t(Y_t, y_t) = -\sum_t Y_t \log y_t$$

(16)

其中, Y_t 表示 t 时刻的真实值, y_t 表示 t 时刻的计算值。

以损失函数最小化为目标,采用梯度下降法求解。以计算输出层的权重 w_4 为例,计算公式表示如下:

$$z_i = w_4 c_i^c + w_5 c_i^w + b$$

(17)

$$\frac{\partial E}{\partial w_4} = \sum_t \frac{\partial E_t}{\partial w_4} = \sum_t \frac{\partial E_t}{\partial z_t} \frac{\partial z_t}{\partial w_4}$$

(18)

梯度的计算过程中采用反向传播算法求解,在计算 w_4 梯度时,采用链式法则,通过中间变量 z_i 求偏导得到梯度 $\frac{\partial E}{\partial w_4}$ 。

梯度方向是函数增长最快的方向,因此梯度的反方向就是函数减少最快的方向。为使损失函数尽快达到最小值, w_4 更新计算公式如式(19)所示:

$$w_4 = w_4 + (-\lambda \frac{\partial E}{\partial w_4})$$

(19)

其中, λ 为学习率,表示调整模型参数更新的速度,它决定模型以多快的速度收敛到最小值。采用上述方法更新模型中所有权重和偏置项,使损失函数达到最小值,完成模型的训练。

3 实验

3.1 军事语料库的构建

通过网络抓取工具从解放军报网站爬取军事新闻语料共 7651 篇,每篇新闻语料在 400 字左右,所有语料爬取完成后,去除语料中无效字符作为军事文本语料。

利用中科院研制的 NLPPIR 软件对获取的语料作分词和命名实体识别,并在此基础上,对标识结果作人工校核,主要是纠正未准确切分和标注的军事类词汇,以确保军事语料的准确性。

3.2 实验环境及参数设置

实验所用的软硬件环境和参数设置分别如表 1 和表 2 所列。

表 1 实验环境

软硬件环境	配置及版本
实验主机	Inter(R) Core(TM) i5-4300M
Python	3.5.2
Anaconda3	4.2.0
Tensorflow	0.12.0

表 2 参数设置

参数	数值
样本固定长度	32
字向量长度	64
词向量长度	64
隐含层结点数	128
双向长短时记忆网络层数	2
学习率	0.0001
Batch_size	32
Epoch	10

3.3 实验数据及处理

实验数据集采用军事领域语料库来评估命名实体识别模型的性能,将语料库划分为训练集、测试集、验证集 3 类,其中训练集为 5355 篇,测试集 1148 篇,验证集 1148 篇。

在军事文本命名实体识别时,应更注重其准确性而不是复杂性,实验采用有监督的学习方法,主要标记 3 类命名实体,分别为人名(用 PER 表示)、地名(用 LOC 表示)、组织机构名(用 ORG 表示)。以 BISO 标注模式标注实体,B 代表实体开始,I 代表实体中间或结束,S 代表单个实体,O 代表不是实体。由于须针对军事语料中每一个字作相关标记,因此将实体标记为 10 类:PER-B,PER-I,PER-S,LOC-B,LOC-I,LOC-S,ORG-B,ORG-I,ORG-S,O。

3.4 实验结果

本文进行 3 组实验,分别训练 BLSTM_CRF 模型^[6]、BLSTM_CRF+attention 命名实体识别模型^[13]以及本文提出的基于注意力机制的命名实体识别模型,并分别对测试集作命名实体识别测试,实验采用准确率、召回率和 F 值作为评价指标。实验结果如表 3 所列。

表 3 命名实体识别对比实验结果

(单位:%)

实验模型	准确率	召回率	F 值
BLSTM-CRF 命名实体识别模型	90.03	84.87	87.37
BLSTM-CRF+attention 命名实体识别模型	92.54	87.05	89.71
基于注意力机制的 命名实体识别模型	93.96	87.83	90.79

3.5 实验分析

与 BLSTM-CRF 模型相比,BLSTM_CRF+attention 命名实体识别模型添加了注意力机制,通过注意力机制能够为不同的上下文信息分配相应大小的权重,使模型更加注重于寻找与当前输出更有关联的信息,达到了更好的识别效果。而本文提出的命名实体识别模型则是在 BLSTM_CRF+attention 模型的基础上,结合了词向量的特征,完善了提取字向量特征不足的问题,因此与 BLSTM_CRF+attention 模型相比,准确率、召回率和 F 值都有提高,说明本文提出的模型能够提升命名实体的识别效果。

结束语 注意力机制在对目标词判断时,能够最大限度地利用有效特征减少无用特征的干扰,因此本文将注意力机制融入命名实体识别模型。同时在实体识别中,改变只以字向量作为输入的方式,以字向量和词向量同时作为输入层的输入,在隐层利用注意力机制分别生成字特征向量和对应的词特征向量,并将二者共同作为输出层的输入,最后通过维特比算法解码,输出最终的命名实体标签。

虽然本文提出的命名实体识别模型可以提升军事文本中命名实体的识别率,但对军事类命名实体的关注度不够。对于军事领域等专业领域而言,特征提取十分重要,因此在后续工作中,将对军事领域语料深入研究,分析军事类命名实体的特点,提取更多特征,使模型识别效果能够真正达到实际应用水平。

参 考 文 献

[1]

俞鸿魁,张华平,刘群,等.基于层叠隐马尔可夫模型的中文命名实体识别[J].通信学报,2006(2):87-94.

[2]

胡文博,都云程,吕学强,等.基于多层条件随机场的中文命名实体识别[J].计算机工程与应用,2009,45(1):163-165,227.