计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.201100007

基于领域适应嵌入的军事命名实体识别

刘 凯[1] 张宏军[2] 陈飞琼[1]
1 陆军工程大学研究生院 南京 210000
2 陆军工程大学指挥控制工程学院 南京 210000
(lkljgc@163.com)

摘 要 为了解决单一军事领域语料不足导致的领域嵌入空间质量欠佳,使得深度学习神经网络模型识别军事命名实体精度较低的问题,文中从字词分布式表示入手,通过领域自适应方法由额外的领域引入更多有用信息帮助学习军事领域的嵌入。首先建立领域词典,将其与CRF算法结合,对收集到的通用领域语料和军事领域语料进行领域自适应分词,作为嵌入训练语料,并将词向量作为特征与字向量拼接,以丰富嵌入信息并验证分词效果;然后对训练所得的通用领域和军事领域的异构嵌入空间进行领域自适应转换,生成领域自适应嵌入,并作为基础模型BiLSTM-CRF层的输入;最后通过CoNLL-2000进行识别评价。实验结果表明,在相同模型下,输入领域适应嵌入比输入一般分词后的语料训练所得的军事领域嵌入,其模型识别的精确率(P)、召回率(R)、综合F1值(F1)分别提高了2.17%,1.04%,1.59%。

关键词:字向量;词向量;中文分词;领域自适应;命名实体识别
中图法分类号 TP391.1

Name Entity Recognition for Military Based on Domain Adaptive Embedding

LIU Kai[1],ZHANG Hong-jun[2] and CHEN Fei-qiong[1]
1 School of Graduate,Army Engineering University of PLA,Nanjing 210000,China
2 College of Command and Control Engineering,Army Engineering University of PLA,Nanjing 210000,China

Abstract In order to solve the poor quality problem of domain embedding space caused by inadequate military corpus which makes low accuracy of applying deep neural network model to military named entity recognition,this paper introduces a domain adaptive method to help learn the embedding of military fields from more useful information of additional fields through distributed representation of words. First,we establish the domain dictionary and combine CRF algorithm to perform domain adaptive word segment with the collected general domain and military areas corpus as training corpus for embedding,and word vectors are used as features and spliced with character vectors to enrich the embedding information and to validate the effect of word segmentation. Then the domain adaptive transformation is carried out to the heterogeneous embedded space of the general domain and the military domain,and the domain adaptive embedding is generated,as the input to BiLSTM-CRF layer of base model. At last,the recognition evaluation is carried out through CoNLL-2000. The experimental results show that,under the same model,the recognition precision rate (P),recall rate (R),and integrated F1 value (F1) of the proposed method are improved by 2.17%,1.04%,and 1.59%,respectively,compared with the military field embedding trained by a corpus which is obtained from general word segmentation.

Keywords Character embedding,Word embedding,Chinese word segmentation,Domain adaptation,Named entity recognition

1 引言

现如今,通过神经网络学习低维稠密的词向量分布式表示已成为研究热点,这样的表示可以捕捉到词语之间的语义和语法关系。词的分布式表示即词嵌入经常以一种无监督方式从大量文本中学习得到,且对于提高模型在命名实体识别、情感分析等下游任务上的性能有很大帮助。

从前期的基于词典和规则的方法,到传统机器学习的命名实体识别方法都已经较为成熟,但利用的特征不能体现数据间良好的语言规律和语义信息,因此识别效果较差。后期提出的基于人工神经网络的深度学习方法,可以自动学习到合适的特征和多层次的表达,在识别模型和效果上均取得了一定突破。自Huang等[1]构建起基于词向量的BiLSTM-CRF这一主流模型,后续研究才进一步展开。之后Lample等[2]融合字级别和词级别的信息,将BiLSTM-CRF作为端到端的模型进行实体识别,在CoNLL-2003数据集上其F1值达到90.94。文献[3]是在RNN-CRF模型结构的基础上,使用attention机制重点改进了词向量与字符向量的拼接,使得模

到稿日期:2020-11-02 返修日期:2021-03-17 本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
通信作者:张宏军(jsnjzhj_lgdx@163.com)

型可以动态地利用词向量和字符向量信息，效果显著提升。文献[4]引入多头注意力机制解析词之间的联系，通过设置权重的方式多模态匹配词典，提高复杂和罕见实体的识别。文献[5]使用语言模型生成的基于上下文相关的字向量作为NER模型输入，来解决数据稀疏和未登录词(Out Of Vocabulary，OOV)问题。近年来，能够学习文本深层编码信息的Bert模型[6]的出现，在迁移学习和预训练模型上为命名实体识别任务提供了新的研究方向。

命名实体识别任务在通用领域研究已经比较成熟，但在军事领域却仍不够深入。提高提取非结构化军事文本信息的能力，对于辅助作战行动和指挥决策部署具有重要意义。军事命名实体指在军事文本中与军事相关的各种命名实体的统称，包括人员、军事地名、军事组织机构、军事武器装备、军事职别职级、部队编制等各类实体。命名实体的识别模型一般采用条件随机场(CRF)和双向长短时记忆网络(BiLSTM)。文献[7]利用word2vec工具训练词向量，利用邻近词作为特征优化CRF模型，提升模型性能。之后，文献[8]提出了特征词筛选算法，用于获取军事领域词向量，但是其需要更多无法直接得到的专业领域语料进行训练。2019年，文献[9]提出了融合字符向量、词向量和词性特征向量的CNN-BiLSTM-CEF模型，并在作战文书命名实体识别中表现出良好的性能。自2019年文献[10]使用注意力机制来提取嵌入向量特征，提高了BiLSTM-CRF网络的识别精度以来，注意力等热点机制开始在军事领域中得到应用。

上述诸多研究表明，字词等向量信息对命名实体识别效果有较为明显的影响，但是为了对人工标注的少量训练语料进行丰富的特征学习，从而更精确地提取军事文本中的有效信息，需要大量语料进行向量生成模型的训练。然而在实际情况中，军事领域语料较为匮乏，通常需要通过收集媒体文本等不同领域的大量语料进行扩充训练，但是因为人工收集到的语料包含更多领域的冗杂信息，导致模型无法学到目标领域中质量更好的嵌入。文献[11-12]提出通过迁移学习中的领域自适应等方法由额外的领域引入更多的有用的信息来帮助学习任务领域的嵌入，其目的是利用源域的丰富数据提高目标域的语言性能。因此，为了利用源域和目标域的不同数据丰富特定领域下的字词嵌入信息，解决单一军事领域语料不足的问题，提高军事领域命名实体识别的准确率，本文提出了建立领域词典并将其与CRF算法结合进行领域自适应分词，并且通过对通用领域和军事领域的异构嵌入空间进行转换建立领域适应嵌入作为模型嵌入层输入，最后采用命名实体识别通用模型BiLSTM-CRF进行效果评估。

2　军事命名实体识别模型

2.1　领域自适应过程

传统分词方法包括基于最大匹配词典的机械分词方法[13]、基于规则的解决歧义的分词方法[14]和基于统计的根据字词共现频率进行分词的方法[15]，这些方法均为单一机械，不能更好地适用于不同领域的分词，尤其是当出现领域专有名词时。因此，本文提出了基于领域词典和CRF算法的自适应分词方法，得到更适应于军事领域的分词语料作为后续嵌入的训练数据。

大多数迁移学习方法都使用通用领域词嵌入作为模型输入，但是当出现不同领域的术语时，这些词可能有不同语义，向量表示也应有所不同，而一些重要领域的特定词可能不存在于通用领域嵌入的词汇中，因此会出现严重的OOV问题。Stenetorp等[16]研究表明，在监督学习任务中，领域特定的词嵌入往往表现得更好，因此我们通过人工采集领域语料，但是当采集的领域语料较少、领域词汇不足时，如军事领域语料，此时需要通过其他领域语料进行迁移学习，补充语料信息。然而，在迁移学习过程中将源域和目标域上的异构嵌入空间进行传递是非常有困难的。我们通过嵌入空间投影方法训练领域自适应嵌入以解决上述问题。

2.1.1　领域自适应分词

根据文献[17]，引入外部词典对命名实体识别有很大用处。在现有的分词方法基础上加载特定领域，如军事领域的专属词典，将提高命名实体识别的精度，使分词方法具有领域适应性，很好地解决了领域内专属名词的识别问题。当领域改变时，不需改变现有的分词模型，只需改变领域词典即可。本节以哈工大提出的基于CRF算法的LTP中文分词系统[1)]为基础，添加自定义的外部词典，以便于对后续通用领域和军事领域嵌入训练语料进行分词。分词流程如图1所示。

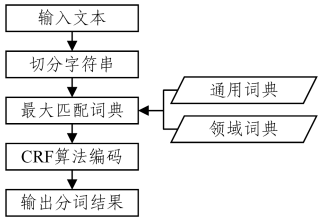


图1　领域自适应分词流程图

Fig. 1　Domain adaptive word segmentation flowchart

首先对输入文本进行字符串切分，切分成原子字符，再对每一个字符按照序列标注方法进行标注，即根据每个字在词中的位置进行B,M,E,S标注[18]，其中B表示当前字是一个词开始，M表示当前字在一个词内部，E表示当前字在一个词结尾，S表示独字词。标注规则为每个字与其后的n个字试组词($n=1,\cdots,5$)，然后与内部通用词典和外部领域词典进行最大匹配。最后经过CRF算法中的维特比解码，输出分词结果。

其中，内部通用词典使用LTP系统默认内部词典，含有8625个词；外部领域词典由我们自行构建，收集2011年版《中国人民解放军军语》7492条词条，军事武器装备词条2627条，军事术语百度百科3936条词条，共计14055个词。

此外，因为我们在准备数据和生成batch样本时需要拆成字的样本，为了使每个字标签更加精确，单字区分度更高，且为了更好地检验分词效果对识别结果的影响，受文献[19]的启发，且根据文献[20]的研究结果分析，词表示可以消除单字的歧义，字符信息可以丰富词语的语义信息，所以我们采用比单独使用字或词嵌入效果更好的拼接字向量及其所在词词向量的方式生成嵌入。本文对输入模型的训练样本进行词语

1) https://ltp.readthedocs.io/zh_CN/latest/appendix.html

最大匹配，然后将分词后训练所得的词向量作为特征与字向量拼接。

2.1.2 领域自适应嵌入

目前的预训练词嵌入通常是在较为庞大的语料库上进行训练的，因此如果重新训练将耗时耗力，又因为如果仅仅将军事领域和通用领域语料合并生成嵌入会造成目标域的语义偏差，所以我们提出将学习到的嵌入从目标嵌入空间投影到源空间来构建领域自适应嵌入。其基本思想为：首先通过计数法对源域和目标域出现的字词进行频率统计，根据预设的频率阈值选择同时在两个领域出现的字词，组成词对，构建源域和目标域数据集的中心词词典；然后采取一定算法计算每个词对的重要程度，并进行排列，通过预设阈值剔除低于阈值的词对，得到中心词筛选词典；最后结合预训练好的源域和目标域嵌入，学习转换矩阵，实现从源域到目标域的线性投影，利用源域丰富信息补充目标域嵌入。

(1)中心词筛选词典构建

受文献[8,21-22]的启发，我们定义如下概念。

定义 1 中心词词典 P 由一系列来自源域和目标域中相对出现频率较高的中心词对(w_s, w_t)组成：

$$P=\{(w_s,w_t)\mid w_s=w_t, f(w_s)\geqslant n_s, f(w_t)\geqslant n_t\}$$

其中，w_s属于源域词汇表V_s，w_t属于目标域词汇表V_t，$f(w)$是数据集中词语 w 出现的频率，n_s和n_t是单词频率阈值，阈值由单词频率排序列表得到。

定义 2 $\bar{f}(w_s)$和$\bar{f}(w_t)$为词典中每一对来自源域和目标域词汇的归一化频率：

$$\bar{f}(w_s)=\frac{f(w_s)}{\max f(w')_{w'\in V_s}}$$

$$\bar{f}(w_t)=\frac{f(w_t)}{\max f(w')_{w'\in V_t}}$$

定义 3 α_i是词对P_i的置信度，表示词对的重要性。P_i代表中心词词典 P 中的第 i 行词对。α_i由归一化的词语频率确定：

$$\alpha_i=\frac{2\cdot\bar{f}(w_s^i)\cdot\bar{f}(w_t^i)}{\bar{f}(w_s^i)+\bar{f}(w_t^i)}$$

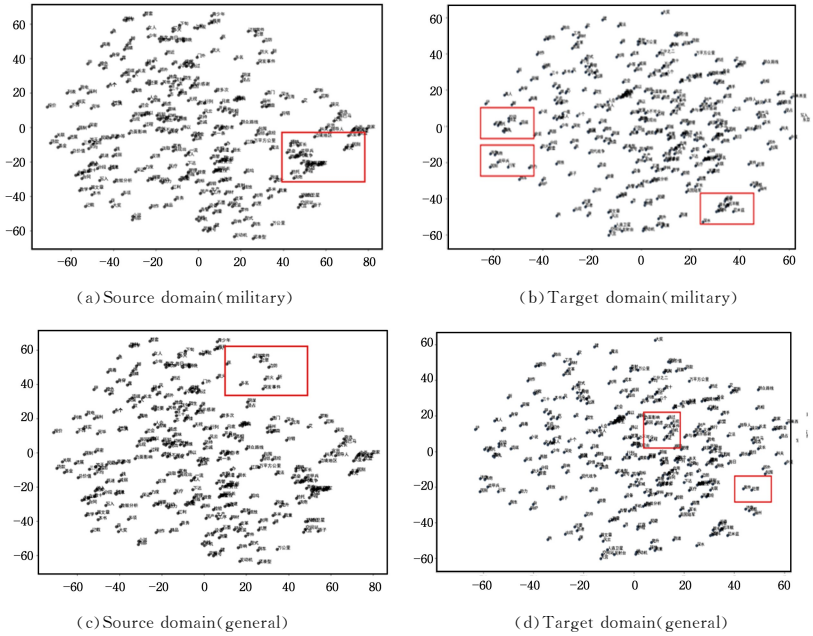
根据计算所得置信水平 α 的值对词对进行排列，当两个词在两个域中都有较高的相对频率时，这对词更重要。因为这些词可能更独立于领域，即在这两个不同的领域表达相同或相似的意思[22]。本文设置投影参数 θ，根据置信水平 α 排列结果，按序按比例选择词对数，控制词典 P 的大小，验证嵌入转换程度对模型识别效果的影响。

定义 4 中心词筛选词典P'，根据投影参数 θ，按照置信度 α，对中心词典 P 进行筛选，即：

$$P'=\{(w_s,w_t)\mid w_s=w_t,\alpha_s\geqslant m_s,\alpha_t\geqslant m_t,|P'|=\theta*|P|\}$$

其中，α_s和α_t分别代表源域和目标域中心词置信度，m_s和m_t为置信度阈值，$|P'|$为中心词筛选词典P'的大小，$|P|$代表中心词词典 P 的大小。

通过构建中心词筛选词典P'得到如(年，年)的词对，词对中的字虽然相同，但是前者来自源域，后者来自目标域，两者的向量表示不同。如图 2 所示，第一行图 2(b)的目标域能够更好地对军事类相似词进行聚类，如“舰艇”“弹丸”，而第二行图 2(c)的源域能够更好地对一般词进行聚类，如“犯罪”。下一步空间投影就是为了最小化这样的中心词嵌入空间差异。



(a)Source domain(military) (b)Target domain(military)

(c)Source domain(general) (d)Target domain(general)

图 2 源域和目标域军事类词与一般类词向量聚类图

Fig. 2 Source domain and target domain military words and general words vector clustering graph

(2)领域空间投影

在给定预训练的领域词嵌入$\boldsymbol{w}_s$和$\boldsymbol{w}_t$，以及中心词筛选词典P'时，根据 Mikolov 等[23]，将单词向量从$\boldsymbol{w}_t$线性投影到$\boldsymbol{w}_s$，以完成领域适应嵌入构建。

我们首先构造两个矩阵$\boldsymbol{W}_S$和$\boldsymbol{W}_T$，这两个矩阵的第 i 行，即$\boldsymbol{w}_s^i$，$\boldsymbol{w}_t^i$，是词典P'中位于第 i 行的词对(w_s^i, w_t^i)的向量表示。

接下来，利用以下函数学习一个变换矩阵 $\boldsymbol{Z}$，以最小化源域和目标域嵌入空间之间的距离：

$$\arg\min_{Z}\sum_{i=1}^{|P'|}\alpha_i\left\|\boldsymbol{W}_T^i\boldsymbol{Z}-\boldsymbol{W}_S^i\right\|^2 \tag{1}$$

学习之后，投影的新嵌入将是$\boldsymbol{W}_T\boldsymbol{Z}$。所以最小二乘拟合多项式（即式(1)）可转化为求目标函数$J(\boldsymbol{Z})$的极值问题：

$$J(\boldsymbol{Z})=\sum_{i=1}^{|P'|}\left\|\boldsymbol{W}_T^i\boldsymbol{Z}-\boldsymbol{W}_S^i\right\|^2 \tag{2}$$

即求：

$$J(\boldsymbol{Z})=\sum_{i=1}^{|P'|}(\boldsymbol{W}_T^i\boldsymbol{Z}-\boldsymbol{W}_S^i)^{\mathrm{T}}(\boldsymbol{W}_T^i\boldsymbol{Z}-\boldsymbol{W}_S^i) \tag{3}$$

的偏导。

本文嵌入维度为100，故$\boldsymbol{W}_T^i$和$\boldsymbol{W}_S^i$的维度均为$|P'|\times 100$，所以，式(3)的导数式等式两边为对称矩阵，存在唯一的转换矩阵$\boldsymbol{Z}$，从而得到转换后嵌入$\boldsymbol{W}_T\boldsymbol{Z}$。最后将未在目标域中出现的源域词添加至嵌入空间，进行源域信息补充，得到最终的嵌入向量矩阵。

此过程即为领域嵌入自适应过程，实现对目标域嵌入的扩充丰富，进而作为模型输入提高命名实体识别的精度。

2.2　基础模型

本节简要介绍了NER任务的BiLSTM-CRF模型结构，如图3所示，其为本文基础模型。

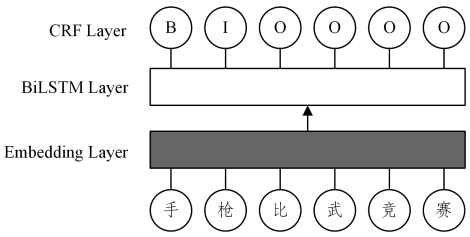


图3　基础模型图

Fig. 3　Basic model

根据Lample等[2]的研究，我们将word2vec[24]预训练后的向量表示作为命名实体识别模型的输入。中间层为能够捕捉长距离依赖和上下文信息的双向长短时记忆网络（BiLSTM），其产生每个标记的隐层状态，生成标记的预测标签概率矩阵，因为需要预测各字词的分类标签，而每个字只有一个标签，所以模型通过在输出层后添加条件随机场（CRF）层进行模型微调和标签约束。模型训练中，通过神经网络的反向传播不断更新模型中所有的权重和偏置项，使损失函数达到最小值，完成训练，从而提高模型的标签预测效果。与目前比较受欢迎的Bert等[6]语言模型相比，BiLSTM＋CRF模型对于本文方法的改进效果评估更简单方便，而且其结构易于优化，训练效率高，比Bert更节省计算资源，所以BiLSTM与CRF搭配的模型在命名实体识别等任务上更受欢迎。基于其通用性和代表性，本文选用该架构作为基础模型。

2.3　训练过程

训练流程图如图4所示。训练流程为：首先利用领域适应方法对嵌入训练语料进行分词；然后经过word2vec工具训练通用领域嵌入（源域）和军事领域嵌入（目标域）；再将源域嵌入投影到目标域嵌入空间，缩小中心词语义空间差距，实现领域自适应，补充目标域嵌入；最后利用转换后的嵌入作为模型嵌入层，经过双向长短时记忆层和条件随机场层进行标签预测和微调。文献[25]提出的ELMO模型表明，利用LSTM的各层状态来表示词向量可以表达更加丰富的信息，LSTM网络第一层可以获取到词语的语法信息，而第二层状态表达词语的语义信息，因此我们采用双层双向长短时记忆层，以更好地学习上下文信息，补充OOV词向量。训练过程中，首先使用参数初始化模型，并使用Adam优化算法不断调整更新所有层的权重。当使用Adam更新不同层中的权重时，我们使用梯度截断及学习率衰减来避免梯度爆炸，并且使用dropout减小模型过拟合问题。

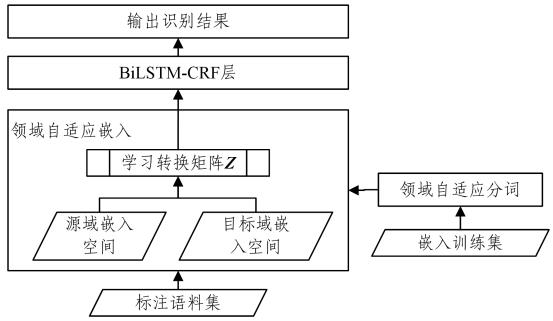


图4　训练流程图

Fig. 4　Training process

当我们输入一个句子时，模型预测的标签路径有很多，但是真正正确的路径只有一条，所以CRF层的损失函数L建立如下：

$$L=-\log\frac{S_{real\,path}}{S_1+S_2+\cdots+S_N}$$

其中，S_N代表第N条预测标签序列分数，$S_{real\,path}$代表真实标签序列分数。每条路径上的分数由句子上的字与标签的发射概率和标签与标签间的转移概率加和计算。最后，加载训练模型，输出提取的实体、类别以及实体在句中的位置边界。

3　实验设置

3.1　实验环境

实验中，实机的主要软硬件参数设置如表1所列。

表1　软硬件参数表

Table 1　Hardware and software parameter

Type	Configuration and version
Hardware	Inter(R) Core(TM) i7-8759H，win10，16GRAM
Software	Python3.6，pytorch1.4

3.2　数据集

实验使用的数据集包括词嵌入训练数据集和模型训练标注数据集。本文采用jieba[1)]工具进行分词，word2vec工具进行词嵌入训练。源域即通用领域语料，其来源于公开wiki2019[2)]中文语料，包含104万个中文词条。目标域即军事领域语料，由scrapy爬虫获取的1637篇军事新闻、4339篇武器装备等军事百度百科资料组成。语料分布如表2所列，共有约565万个有效词，39万条长句。由于目前缺乏专业的军事领域标注语料库，因此，本次实验采用的军事领域语料库经

1) https://github.com/fxsjy/jieba/

2) https://github.com/brightmart/nlp_chinese_corpus

人工标注生成。首先利用爬取到的百度百科信息经过人工筛选和标签标注建立词典库；然后采用双向最大匹配算法(BMM)对爬取得到的410篇军事新闻语料进行实体标注；最后得到标注数据集分布，如表3所列。

表2 嵌入训练数据集

Table 2 Embedding training dataset

数据集		篇数	词数×10^4(≈)	句数×10^4(≈)
中外军事新闻	陆军	266	66	9
	海军	222		
	空军	210		
	太空作战	504		
	网络作战	183		
	电磁作战	103		
	核武器作战	149		
军事百科	武器装备	3025	499	30
	弹药	751		
	其他	563		
总计		5976	565	39

表3 标注数据集

Table 3 Annotated dataset

Dataset	Number(sentences)
Train_set	27444
Test_set	6825
Dev_set	4640

3.3 命名实体标签设置

因为通用的命名实体识别一般包括人名、地名、组织机构名3类，所以本文也对人名、地名和军事组织机构进行实体识别，另外还添加了军事领域特有命名实体，如军事武器装备和弹药。标签集采用BIO标签，其中B代表实体首部、I代表实体内部、O代表非实体，具体如表4所列。

表4 命名实体识别标签

Table 4 Named entity recognition tags

Entity type	Entity head	Entity interior and tail
人名	B-PER	I-PER
军事地名	B-LOC	I-LOC
军事组织机构名	B-ORG	I-ORG
武器装备	B-WEAP	I-WEAP
军事弹药	B-AMM	I-AMM

3.4 评价指标

对于NER的评价指标，我们使用CoNLL-2000评估脚本进行精确率(P)、召回率(R)和综合指标F1评分。CoNLL-2000评估脚本是一种精准匹配的方法，只有当实体边界和实体类别同时被标记正确时，才能认为实体识别正确。

$$P=\frac{M_{\text{正确识别实体数}}}{N_{\text{识别实体数}}}\times 100\%$$

$$R=\frac{M_{\text{正确识别实体数}}}{N_{\text{实体总数}}}\times 100\%$$

$$F1=\frac{2\times P\times R}{P+R}\times 100\%$$

4 实验结果与分析

首先，本文设置学习率为0.004，分别比较投影参数θ为0，0.2，0.4，0.6，0.8，1.0时的评价指标，结果如图5所示。然后选择最优θ值进行下一步实验。

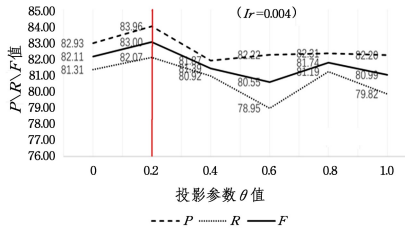


图5 不同投影参数θ实验结果对比

Fig. 5 Different projection parameters θ experimental results

由图5可知，θ值取1时，代表仅使用自适应分词后的单一军事语料，此时，精度降低较大，说明将中心词全部转换，反而会丢失信息，而当θ值为0.2时，P，R，F1值的综合效果最好，所以选择此值反映嵌入空间转换程度。

最后在$\theta=0.2$，学习率$lr=0.004$，$batchsize=128$，隐层结点数为256的情况下，采用BiLSTM+CRF基础模型进行4组对比实验。实验1－实验4分别为对基础模型输入结巴分词后语料训练所得的军事语料嵌入(jieba+mili)、结巴分词后语料训练所得的领域自适应嵌入(jieba+dominadaption emdedding(DAE))、领域自适应分词后语料训练所得的军事语料嵌入(Domin Adaption Split(DAS)+mili)和领域自适应分词后语料训练所得的领域自适应嵌入(DAS+DAE)。实验结果如表5所列。

表5 实验结果对比

Table 5 Comparison of experimental results

(单位：%)

Model input	Precision rate	Recall rate	F1
Jieba+mili	81.79	81.03	81.41
Jieba+DAE	82.57	81.23	81.90
DAS+mili	82.93	81.31	82.11
DAS+DAE	83.96	82.07	83.00

理想情况下，精确率和召回率两者越高越好，且在命名实体识别任务中，对于两者都有较高的要求，但是在实际情况下，精确率较高时，召回率往往比较低，反之亦然。所以综合考虑精确率和召回率一般采用F1值。表5中，实验1和实验3、实验2和实验4的结果表明，经过领域自适应分词后F1值可提升0.7%～1.1%；实验1和实验2、实验3和实验4分别验证了经过领域自适应嵌入后F1值可提升0.5%～0.9%。综合考虑，本文方法相比仅使用jieba分词后的单一军事语料训练所得嵌入的精确度、召回率和F1值均有提升，且综合指标F1值达到83%，改进效果较为显著。

结束语 针对军事命名实体识别的需求及军事领域语料的不足，嵌入质量较低且信息缺失的问题，本文引入通用语料进行语料扩充，但是为了使通用多领域语料更适应于军事领域，本文在命名实体识别的基础模型上不仅利用词典与CRF算法进行领域自适应分词，使得语料更具有目标域特征，而且还通过领域迁移方法训练领域自适应嵌入，使用通用领域嵌入丰富军事领域嵌入，通过线性投影，缩小领域间嵌入空间的差异，既保留了军事信息，又能学习通用领域丰富的知识，同时引入字词向量拼接结合的方式，既丰富了嵌入信息又能检验分词效果，有助于更好地识别军事命名实体，最终总体效果

提升 1.30%±0.003。由于缺乏专业的军事语料库，本文所用语料库均为人工获取和标注，故而会出现标注数据标签分布不均匀的问题，如文本 AMM 实体远少于 WEAP 实体，此外军事复合实体（如“导弹驱逐舰”，既包含弹药 AMM 实体“导弹”，又包含武器装备 WEAP 实体“驱逐舰”）的识别也影响着模型的识别精度。因此，下一步将继续探索语义更加准确丰富的嵌入形式，并进一步在复杂实体层面进行标注语料构建。

参 考 文 献

[1] HUANG Z H, XU W, YU K. Bidirectional LSTM-CRF models for sequence tagging[EB/OL]. (2015-08-09)[2020-10-01]. https://arxiv.org/pdf/1508.01991.

[2] LAMPLE G, BALLESTEROS M, SUBRAMANIAN S, et al. Neural architectures for named entity recognition[C]// Proceedings of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics. 2016: 260-270.

[3] REI M, CRICHTON G, PYYSALO S. Attending to characters in neural sequence labeling models[C]// Proceedings of the 26th International Conference on Computational Linguistics. 2016: 309-318.

[4] XU K, WANG Q, LI Z Z, et al. Biomedical named entity recognition based on BiGRU network with multi-head attention mechanism[J]. Computer applications and software, 2020, 37(5): 151-232.

[5] ZHANG D, CHEN W L. Chinese Named Entity Recognition Based on Contextualized Char Embeddings[J]. Computer Science, 2021, 48(3): 233-238.

[6] DEVLIN J, CHANG M W, LEE K, et al. Bert: pre-training of deep bidirectional transformers for language understanding [C]// Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2019: 4171-4186.

[7] JIANG W Z, GU J J, HU W X, et al. Military named entity recognition based on multi-models[J]. Ordnance Industry Automation, 2011, 30(10): 90-93.

[8] QIN J, CAO L, PENG H, et al. A domain feature word vector description method for military texts[J]. Computer Engineering, 2016, 42(8): 160-165.

[9] ZHANG X H, CAO X W, GAO Y. Named Entity Recognition for Combat Documents Based on Deep Learning[J]. Command Control & Simulation, 2019, 41(4): 22-16.

[10] SHAN Y D, WANG H J, HUANG H, et al. Study on Named Entity Recognition Model Based on Attention Mechanism-Taking Military Text as Example[J]. Computer Science, 2019, 46(z1): 111-114.

[11] PAN S J, QIANG Y. A Survey on Transfer Learning[J]. IEEE Transactions on Knowledge and Data Engineering, 2010, 22(10): 1345-1359.

[12] WEISS K, KHOSHGOFTAAR T M, WANG D D. A survey of transfer learning[J]. Journal of Big Data, 2016, 3(1): 9.

[13] GUO T K. A research on Chinese Word Segmentation based on Dictionary[D]. Harbin: Harbin University of Science and Technology, 2010.

[14] ZHANG J. A Chinese Word Segmentation Method Based on Rules[J]. Computer and Modernization, 2005(4): 18-20.

[15] ZHAO Y Z. A Chinese word segmentation method based on word frequency statistics[J]. Science and Technology, 2016, 26(10): 283.

[16] STENETORP P, SOYER H, PYYSALO P, et al. Size(and domain) matters: Evaluating semantic word space representations for biomedical text[C]// Proceedings of the 5th International Symposium on Semantic Mining in Biomedicine. 2012.

[17] ZHANG M S, CHE W X, LIU T. Combining statistical model and dictionary for domain adaption of Chinese word segmentation[J]. Journal of Chinese Information Processing, 2012, 26(2): 8-12.

[18] XUE N. Chinese word segmentation as character tagging[J]. International Journal of Computational Linguistics and Chinese Language Processing, 2003, 8(1): 28-48.

[19] XIE Z N. Research on Chinese name entity recognition algorithm [D]. Hangzhou: Zhejiang University, 2017.

[20] LI W K, LI W, WU Y F. Combination methods of Chinese character and word embeddings in deep learning[J]. Journal of Chinese Information Processing, 2017, 31(6): 140-146.

[21] TAN L C, ZHANG H T, SMUCKER M, et al. Lexical comparison between wikipedia and twitter corpora by using word embeddings[C]// Proceedings of ACL. 2015.

[22] LIN B Y, LU W. Netural adaptation layers for cross-domin named entity recognition[C]// Proceedings of the 2018 Conference on EMNLP. 2018: 2012-2022.

[23] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed Representations of Words and Phrases and their Compositionality[C]// Proceedings of Neural Information Procesing Systems Foundation. 2013.

[24] MIKOLOV T, CORRADO G, CHEN K, et al. Efficient Estimation of Word Representations in Vector Space[C]// Proceedings of the ICLR. 2013: 1-12.

[25] PETERS M, NEUMANN M, IYYER M, et al. Deep contextualized word representations[C]// Proceedings of NAACL-HLT. 2018: 2227-2237.

LIU Kai, born in 1996, postgraduate, His main research interests include natural language processing and so on.

ZHANG Hong-jun, born in 1963, professor, Ph.D supervisor. His main research interests include military modeling & simulation and data engineering.

（责任编辑：喻藜）