文章编号：2095-2163(2020)02-0123-05 中图分类号：TP391 文献标志码：A

基于多神经网络协同训练的命名实体识别

王 栋，李业刚，张 晓
（山东理工大学 计算机科学与技术学院，山东 淄博 255049）

摘 要：为了提高命名实体识别模型的系统实用性，有效利用互联网中海量未经标注的数据，提出了一种基于多神经网络协同训练的命名实体识别模型。该模型融合了循环神经网络和协同训练的优势，首先利用少量的有标记数据训练 3 种不同的神经网络获得初始识别模型，然后在大量无标注数据上对 3 种神经网络模型进行协同训练以优化模型。实验结果表明，本文模型能够有效地训练大量的无标记数据，与传统的协同训练和单一神经网络识别模型相比，模型的整体性能得到了显著提升。

关键词：命名实体识别；循环神经网络；协同训练

Named entity recognition based on tri-training of multiple neural network

WANG Dong, LI Yegang, ZHANG Xiao
(College of Computer Science and Technology, Shandong University of Technology, Zibo Shandong 255049, China)

【Abstract】 In order to improve the system practicability of the named entity recognition model and effectively utilize the massive unlabeled data in the Internet, this paper proposes a named entity recognition model based on tri-training of multi-neural network. The model combines the advantages of recurrent neural network and tri-training. Firstly, three different neural networks are trained with a small amount of labeled data to obtain the initial recognition model, then the tri-training of three neural network named entity recognition models are performed on a large number of unlabeled data to optimize the model. The experimental results show that the model can effectively train a large amount of unlabeled data, and the overall performance of the model is significantly improved compared with the traditional tri-training and single neural network recognition model.

【Key words】 named entity recognition; recurrent neural network; tri-training

0 引 言

命名实体识别(Named Entity Recongition，NER)是自然语言处理领域(Natural Language Processing，NLP)中经常用到的实用性技术，在事件抽取[1]、信息检索[2]、情感分析[3]等许多任务中发挥着重要作用，旨在从文本数据中提取指定的实体信息。传统的 NER 多采用机器学习中的监督学习模型，该类模型需要依赖大量的特征工程和语言学规则。随着信息时代的到来，存在于互联网中的文本数据越来越多，此时采用传统方法的 NER 技术处理如此巨大数据量的文本信息将会十分困难。近年来，深度学习技术(Deep Learning，DL)受到了研究者的广泛关注，基于深度学习的 NER 技术取得了丰硕的成果，深度学习中的神经网络模型可以有效地自动捕获序列文本中的词级和字符级特征，避免了对特征工程的依赖和人工添加语言学规则，显著地提高了 NER 的实用性。然而，基于深度学习的命名实体识别方法与统计机器学习方法一样，模型的训练需要大量的有标记数据，以确保识别精度的准确性。但是有标记数据需要人工进行标注，此过程必然会耗费大量的相关成本。如果有标记数据规模小，则难以获得较高的识别精度。

针对有标记语料数据的匮乏，充分利用海量无标记语料数据，本文提出了一种多神经网络协同训练模型(Tri-training for Multiple Neural Network，TMNN)。首先采用 3 种不同的神经网络(LSTM 网络，BLSTM 网络，GRU 网络)初始化为 3 种不同的 NER 识别模型，然后基于 Tri-training 算法，利用少量有标记序列文本数据和大量无标记序列文本数据对 3 种 NER 模型进行协同训练，最后融合 3 种 NER 模型对文本数据进行标注。实验表明该模型在简历命名实体识别上取得了良好的效果。

1 相关工作

1.1 命名实体识别

命名实体识别任务于 1996 年在 MUC-6 会议上首次提出，旨在识别出序列文本数据中的实体类

基金项目：国家自然科学基金面上项目(61671064)。

作者简介：王 栋(1994-)，男，硕士研究生，主要研究方向：命名实体识别、自然语言处理；李业刚(1975-)，男，博士，副教授，硕士生导师，主要研究方向：语言信息处理、机器学习、机器翻译等；张 晓(1995-)，男，硕士研究生，主要研究方向：命名实体识别、自然语言处理。

通讯作者：李业刚 Email：389775268@qq.com

收稿日期：2019-11-10

信息。早期的研究者采用统计机器学习方法对序列文本中的实体类数据进行识别,此类方法需要研究者人工制定相应的语言规则模板。目前,NER 任务的研究赢得了众多学者的青睐与重视。究其原因,一方面 NER 是自然语言处理的关键技术,是信息提取和信息检索的基础。另一方面,随着深度学习的不断发展,将深度学习技术应用在 NER 任务上,也已成为时下学界的研究重点。

深度学习中的循环神经网络(Recurrent Neural Network,RNN)能够捕捉句子的上下文信息,尤其适用于序列任务。随后具有改进结构的长短期记忆网络(Long Short-Term Memory,LSTM)逐渐成为解决序列问题的主流方法,常见的基于深度学习的 NER 模型结构如图 1 所示。文献[4]使用 LSTM 网络和 CNN 网络组成了一种混合结构模型,该模型可以分别获取字符和词级别的特征信息,避免了对特征工程的需求。文献[5]的 LSTM 模型对输入模型的序列处理了两次,第一次以提取文本信息,第二次用来消除歧义。文献[6]将 LSTM 网络和条件随机场(Conditional Random Field,CRF)进行了联合,模型获得了良好的性能。

综合以上学者的研究在分析后可知,将深度学习技术与传统的机器学习方法进行融合所产生的混合模型具有良好的性能。一方面,深度学习技术可以自动获取文本序列的特征信息,减少人工干预。另一方面,用机器学习算法对深度学习识别模型进行优化和校正可以获得更优秀的识别效果。

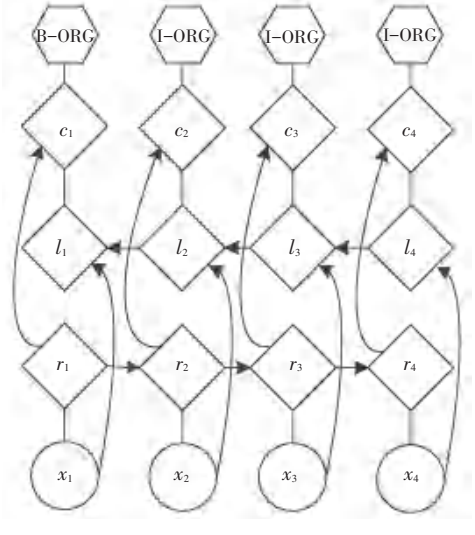

图 1 基于深度学习的命名实体识别图

Fig. 1 Named entity recognition based on deep learning

1.2 LSTM 网络与 BLSTM 网络

长短期记忆网络(Long Short-Term Memory,LSTM)是一种能够捕获序列文本特征信息的 RNN 改进模型,在语音识别[7]、机器翻译[8]、语言建模[9]等多种任务中均有着良好表现。相较于传统的 RNN 模型,LSTM 模型解决了早期 RNN 模型存在的长期依赖问题,同时避免了梯度爆炸和消失的问题。尽管文本序列和语音序列的任务不同,但 LSTM 网络的结构十分适合于处理序列化的数据,LSTM 网络记忆单元结构如图 2 所示。LSTM 网络具有 3 种门结构,分别是:输入门、遗忘门、输出门。其中,输入门控制当前输入的信息,遗忘门决定保留上层传来的信息量,输出门控制网络输出的信息。通过 3 种门结构,LSTM 可以有效地控制记忆信息。LSTM 网络的公式描述如下所示:

$$f_t = \sigma(\boldsymbol{W}_f \cdot [h_{t-1}, s_t] + b_f), \tag{1}$$

$$i_t = \sigma(\boldsymbol{W}_i \cdot [h_{t-1}, s_t] + b_i), \tag{2}$$

$$\tilde{C}_t = \tanh(\boldsymbol{W}_c \cdot [h_{t-1}, s_t] + b_c), \tag{3}$$

$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t, \tag{4}$$

$$o_t = \sigma(\boldsymbol{W}_o \cdot [h_{t-1}, s_t] + b_o), \tag{5}$$

$$h_t = o_t \odot \tanh C_t. \tag{6}$$

其中,$\boldsymbol{W}_f, \boldsymbol{W}_i, \boldsymbol{W}_o$ 为权值矩阵;f, i, o 分别为遗忘门、输入门、输出门;C 为记忆单元;$\tilde{C}_t$ 为候选向量表示当前的细胞状态;h_t 是 t 时刻的隐藏层的输出。

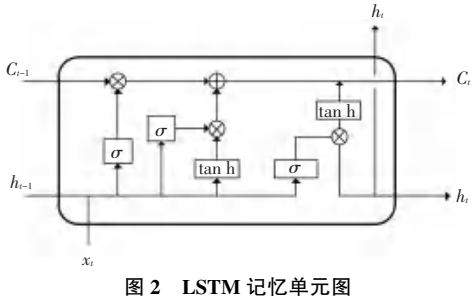

图 2 LSTM 记忆单元图

Fig. 2 LSTM memory unit diagram

双向长短期记忆网络(Bidirectional Long Short-Term Memory,BLSTM)的设计原理是将一个前向的 LSTM 网络和一个后向的 LSTM 网络连接到同一输出,以此来获取前向和后向的信息[10]。相较于单向的 LSTM 网络,该网络结构可以更充分地利用序列化文本的上下文信息,双向网络的输出为前向和后向网络的输出拼接,该种输出的公式描述如下所示:

$$B_t = [\overrightarrow{h}_t, \overleftarrow{h}_t]. \tag{7}$$

1.3 GRU

门控循环单元(Gated Recurrent Unit,GRU)也属于 RNN 网络的一种变体网络模型[11]。该网络具有更简洁的门结构,相较于 LSTM 网络依靠 3 种门

结构来实现信息的更新与保留，GRU 网络则依靠更新门与复位门来控制记忆信息，更新门负责控制 $t-1$ 时刻时记忆单元存储的信息量，复位门负责结合当前输入的信息与历史记忆信息，2 种门结构共同决定了 GRU 网络的输出表示。GRU 网络记忆单元结构图如图 3 所示。GRU 网络具有结构更简单，参数更少，计算速度更快的优势。GRU 网络的公式描述如下所示：

$$r_t = \sigma(\boldsymbol{W}_r \cdot [h_{t-1}, s_t]), \tag{8}$$

$$z_t = \sigma(\boldsymbol{W}_z \cdot [h_{t-1}, s_t]), \tag{9}$$

$$\tilde{h}_t = \tanh(\boldsymbol{W}_h \cdot [r \odot h_{t-1}, s_t]), \tag{10}$$

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t. \tag{11}$$

其中，$\boldsymbol{W}_r$、$\boldsymbol{W}_z$、$\boldsymbol{W}_h$ 为权重矩阵；z_t 为更新门；r_t 为复位门；σ 为激活函数；$\odot$ 表示向量之间的点乘运算；h_t 为 t 时刻 GRU 的输出表示。

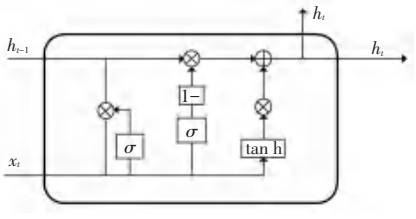

图 3　GRU 记忆单元图

Fig. 3　GRU memory unit diagram

1.4　协同训练

协同训练算法是一类典型的半监督学习算法，可以将无标记的数据自动训练为有标记的数据，使得海量无标记的数据得以利用，减少了对有标记数据的依赖，训练的过程中仅仅使用了少量的有标记数据。文献[12]提出的 Co-training 算法通过在 2 个视图上利用有标记的数据分别初始化分类器，并使用 2 个分类器对无标记的数据进行标注，同时将每个分类器标注后的数据作为另一个分类器的输入，从而达到更新训练集的目的。随后，文献[13]提出了 Tri-training 算法，增加了第三个分类器。该算法通过对有标记数据集重复取样生成训练集，由此训练得到 3 个分类器。在随后的训练过程中，3 个分类器中用到的训练数据皆由其他两个分类器合作提供。在对数据进行标注时，Tri-training 算法不同于 Co-training 算法仅仅使用一个分类器进行标注，而是采用投票法将 3 个分类器联合起来对数据进行标注。上述过程不再需要分类器的差异性，因此使得 Tri-training 算法具有了更强的实用性。

综合上述研究，本文融合神经网络模型和 Tri-training 算法，提出了多神经网络协同训练模型。首先选取 3 种不同的神经网络模型，作为 Tri-training 算法的 3 个初始模型，为使初始识别模型具有一定的差异性，本文实验中分别选取了 LSTM 网络、BLSTM 网络及 GRU 网络。TMNN 模型在训练的过程中使用了少量的有标记数据和大量无标记数据，克服了缺乏有标记语料的困难。

2　多神经网络协同训练模型

基于上述的相关工作，本文提出一种多神经网络协同训练模型 TMNN，首先选取了 3 种不同的神经网络模型，彼此都具有一定的差异性。模型训练时使用少量有标记的数据 L 以及大量未标记的数据 U 对 3 种初始模型进行协同训练。首先对 L 进行重复采样，得到 3 个不同的训练集 $L_{1,2,3}$。然后利用训练集分别训练 3 种初始识别模型 $H_{1,2,3}$。

在协同训练的过程中，各神经网络识别模型所更新的有标记数据由其余两个识别模型协同提供。假设一个无标记的数据 x，如果 H_2 和 H_3 对 x 的识别相同，则认为该识别结果准确。如果 H_1 对 x 的识别与 H_2 和 H_3 不相同，则该识别结果不准确。每一轮训练，待标记的数据从 U 中获得，直至 U 为空。训练结束后，获得的模型 $H_{1,2,3}$ 基于投票法对数据进行重新标注，其计算公式如下所示：

$$H(x) = \underset{y \in label}{\operatorname{argmax}} \frac{P_i(L) \times \sum_{i=1}^{3} \theta(y, H_i(x))}{\sum_{i=1}^{3} P_i(L)}. \tag{12}$$

其中，L 为少量的有标记数据集；P 为初始模型的识别精度；θ 为用于判断标注结果的函数。

TMNN 的算法步骤详见如下。

算法：多神经网络协同训练模型 TMNN

输入：有标记数据集 L；无标记数据集 U；初始模型 H_1、H_2、H_3

输出：最后的 NER 结果

Step 1　$L \rightarrow L_1$、L_2、L_3、T

$[H_1, L_1] \rightarrow C_1$，$[H_2, L_2] \rightarrow C_2$，$[H_3, L_3] \rightarrow C_3$

Step 2　Repeat

$T \xrightarrow{Bootstrap} L_1^i$，$L_2^i$，$L_3^i$

从 U 中选取待标记数据至 U^i

利用 C_1^i，C_2^i，C_3^i 进行标记，并且得到更新数据集 V_1、V_2、V_3

$L_1^i \cup V_1 \rightarrow L_1^{i+1}$、$L_2^i \cup V_2 \rightarrow L_2^{i+1}$、$L_3^i \cup V_3 \rightarrow L_3^{i+1}$。其中，$L_{1,2,3}^{i+1}$ 为新的训练集

$[H_1, L_1^{i+1}] \rightarrow C_1^{i+1}$

$[H_2, L_2^{i+1}] \rightarrow C_2^{i+1}$

$[H_3, L_3^{i+1}] \rightarrow C_3^{i+1}$

$L_1^{i+1} \cup L_2^{i+1} \cup L_3^{i+1} \rightarrow T$

$H_{1,2,3}$ 依据投票法对 T 中数据重新标注

until U 为空

3 实验

3.1 实验设置

本文在新浪财经随机选取 1 024 份上市公司的高管简历中文文本数据作为实验的语料,该语料包括了姓名、学历、籍贯、毕业院校等 8 种实体信息,8 种实体描述见表 1。数据集的规模为 16 565 条,实验过程中对语料随机选取 20%作为测试集,20%作为有标记的训练集 L, 60%的数据集作为未标注集 U。为了避免神经网络模型输入差异性对实验效果的影响,实验的过程中统一使用[-0.25,0.25]区间内随机初始化的方式得到的字向量作为 3 种初始化模型的输入。

表 1 简历实体类别表

Tab. 1 Resume entity category table

实体	示例
民族	如“汉族”
学校	如“山东大学”
籍贯	如“山东济南”
姓名	如“张三”
职位	如“经理”
出生日期	如“1994 年”
性别	“男”“女”
学历	如“本科”

3.2 评价指标与标注策略

本文采用 *BIO* 的标注策略,该标注策略中 B 表示实体的起始部, I 表示的是实体的非起始部, O 表示其他。并且采用准确率(*Precision*)、召回率(*Recall*)、F_1 值作为模型识别性能评价指标。本文评价指标的计算公式如下所示:

$$P = \frac{\text{正确识别的命名实体个数}}{\text{实际识别的命名实体个数}} \times 100\%, \quad (13)$$

$$R = \frac{\text{正确识别的命名实体个数}}{\text{标注的命名实体总数}} \times 100\%, \quad (14)$$

$$F_1 = \frac{2PR}{P+R} \times 100\%. \quad (15)$$

3.3 实验结果与分析

为了分析 TMNN 模型的性能,本文对比分析了 TMNN 模型与传统协同训练方法和 3 种单一神经网络 NER 模型(LSTM-CRF 模型、GRU-CRF 模型、BLSTM-CRF 模型)在相同数据集上的识别效果。其中,传统协同训练选用条件随机场 CRF 作为初始分类器,实验结果见表 2。从表 2 中可以看出,多神经网络协同训练模型的识别质量远高于传统协同训练算法。究其原因,传统协同训练对特征工程和语言学规则具有较高的依赖性,模型泛化性能较差,识别质量低,而本文提出的 TMNN 的初始识别模型分别选用了 3 种不同的神经网络模型,这三种神经网络可以自动提取文本数据的内部特征,避免了人工添加过多的特征工程和语言学规则,从而显著地提高了 NER 的精度。相较于 3 种单一的神经网络模型,多神经网络协同训练模型 TMNN 通过使用不同的神经网络提取到具有差异化的特征,并且通过协同训练模型,达到持续优化模型的目的,对比 3 种单一的神经网络模型的 F_1 值分别有了 3.35%、2.58%、1.25%的提高,系统性能显著提升。

表 2 实验结果对比表

Tab. 2 Experimental result comparison table

模型	准确率/%	召回率/%	F_1/%
协同训练	-	-	80.55
LSTM-CRF	87.37	84.80	86.07
GRU-CRF	87.39	86.30	86.84
BLSTM-CRF	89.45	86.93	88.17
TMNN	90.78	88.10	89.42

图 4 给出本文提出的 TMNN 模型和模型所使用的 LSTM 网络、BLSTM 网络、GRU 网络在训练过程中识别精度的变化趋势。从图 4 中可以看出, 当 *iteration* 大于 5 时,4 种模型的 F_1 值趋于稳定。并且在训练过程中,多神经网络协同训练模型的性能都要优于其他三种神经网络模型。综合上述实验结果可以看出,多神经网络协同训练模型具有良好的稳定性和系统性能,并且模型的实用性也有了显著的提高。

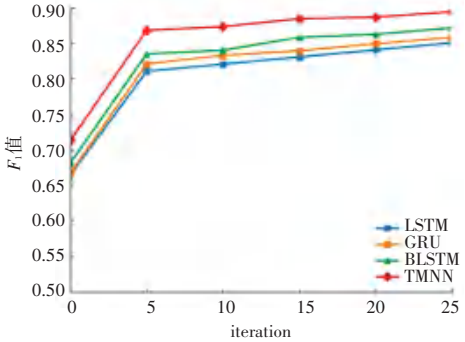

图 4 TMNN 和 3 种神经网络识别效果对比图

Fig. 4 Comparison of recognition effects between TMNN and three neural network

4　结束语

本文提出的多神经网络协同训练模型，将神经网络和协同训练算法各自的优势相结合，使用神经网络模型自动提取序列文本的内部特征，有效利用了协同训练算法对无标记数据进行训练，达到了利用大量无标记数据进行命名实体识别的目的，TMNN 模型降低了对人工标记数据的需要，模型的系统实用性得到了提高。实验表明，本文模型具有良好的系统性能，在实际应用中优于已有的其它模型。

随着专业领域语料越来越多，识别专业领域命名实体的需求越来越大。下一步将探索专业领域的命名实体识别方法，以提高命名实体识别的跨领域适应性，进一步增强模型对于专业领域文本数据的学习能力，从而达到更好的专业领域识别效果，提高命名实体识别的应用范围。

参考文献

[1] 吴文涛，李培峰，朱巧明. 基于混合神经网络的实体和事件联合抽取方法[J]. 中文信息学报，2019，33(8)：77.

[2] 卜质琼，郑波尽. 基于 LDA 模型的 Ad hoc 信息检索方法研究[J]. 计算应用研究，2015，32(5)：1369.

[3] 赵冬梅，李雅，陶建华，等. 基于协同过滤 Attention 机制的情感分析模型[J]. 中文信息学报，2018，32(8)：128.

[4] CHIU J P C，NICHOLS E. Named entity recognition with bidirectional LSTM-CNNs[C]//Transactions of the Association for Computational Linguistics. Stroudsburg：ACL，2016，4：357.

[5] HAMMERTON J. Named entity recognition with long short-term memory [C]//Proceedings of the 7th conference on Natural Language Learning. Edmonton，AB，Canada：ACL，2003：172.

[6] LAMPLE G，BALLESTEROS M. Neural architectures for named entity recognition [C]// Conference of the North American Chapter of the Association for Computational Linguistics：Human Language Technologies. Stroudsburg：ACL，2016：260.

[7] GRAVES A，MOHAMED A R，HINTON G. Speech recognition with deep recurrent neural networks[C]// Proceedings of the 2013 IEEE International Conference on Acoustics，Speech and Signal Processing. Vancouver，BC，Canada：ieee，2013：6645.

[8] 谭敏，段湘煜，张民. 基于领域特征的神经机器翻译领域适应方法[J]. 中文信息学报，2019，33(7)：56.

[9] MIKOLOV T，KARAFIÁT M，BURGET L，et al. Recurrent neural network based language model[C]// Proc of the 11th Annual Conference of the International Speech Communication Association. Makuhari，Japan：Interspeech，2010：1045.

[10] 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.

[11] 孙媛，王丽客，郭莉莉. 基于改进词向量 GRU 神经网络模型的藏语实体关系抽取[J]. 中文信息学报，2019，33(6)：35.

[12] 周志华，王珏. 机器学习及其应用[M]. 北京：清华大学出版社，2009.

[13] ZHOU Z H，LI M. Tri-training：Exploiting unlabeled data using three classifiers[J]. IEEE Transactions on Knowledge and Data Engineering，2005，17(11)：1529.

(上接第 122 页)

[3] CARUANA R. Multitask learning[M]. Machine Learning，1997，28(1)：41.

[4] YANG H Q，KING I，LYU M. Multi-task learning for one-class classification [C]// Proceedings of the International Joint Conference on Neural Networks (IJCNN). Barcelona，Spain：IEEE，2010：1.

[5] HE X，MOUROT G，MAQUIN D，et al. Multi-task learning with one-class SVM[J]. Neurocomputing，2014，133(6)：416.

[6] XUE Yongjian，BEAUSEROY P. Multi-task learning for one-class SVM with additional new features[C]// Proceedings of the 23rd International Conference on Pattern Recognition (ICPR). Cancún，Mexico：dblp，2016：1571.

[7] XU Shuo，AN Xin，QIAO Xiaodong，et al. Multi-task least-squares support vector machines [J]. Multimedia Tools Applications，2014，71(2)：699.

[8] LI Ya，TIAN Xinmei，SONG Mingli，et al. Multi-task proximal support vector machine[J]. Pattern Recognition，2015，48(10)：3249.

[9] FUNG G，MANGASARIAN O L. Proximal support vector machine classifiers [C]// Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco：ACM，2001：77.

[10] XIE Xijiong，SUN Shiliang. Multitask twin support vector machines [M]//HUANG T，et al. Proceedings of the 19th International Conference on Neural Information Processing ICONIP - Volume Part II. Heidelberg/Berlin：Springer-Verlag，2012：341.

[11] MEI Benshan，XU Yitian. Multi-task least squares twin support vector machine for classification [J]. Neurocomputing，2019，338：26.

[12] BURGES C J. A tutorial on support vector machines for pattern recognition[J]. Data Mining Knowledge Discovery，1998，2(2)：121.