doi:10.3969/j.issn.1673-9833.2019.02.009

基于自注意力深度学习的微博实体识别研究

徐 啸，朱艳辉，冀相冰

（湖南工业大学 计算机学院，湖南 株洲 412007）

摘 要：命名实体识别是自然语言处理的重要基础，随着神经网络的快速发展，深度学习的各种方法被应用于文本处理的各个方向。引入自注意力机制，结合深度学习方法，提出一种基于自注意力的双向长短期记忆条件随机场（SelfAtt-BiLSTM-CRF）方法来识别微博中的实体，利用自注意力机制，获取词与词之间的依赖关系，进一步提高模型的识别能力。实验表明，所提出的方法取得了较好的识别效果。

关键词：实体识别；自注意力；深度学习；神经网络

中图分类号：TP391 **文献标志码**：A **文章编号**：1673-9833(2019)02-0048-05

引文格式：徐 啸，朱艳辉，翼相冰 . 基于自注意力深度学习的微博实体识别研究 [J]. 湖南工业大学学报，2019，33(2)：48-52.

Research on Microblog Entity Recognition Based on Self-Attention Deep Learning

XU Xiao，ZHU Yanhui，JI Xiangbing

（College of Computer，Hunan University of Technology，Zhuzhou Hunan 412007，China）

Abstract：Named Entity Recognition (NER) is an important basis for natural language processing. With the rapid development of neural networks, various methods of deep learning have been pervasively applied to text processing. By introducing the self-attention mechanism, as well as combined with deep learning method, a self-attention-based bidirectional long-term and short-term memory conditional random field (SelfAtt-BiLSTM-CRF) method has been proposed to identify entities in microblogs. By utilizing the self-attention mechanism to obtain the dependency between words, this method helps to further improve the recognition ability of the model. Experiments show that the method proposed in this paper has achieved a satisfying recognition effect.

Keywords：entity recognition；self-attention；deep learning；neural networks

1 研究背景

命名实体识别（named entity recognition，NER）是自然语言处理（natural language processing，NLP）的重要基础，NER 对于信息抽取和实体链接等高级应用非常重要和有用。NER 最初在 1995 年 MUC-6 上提出的，传统的 NER 是对人名、地名、组织机构名的识别 [1]。20 世纪 90 年代初期，国内一些学者开始了对中文命名实体识别的研究。传统的命名实体识别方法有最大熵模型 [2]、隐马尔可夫模型 [3]、随机条件场模型 [4] 等，但这些方法都需要人工定义特征模板。随着深度学习技术的发展，深度神经网络（deep

收稿日期：2018-10-24

基金项目：国家自然科学基金资助项目（61402165），湖南省自然科学基金资助项目（2018JJ2098），湖南工业大学重点基金资助项目（17ZBLWT001KT006）

作者简介：徐 啸（1992-），男，江苏泰兴人，湖南工业大学硕士生，主要研究方向为自然语言处理，E-mail：1151536817@qq.com

neural network，DNN）、卷积神经网络（convolutional neural network，CNN）、循环神经网络（recurrent neural network，RNN）、长短期神经网络（long short term memory neural network，LSTM）、双向长短期神经网络（bi-direction long short term memory neural network，BiLSTM）等模型都被提出并应用于各个方面，在命名实体识别领域也取得了不错的结果。

微博是近年来发展较快且影响较大的网络全民媒体平台形式[5]，目前，对于长文本命名实体识别的研究逐渐趋于成熟，而对于微博这种短文本命名实体识别的准确率还有待进一步提高。与长文本相比，微博文本字数限定在 140 字以内，一般平均每条微博字数为 50 字左右，属于短文本，特征稀疏。此外，微博文本较为随意，其中的新词不断涌现，更增加了识别命名实体的难度。

近年来，注意力模型结合深度学习技术被广泛应用在语音识别、图像处理和自然语言处理等领域。注意力机制的核心思想是从众多信息中选择出对当前任务目标更关键的信息。2017 年 6 月 google 机器翻译团队提出了一种自注意力（Self-Attention）机制[6]。自注意力机制通常不会使用其他额外的信息，但是它能使用自注意力关注本身并从句子中抽取更多相关信息。本文提出了一种融合自注意力机制的深度神经网络结构，用来处理微博中的命名实体识别，该结构能在一定程度上提高识别命名实体的效果。

2　基于自注意力深度学习的命名实体识别模型

构建的用于命名实体识别的 SelfAtt-BiLSTM-CRF 网络模型整体结构如图 1 所示。训练语料的文本向量作为输入，经过 BiLSTM 层、Self-Attention 层和 CRF 层训练模型的参数。

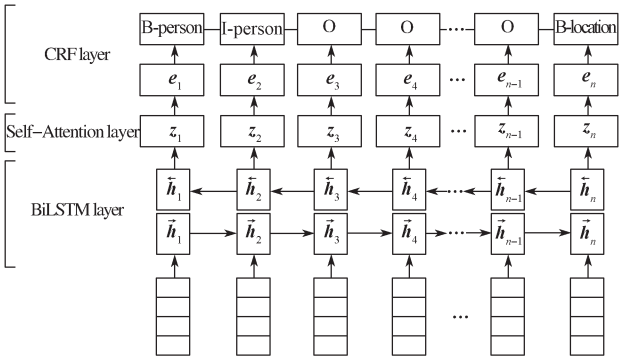

图 1　SelfAtt-BiLSTM-CRF 神经网络模型

Fig. 1　SelfAtt-BiLSTM-CRF neural networks model

2.1　词向量的文本表示

词向量[7]（亦称分布式词表示）可以从一个大的未标注语料库中捕捉到词的语义和句法信息，和 BOW（bag-of-word）相比，词向量具有低维性和密集性等特点。近年来，Word2vec[8] 和 GloVe[9] 被广泛应用于 NLP 的各个领域。本研究使用 Word2vec 从两个方面获取词向量。首先，一个词语是不是命名实体与它所在的上下文内容是有关联的，所以利用语言模型在大规模无标注的语料上训练词向量，Word2vec 有 2 种实现模型：Skip-gram 和 CBOW。Skip-gram 由当前的词语来预测上下文的词语，而 CBOW 则由上下文的词语来预测当前的词语[10]。本文采用 CBOW 模型来训练语料，得到的词向量记为 $\boldsymbol{x}_t^{\mathrm{e}}$。其次，词语中的每一个字对命名实体的识别也是有影响的，这些字符级特征也包含着实体的丰富结构特征，如前缀和后缀。在一些噪声较多的文本中，人们更倾向于使用缩写或者昵称来描述实体，这些特征都有助于提高命名实体识别的效果。使用 Word2vec 获取每个字的字符级向量，记为 $\boldsymbol{x}_i^{\mathrm{c}}$。最后将词语向量与其对应的字符级向量串联，记为 $\boldsymbol{x}_t=\left[\boldsymbol{x}_t^{\mathrm{e}};\boldsymbol{x}_t^{\mathrm{c}}\right]$，作为 SelfAtt-BiLSTM-CRF 网络模型的输入。

2.2　BiLSTM 层

基于 LSTM 的神经网络在序列标注问题上表现良好，主要是因为它能够获取上下文的信息。然而，单向的 LSTM 中的隐藏状态只能接收前文的信息，通过 BiLSTM 能让隐藏状态获取到过去和未来的上下文信息。在句子中，命名实体的识别准确性取决于词的上下文，每一句话中命名实体的前后 2 个词语对预测标签都有很大作用，如果能够获得文本中过去和未来的上下文信息，对命名实体识别有很大的帮助。BiLSTM 模型结构如图 2 所示。图中 $\boldsymbol{x}_i$ 表示模型在 i 时刻的输入，前向 LSTM 在 i 时刻的输出为 $\overrightarrow{\boldsymbol{h}}_i$，反向 LSTM 在 i 时刻的输出为 $\overleftarrow{\boldsymbol{h}}_i$，则 BiLSTM 在 i 时刻的输出表示定义为 $\boldsymbol{h}_i=\left[\overrightarrow{\boldsymbol{h}}_i;\overleftarrow{\boldsymbol{h}}_i\right]$，即将 $\overrightarrow{\boldsymbol{h}}_i$ 与 $\overleftarrow{\boldsymbol{h}}_i$ 拼接在一起。

这一层的输入是词向量训练层输出的词向量 $\boldsymbol{x}_i$，表示为 $(x_1, x_2, \cdots, x_n)$，该层的输出是每个输入向量的隐藏序列，表示为 $(\boldsymbol{h}_1, \boldsymbol{h}_2, \cdots, \boldsymbol{h}_n)$。多个句子的词向量组成矩阵记为 $\boldsymbol{H}$。

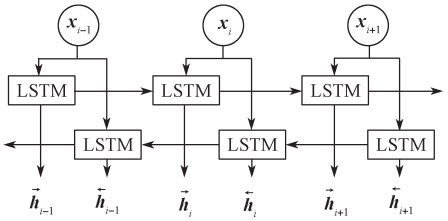

图 2　BiLSTM 模型结构

Fig. 2　Structure of BiLSTM model

2.3 Self-Attention 层

自注意力机制已经成功地应用于许多任务中，包括阅读理解、抽象摘要、文本蕴含、机器翻译等。本研究构建的 Self-Attention 层结构如图 3 所示。上一层 BiLSTM 在每个时刻的输出构成一个向量矩阵 $\boldsymbol{H}$，与通过训练而得的 3 个权重矩阵 $\boldsymbol{W}_i^Q$、$\boldsymbol{W}_i^K$、$\boldsymbol{W}_i^V$进行 h 次的映射，对每一次映射得到的 $\boldsymbol{Q}$、$\boldsymbol{K}$ 和 $\boldsymbol{V}$，并行地执行注意力函数生成输出，并将这 h 个输出拼接在一起，再进行一次注意力映射，得到最终的输出矩阵 $\boldsymbol{Z}$，每个时序为 $\boldsymbol{z}_t$。

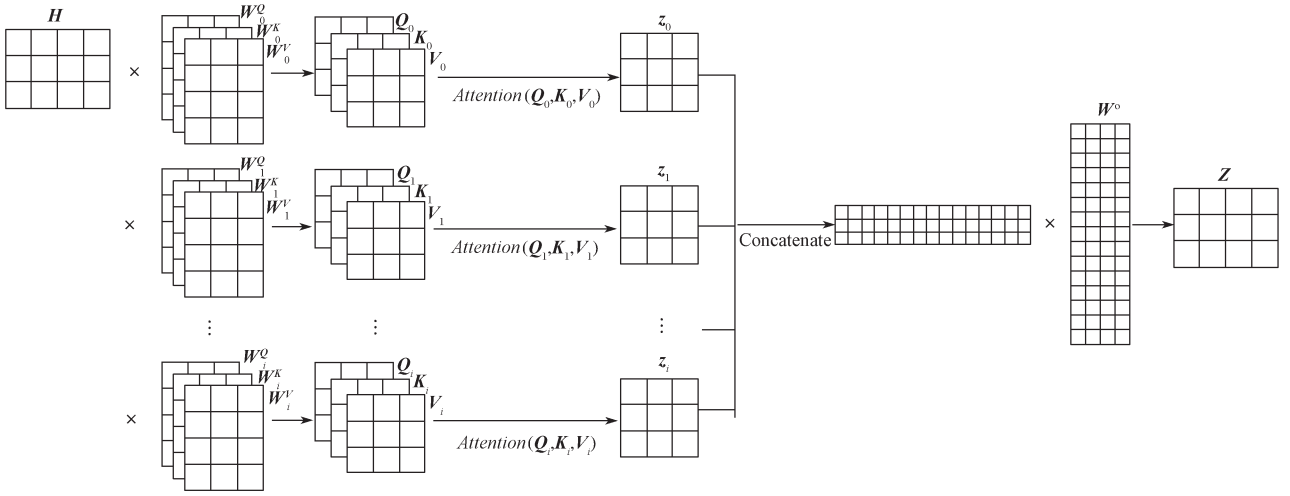

图 3 Self-Attention 层结构图

Fig. 3 Self-Attention layer structure diagram

最常用的 2 个注意力函数是加法注意力函数和点乘注意力函数，这里使用的是点乘注意力函数，具体算法步骤如下。

步骤 1 根据 Query 和 Key 计算 2 个词语向量的相似性或者相关性。将 Query 和 Key 中每个词语的向量进行点乘。

步骤 2 对步骤 1 的点乘结果进行 *softmax* 归一化处理，计算出权重系数。

步骤 3 根据权重系数对 Value 进行加权求和。

其中归一化处理过程中，将点乘的结果送入 *softmax* 函数后，就会进入函数的极大值区间，导致梯度计算时得到的梯度值极小，模型的训练变慢，所以为了中和这种影响，除了常规的点乘注意力机制外，添加上一个伸缩因子 $\frac{1}{\sqrt{d_k}}$。该函数实现了从 Query 和一系列键值对 Key-Value 到一个输出的映射，具体的公示表示为

$$Attention(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=softmax\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V}。\tag{1}$$

2.4 CRF 层

Self-Attention 层输出每个词对应各个标注的分数，选择分数最高的作为该词的标注，但是在预测过程中，句首词语的标注不可能是 I 标签。为了防止这种非法标注的出现，通过添加 CRF 层为最终的预测标注添加限制。用 tanh 函数来预测每个可能的标注作为网络输出分数的单词的置信分数：

$$\boldsymbol{e}_t=\tanh(\boldsymbol{W}_e\boldsymbol{z}_t)。\tag{2}$$

式中 $\boldsymbol{W}_e$ 为训练时学习到的参数矩阵。

不对标记决策进行独立建模，而是在所有可能的标记路径中添加 CRF 层来解码最佳标记路径。记 $\boldsymbol{P}$ 为网络输出的分数矩阵，$\boldsymbol{P}$ 的第 t 列为向量 $\boldsymbol{e}_t$，P_{i,y_i} 为矩阵 $\boldsymbol{P}$ 中的一个分数，对应句子中第 i 个词语可能为标签 y_i 的分数。此外，引入一个标签转移矩阵 $\boldsymbol{T}$，T_{y_{i-1},y_i} 代表的是从标签 y_{i-1} 成功转移到标签 y_i 的分数。给定句子的词语序列 X，由模型得到预测序列 y，则预测结果正确的可能性的量化定义如式（3）所示：

$$s(X,y)=\sum_{i=1}^{n}\left(T_{y_{i-1},y_i}+P_{i,y_i}\right)。\tag{3}$$

使用 *softmax* 函数对所有可能的标记路径进行归一化处理，从而得到路径 y 的条件概率：

$$p(y\mid X)=\frac{\mathrm{e}^{s(X,y)}}{\sum\limits_{\tilde{y}\in Y}\mathrm{e}^{s(X,\tilde{y})}}。\tag{4}$$

式中：$\tilde{y}$ 为真实的标记值；Y 为所有可能的标注序列。

在训练阶段，模型的目标是使正确标注的序列的概率最大。使用 Viterbi 算法求得所有序列上打分最高的序列 B，并将其作为最终的命名实体识别的标注结果，如式（5）所示：

$$B=\underset{\tilde{y}}{\operatorname{argmax}}\ s(X,\tilde{y})。\tag{5}$$

3　实验

3.1　实验语料与标注

本文使用李刚等 [11] 提供的 3 000 条微博训练语料和 2 000 条微博测试语料，使用八爪鱼爬虫软件爬取微博内容，记为 Text1。此外，本文还从搜狗输入法词库中下载人名地名机构名，下载的文件是 scel 格式，需要将其转换成 txt 格式，然后与部分微博和新闻内容混合成一个文本，记为 Text2。

本研究使用 BIO 模式来标注训练数据，B 表示实体开始，I 表示实体中间和实体结束，O 表示不是实体。具体实体类别及其标注方式如表 1 所示。

表 1　实体标注方法

Table 1　Entity labeling method

实体类别	标注编码
人名	B-person/I-person
地名	B-location/I-location
组织机构名	B-organization/I-organization
杂项	B-misc/I-misc
非实体	O

3.2　环境配置

实验软硬件环境配置如表 2 所示。

表 2　实验软硬件环境

Table 2　Software and hardware experimental environment

项　目	环　境
系统	Ubuntu16.04 LTS
GPU	NVIDIA Quadro K1200
硬盘	1 TB
内存	16 GB
Python 版本	Python3.7.0
TensorFlow 版本	TensorFlow1.4.0

3.3　实验与结果分析

本文采用准确率 P(precision)、召回率 R(recall)、F 值（F-Measure）作为评价标准，在本文实验中的具体定义如下：

$$\begin{cases} P=\dfrac{A}{A+B}, \\ R=\dfrac{A}{A+C}, \\ F=\dfrac{2\times P\times R}{P+R}。 \end{cases} \tag{6}$$

式中：A 为正确识别命名实体个数；B 为错误识别命名实体个数；C 为未识别出来的命名实体个数。

准确率表示正确识别出的命名实体个数占识别出来的命名实体总个数的比例。召回率表示正确识别出的命名实体个数占语料中总的命名实体个数的比例，也就是正确识别的数量程度。为了提高识别效果，减少识别方法的选择对召回率和准确率的意外影响，需对召回率和准确率进行综合权衡，取两者的加权平均值，即为 F 值。

3.3.1　参数优化实验

课题组使用 Word2vec 对 Text1 训练词向量，具体训练参数如表 3 所示。

表 3　Word2vec 参数表

Table 3　Word2vec parameter table

参　数	取　值
算法	CBOW
窗口大小	10
向量维数	300
最小词频	2
学习速率	0.001
迭代次数	10

此外，将表 3 中窗口大小 10 更改为 7，对 Text2 训练字符向量。分别使用词向量和词向量 + 字符向量作为 SelfAtt-BiLSTM-CRF 模型的输入进行实验。实验结果如表 4 所示。

表 4　有无字符级向量的实验结果对比

Table 4　Comparison of experimental results with or without character level vectors　　%

方　法	P	R	F
词向量	62.73	63.44	63.08
词向量 + 字符向量	73.76	68.67	71.12

结果表明，有字符向量的情况下，P、R、F 值均有显著的提升，F 值提升了 8.04%，由此得知，在文本中噪声多的情况下，使用词向量和字符向量的混合向量，能够提高模型的鲁棒性。

在 Self-Attention 层中，对不同的映射次数进行实验，得到如图 4 所示的实验结果。结果表明，当映射次数 h 为 5 时，识别的效果最好。

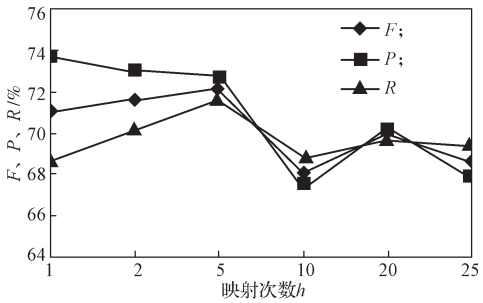

图 4　Self-Attention 层中不同映射次数下实验结果对比

Fig. 4　Comparison of experimental results under different mapping times in Self-Attention layer

3.3.2　对比实验与分析

为了验证方法的有效性，使用相同的训练语料和测试语料，分别对 CRF、BiLSTM、BiLSTM-CRF

和 SelfAtt-BiLSTM-CRF 方法进行实验。表 5 分别给出了这 4 种模型的识别效果以及每个模型每轮训练的时间。

表 5 不同方法的实验结果对比

Table 5 Comparison of experimental results of different methods

模型名称	*P*/%	*R*/%	*F*/%	*t*/s
CRF	77.05	52.84	62.69	51
BiLSTM	69.39	67.98	68.68	54
BiLSTM-CRF	72.78	68.70	70.68	58
SelfAtt-BiLSTM-CRF	72.75	71.54	72.14	60

由表 5 的结果可以看出，SelfAtt-BiLSTM-CRF 模型对微博命名实体识别的 *F* 值高于其他模型的结果，此外，各个模型每轮训练所需的时间都在 50~60 s 浮动，计算速度相差不大。综合实验结果表明，基于自注意力深度学习的微博命名实体识别方法优于其他方法。

4 结语

本文提出了一种基于自注意力的深度学习模型，利用词与词之间的联系来提升微博中命名实体识别的准确率。网络中的新词层出不穷，这提升了微博中命名实体识别的难度，在下一步的研究工作中，课题组将针对微博新词的命名实体识别展开研究，以获得更好的识别性能。

参考文献：

[1] 张祥伟，李 智 . 基于多特征融合的中文电子病历命名实体识别 [J]. 软件导刊，2017，16(2)：128-131.
ZHANG Xiangwei，LI Zhi. Chinese Electronic Medical Record Named Entity Recognition Based on Multi-Feature Fusion[J]. Software Guide，2017，16(2)：128-131.

[2] BORTHWICK A. A Maximum Entropy Approach to Named Entity Recognition[D]. New York：New York University，1999.

[3] SUN J，GAO J，ZHANG L，et al. Chinese Named Entity Identification Using Class-Based Language Model[C]// 19th International Conference on Computational Linguistics. Taibei：Association for Computational Linguistics，2002：1-7.

[4] MAO X，HE S K，BAO S C，et al. Chinese Word Segmentation and Named Entity Recognition Based on Conditional Random Fields[C]//Proceedings of Sixth SIGHAN Workshop on Chinese Language Processing. Hyderabad：[s. n.]，2008：90-93.

[5] 邱泉清，苗夺谦，张志飞 . 中文微博命名实体识别 [J]. 计算机科学，2013，40(6)：196-198.
QIU Quanqing，MIAO Duoqian，ZHANG Zhifei. Named Entity Recognition on Chinese Microblog[J]. Computer Science，2013，40(6)：196-198.

[6] VASWANI A，SHAZEER N，PARMAR N，et al. Attention Is All You Need[C]//Advances in Neural Information Processing Systems 30. Long Beach：NIPS，2017：5998-6008.

[7] MIKOLOV T，YIH W，ZWEIG G. Linguistic Regularities in Continuous Space Word Representations[C]//Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics：Human Language Technologies. Atlanta：Association for Computational Linguistics，2013：746-751.

[8] MIKOLOV T，CHEN K，CORRADO G，et al. Efficient Estimation of Word Representations in Vector Space[EB/OL]. [2018-07-20]. https://arxiv.org/abs/1301.3781.

[9] 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 (EMNLP). Doha：Association for Computational Linguistics，2014：1532-1543.

[10] 冯艳红，于 红，孙 庚，等 . 基于 BLSTM 的命名实体识别方法 [J]. 计算机科学，2018，45(2)：261-268.
FENG Yanhong，YU Hong，SUN Geng, et al. Named Entity Recognition Method Based on BLSTM[J]. Computer Science，2018，45(2)：261-268.

[11] 李 刚，黄永峰 . 一种面向微博文本的命名实体识别方法 [J]. 电子技术应用，2018，44(1)：118-120.
LI Gang，HUANG Yongfeng. An Approach to Named Entity Recognition Towards Micro-Blog[J]. Application of Electronic Technique，2018，44 (1)：118-120.

（责任编辑：申 剑）