Computer Science and Application 计算机科学与应用, 2021, 11(3), 628-634
Published Online March 2021 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2021.113064

Hans汉斯

基于深度学习的命名实体识别算法

陈　娟[1]，王卓薇[2]，程良伦[3]

[1]广东省信息物理融合系统重点实验室，广东 广州
[2]广东工业大学，广东 广州
[3]广东工业大学计算机学院，广东 广州
Email: 1269865878@qq.com

收稿日期：2021年2月23日；录用日期：2021年3月17日；发布日期：2021年3月24日

摘　要

命名实体识别(Named Entity Recognition, NER)的定义是从自由文本中识别出属于预定义类别的文本片段(如人名、地理位置名、机构组织名等)。命名实体识别一直是许多自然语言应用的基础，例如问题回答、提取文本摘要和知识库建立。早期的NER系统在实现良好性能方面取得了巨大的成功，其代价是人类工程在设计特定领域的特征和规则方面付出的代价。近年来，非线性处理的连续实值向量表示和语义组合使得深度学习在命名实体识别系统中发挥了很好的作用。在本文中，我们提供了一种基于深度学习的命名实体识别算法。首先我们随机初始化训练集中的每个字特征，并在获取该字典句子中每个字的特征之后，利用周期卷积来得到其固定长度的特征，以此作为句子特征;随后训练数据自动编码器，通过栈式自动编码器得到高层句子的特征;最后通过高层句特征与字特征的组合训练字的标注网络模型来得到未知字的标注值，再进行实体扩展(分类，属性，副标题)，最后利用马尔科夫逻辑网络优化整体识别效果。

关键词

知识图谱，深度学习，实体识别

Named Entity Recognition Algorithm Based on Deep Learning

Juan Chen[1], Zhuowei Wang[2], Lianglun Cheng[3]

[1]Guangdong Provincial Key Laboratory of Cyber-Physics Fusion System, Guangzhou Guangdong
[2]Guangdong University of Technology, Guangzhou Guangdong
[3]School of Computer Science, Guangdong University of Technology, Guangzhou Guangdong
Email: 1269865878@qq.com

Received: Feb. 23rd, 2021; accepted: Mar. 17th, 2021; published: Mar. 24th, 2021

文章引用: 陈娟, 王卓薇, 程良伦. 基于深度学习的命名实体识别算法[J]. 计算机科学与应用, 2021, 11(3): 628-634.
DOI: 10.12677/csa.2021.113064

Abstract

Named entity recognition is the task of identifying rigid indicators from text belonging to predefined semantic types (such as person, location, organizations, and so on). NER has been the basis for many natural language applications, such as question answering, text summarization and machine translation. Early NER systems had great success in achieving good performance at the cost of human engineering in designing domain-specific features and rules. In recent years, deep learning has been used in NER systems through continuous real-valued vector representations and semantic combinations of nonlinear processing, resulting in the most advanced performance. In this article, we provide an entity recognition technique based on deep learning. Firstly, each word feature in the training set is randomly initialized, and the feature of each word in the sentence is obtained based on the dictionary. Then, the fixed-length feature is obtained by periodic convolution of different length of sentence features, which are used as sentence features. Then the data autoencoder is trained to get the features of high-level sentences through the stack autoencoder. Finally, the combination of high level sentence features and word features is used to train the annotation network model of words, and the annotation value of unknown words is obtained based on the annotation model, and then the entity expansion (classification, attribute, subtitle) is carried out. Finally, the overall recognition effect is optimized by using Markov logic network.

Keywords

Knowledge Graph, Deep Learning, Entity Recognition

1. 引言

命名实体识别(Named Entity Recognition, NER)定义是从自由文本中识别出属于预定义类别的文本片段，如人名、地理位置名、机构组织名等[1]。NER 不仅是信息抽取的独立工具，而且在各种自然语言处理(NLP)的分支中充当着重要角色，如文本分析、信息抽取，文本摘要生成，问答系统，建立知识图谱等。近年来，深度学习(DL，也称为深度神经网络)被各个领域广泛应用，并取得成功。从 Collobert 等人开始[2] [3] [4]，基于 DL 的 NER 系统，其具有最少的特征工程，正在迅猛的发展。在过去的几年中，NER 慢慢引入深度学习，并得到了很好的效果[5] [6] [7] [8] [9]。这一现象鼓舞我们深入研究 NER 中的深度学习技术。Nadeau 和 Sekine [3]可以说是最成熟的技术，发表于 2007 年。该论文讲述了从手工制定的规则到机器学习的技术发展趋势。2013 年，Patawar 和 Potey [10]在 2015 年进行了简短回顾。最近的两项研究分别涉及新领域[11]和复杂的实体提及[12]。总之，现有的研究主要包括基于特征的机器学习模型，与这项工作更为紧密的是最近的两次调查。在 2018 年，Goyal 等[13]调查了 NER 的发展和进步。但是，它们不包括深度学习技术的最新进展。Yadav 和 Bethard [12]根据句子中的单词表示，对 NER 的最新进展进行了研究。这项研究主要技术是输入的分布式表示形式(例如，字符级和单词级嵌入)，而不是采用查看上下文编码器和标签解码器。在本文中，首先我们随机初始化训练集中的每个字特征，并在获取该字典句子中每个字的特征之后，利用周期卷积来得到其固定长度的特征，以此作为句特征;随后训练数据自动编码器，通过栈式自动编码器得到高层句子的特征；最后通过高层句特征与字特征的组合训练字的标注网络模型来得到未知字的标注值，再进行实体扩展(分类，属性)，最后利用马尔科夫逻辑网络优化整体识别效果。

2. 命名实体识别研究历史

对英文文本的实体识别的研究领先于中文文本命名实体识别的研究，早在 1991 年，Rau 在第七届 IEEE 人工智能应用会议上，发布了一个实体识别系统，该系统可以抽取出文本中的公司名称[14]，在当时引起了轰动。从此命名实体识别就开始被引入 MUC-6 [15]，MUC-7 的 MET-2 [16]等一系列会议中。在 20 世纪 90 年代初期，孙茂松等[17]开始研究中文命名实体识别，最开始研究的范围比较窄，仅仅是对文本中的人名进行识别。而后张小衡等[18]开始扩大实体识别的范围，建立了高校名数据集，并采用人工规则进行了实验，完成了对文本中的组织机构名称的抽取。2000 年，ZHANG, ZHOU 等[19]在 ACL 会议上发表了一个抽取命名实体及它们之间关系的系统。在算法方面，命名实体识别最早期的方法是基于规则，基于字典。后来发展成传统机器学习的方法，比如 CRF。经过专家们的不断努力，实体识别引入了深度学习的方法，比如 RNN-CRF, CNN-CRF 等模型[20] [21] [22] [23] [24]。

3. 关键技术

3.1. 字词结合向量

3.1.1. 字特征向量

基于字特征向量的模型缺点是单独输入字符，并没有考虑到相邻字符，句子，段落之间所存在的语义关系。其模型结构如图 1 所示，在最开始的时候字向量公式为

$$X_j^c = e^c\left(c_j\right)$$

$$h_j^c = \left[\vec{h}_j^c; \overleftarrow{h}_j^c\right]$$

将字符序列 $c_1, c_2, \cdots, c_m$，输入 Lstm-CRF 模型中。每个字符 c_j 用第一个公式表示，e^c 表示字符嵌入查找表，即计算字向量的操作。一个双向 LSTM 应用于 $x_1, x_2, \cdots, x_j$，可以得到双向 lstm 的两个方向隐含层的输出，即分别在从左到右和从右到左的方向上的两组不同的参数。根据第二个公式将其参数进行合并，得到的就是第 j 个字符的输出，再将其输入 CRF 中。2017 年，CHEN [25]等人改进了该模型
如公式三所示

$$X_j^c = \left[e^c\left(c_j\right); e^b\left(c_j, c_{j+1}\right)\right]$$

在计算输入向量的时候把这个字的向量和下一个字符的向量进行了合并，以此来加强字符间的语义关系。

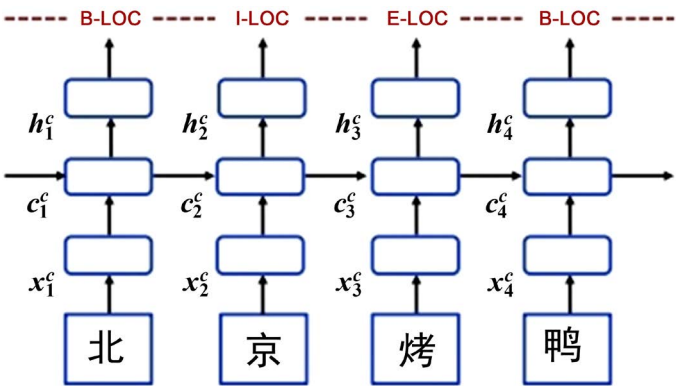

Figure 1. Character feature vector model structure
图 1. 字特征向量模型结构

3.1.2. 词特征向量

词特征向量类似字特征，它需要词嵌入来表示每个词，其模型结构如图 2 所示，公式如下：

$$X_i^w = \left[e^w\left(w_i\right); X_i^c \right]$$

$$X_i^c = \left[\overrightarrow{h}^c_{t(i,len(i))}; \overleftarrow{h}^c_{t(i,1)} \right]$$

该模型用了两层 LSTM，第一层是公式七中计算每个词所有字向量的输出。第二层双向 Lstm 用来学习隐藏状态 $\overrightarrow{h}^c_{t(i,1)},\cdots,\overrightarrow{h}^c_{t(i,len(i))}$，$len(i)$表示 w_i 中的字符数。公式中 $\overrightarrow{h}^c_{t(i,len(i))}$ 表示第 i 个词最后一个字的正向的隐含层 h；$\overleftarrow{h}^c_{t(i,1)}$ 则表示第 i 个词第一个字的反向的隐含层 h [26]。

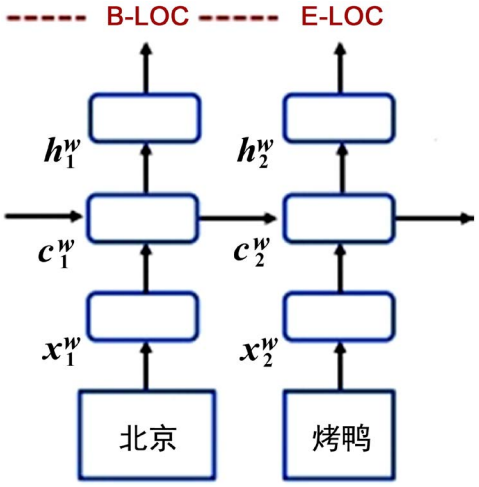

Figure 2. Word feature vector model structure
图 2. 词特征向量模型结构

3.1.3. 字词结合特征

单词－字符格模型可以看作是基于字符的模型的扩展，集成了基于单词的单元和用于控制信息流的其他门。其模型结构如图 3 所示，模型的输入是字符序列 $c_1,c_2,\cdots,c_m$ 以及与词典 D 中的单词匹配的所有字符子序列。该模型涉及四种类型的向量，即输入向量，输出隐藏向量，单元向量和门向量。作为基本组成部分，字符输入向量用于表示每个字符 c_j，该模型的基本递归结构是使用字符单元向量 c_j^c 和一个隐藏元素构造的每个 c_j 上的向量 h_j^c，其中 c_j^c 用于记录从句子开头到 c_j 的循环信息流，h_j^c 用于进行 CRF 序列标记[27]。该模型公式如下：

$$x_j^c = e^c\left(c_j\right)$$

$$\begin{bmatrix} \boldsymbol{i}_j^c \\ \boldsymbol{o}_j^c \\ \boldsymbol{f}_j^c \\ \tilde{\boldsymbol{c}}_j^c \end{bmatrix} = \begin{bmatrix} \sigma \\ \sigma \\ \sigma \\ \tanh \end{bmatrix} \left(\boldsymbol{W}^{c\top} \begin{bmatrix} \boldsymbol{x}_j^c \\ \boldsymbol{h}_{j-1}^c \end{bmatrix} + \boldsymbol{b}^c \right)$$

$$\boldsymbol{c}_j^c = \boldsymbol{f}_j^c \odot \boldsymbol{c}_{j-1}^c + \boldsymbol{i}_j^c \odot \tilde{\boldsymbol{c}}_j^c$$

$$\boldsymbol{h}_j^c = \boldsymbol{o}_j^c \odot \tanh\left(\boldsymbol{c}_j^c\right)$$

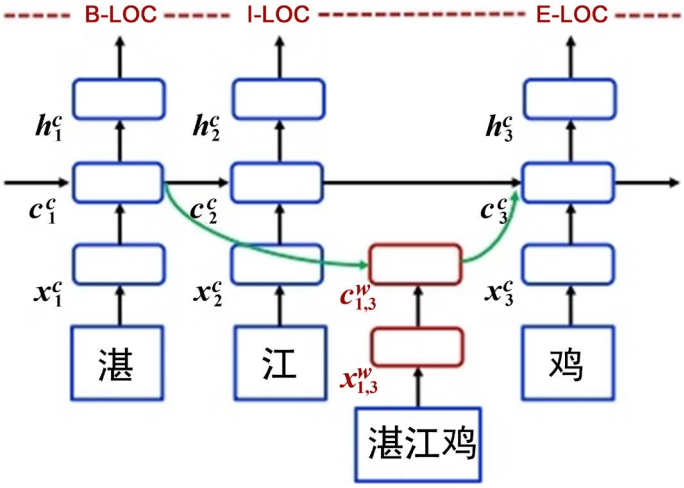

Figure 3. Lattice feature vector model structure
图 3. 单词－字符格模型结构

3.2. 实体扩展

本文制作了中国地理特产数据集，实体扩展是指针对某一实体类别比如梨，给出了香梨，鸭梨等种子实体，输出是梨这个类别里其他未知实体，比如雪花梨，水晶梨等。本文采用的是实体扩展方法是基于模板的实体抽取，我们的目标实体(雪花梨，水晶梨)与种子实体(香梨，鸭梨)同属于梨这个语义类，首先我们预定义好指示上下文关系的语义模板，再分析种子实体(香梨，鸭梨)所处的上下文得到模板，然后基于 Booststrapping 策略，反复迭代，得到更多的种子模板，以模板为特征，计算候选实体的置信度。

3.3. 马尔科夫逻辑网络(MLN)优化整体识别效果

马尔科夫逻辑网络是一种统计关系学习模型，之所以引入该网络，是因为我们的模型针对我们的实体，从分类属性方面进行了扩展，但是这两个方面的扩展是相互独立的，所以我们在扩展之后，引入了马尔科夫逻辑网络来提高我们实体识别的精确度。我们将模型得到的规则转化为子句的集合，将每一个子句看成一个节点，而每个集合中子句的关系即为连边，至此我们就构成了马尔科夫逻辑网。其概率计算的公式如下：

$$P(X=x)=\frac{1}{z}\exp\left(\sum_{i=1}^{F} w_i n_i(x)\right)$$

这个公式中，$n_i(x)$ 是某一个规则 F_i 的取值为真的时候所对应闭规则的个数。如果我们的规则权重越大，那么必然 $n_i(x)$ 就会越大，也就说明我们所取的 x 在 F_i 下越可信。然后我们把当前取值 x 在所有规则下的可信度相乘，再除以归一化因子，就得到了当前取值的概率[28] [29]。

4. 实验结果分析

4.1. 实验数据

本文人工构建了一个中国地理特产介绍的命名实体识别数据集，我们的中国地理特产数据集包括训练集，开发集，测试集，训练集里有 34906 个字，开发集里有 4396 个字，测试集里有 4768 个字。

4.2. 实验环境

本研究中的实验环境为 ubuntu16.04 操作系统，Python3.6，深度学习框架为 Pytorch1.4.0。

4.3. 实验设计与结果

本文模型主要识别中国地理特产网数据集中的特产名，地理位置名以及组织机构名，为了验证本文模型的性能，我们设计了三个实验，主要模型为 Lattice LSTM-CNN-CRF，而对比实验模型我们用了 LSTM-CNN-CRF 模型和 BiLSTM-CNN-CRF 模型，本文实验结果是从准确率、召回率和 F1 值三个方面来进行。见下表 1。

Table 1. Experimental results
表 1. 实验效果

模型	准确率	召回率	F1 值
LSTM-CNN-CRF	88.90	87.90	88.10
BiLSTM-CNN-CRF	89.80	89.80	89.20
Lattice LSTM-CNN-CRF	92.82	92.10	92.56

5. 结束语

在本文中，首先我们随机初始化训练集中的每个字特征，并在获取该字典句子中每个字的特征之后，利用周期卷积来得到其固定长度的特征，以此作为句特征；随后训练数据自动编码器，通过栈式自动编码器得到高层句子的特征；最后通过高层句特征与字特征的组合训练字的标注网络模型来得到未知字的标注值，再进行实体扩展(分类，属性)，最后利用马尔科夫逻辑网络优化整体识别效果。

基金项目

《工业过程数据实时获取与知识自动化》，国家自然科学基金委员会资助项目，项目编号：U17012621006336。

参考文献

[1] Christian, B., Heath, T. and Berners-Lee, T. (2009) Linked Data—The Story So Far. *International Journal on Semantic Web and Information Systems*, **5**, 1-22. https://doi.org/10.4018/jswis.2009081901

[2] Bollacker, K., Cook, R. and Tufts, P. (2007) Freebase: A Shared Database of Structured General Human Knowledge. *AAAI*, **7**, 1962-1963.

[3] Nadeau, D. and Sekine, S. (2007) A Survey of Named Entity Recognition and Classification. *Lingvisticæ Investigationes*, **30**, 3-26. https://doi.org/10.1075/li.30.1.03nad

[4] Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K. and Kuksa, P. (2011) Natural Language Processing (Almost) from Scratch. *Journal of Machine Learning Research*, **12**, 2493-2537.

[5] Huang, Z., Xu, W. and Yu, K. (2015) Bidirectional LSTM-CRF Models for Sequence Tagging. arXiv preprint arXiv:1508.01991.

[6] Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K. and Dyer, C. (2016) Neural Architectures for Named Entity Recognition. *Proceedings of the* 2016 *Conference of the North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies*, San Diego, June 2016, 260-270. https://doi.org/10.18653/v1/N16-1030

[7] Chiu, J.P. and Nichols, E. (2016) Named Entity Recognition with Bidirectional LSTM-CNNs. *Transactions of the Association for Computational Linguistics*, **4**, 357-370. https://doi.org/10.1162/tacl_a_00104

[8] Peters, M.E., Ammar, W., Bhagavatula, C. and Power, R. (2017) Semi-Supervised Sequence Tagging with Bidirectional Language Models. *Proceedings of the* 55*th Annual Meeting of the Association for Computational Linguistics*, **1**, 1756-1765. https://doi.org/10.18653/v1/P17-1161

[9] Patawar, M.L. and Potey, M. (2015) Approaches to Named Entity Recognition: A Survey. *International Journal of Innovative Research in Computer and Communication Engineering*, **3**, 12201-12208.

[10] Saju, C.J. and Shaja, A. (2017) A Survey on Efficient Extraction of Named Entities from New Domains Using Big Data Analytics. 2017 *Second International Conference on Recent Trends and Challenges in Computational Models* (*ICRTCCM*), Tindivanam, 3-4 February 2017, 170-175. https://doi.org/10.1109/ICRTCCM.2017.34

[11] Dai, X. (2018) Recognizing Complex Entity Mentions: A Review and Future Directions. *Proceedings of ACL* 2018, *Student Research Workshop*, Melbourne, July 2018, 37-44. https://doi.org/10.18653/v1/P18-3006

[12] Yadav, V. and Bethard, S. (2018) A Survey on Recent Advances in Named Entity Recognition from Deep Learning Models. *Proceedings of the* 27*th International Conference on Computational Linguistics*, Santa Fe, 20-26 August 2018, 2145-2158.

[13] Goyal, A., Gupta, V. and Kumar, M. (2018) Recent Named Entity Recognition and Classification Techniques: A Systematic Review. *Computer Science Review*, **29**, 21-43. https://doi.org/10.1016/j.cosrev.2018.06.001

[14] Rau, L.F. (1991) Extracting Company Names from Text. *Proceedings of the* 7*th IEEE Conference on Artificial Intelligence Applications*, Miami Beach, 24-28 February 1991, 29-32.

[15] Grishman, R. and Sundheim, B. (1996) Message Understanding Conference-6: A Brief History. *Proceedings of the* 16*th International Conference on Computational Linguistics*, **1**, 466-471. https://doi.org/10.3115/992628.992709

[16] Chinchor, N.A. (1998) Overview of MUC-7/MET-2. *Proceedings of the* 7*th Message Understanding Conference*.

[17] 孙茂松, 黄昌宁, 高海燕, 等. 中文姓名的自动辨识[J]. 中文信息学报, 1995, 9(2): 16-27.

[18] 张小衡, 王玲玲. 中文机构名称的识别与分析[J]. 中文信息学报, 1997, 11(4): 21-32.

[19] Zhang, Y. and Zhou, J.F. (2000) A Trainable Method for Extracting Chinese Entity Names and Their Relations. In: *Proceedings of the* 2*nd Chinese Language Processing Workshop*, Hong Kong, October 2000, 66-76. https://doi.org/10.3115/1117769.1117780

[20] Bikel, D.M., Schwarta, R. and Weischedel, R.M. (1999) An Algorithm that Learns What's in a Name. *Machine Learning*, **34**, 211-231. https://doi.org/10.1023/A:1007558221122

[21] Liao, W. and Veeramachaneni, S. (2009) A Simple Semi-supervised Algorithm for Named Entity Recognition. *Proceedings of the NAACL HLT* 2009 *Workshop on Semi-Supervised Learning for Natural Language Processing*, Boulder, June 2009, 58-65. https://doi.org/10.3115/1621829.1621837

[22] Ratinov, L. and Roth, D. (2009) Design Challenges and Misconceptions in Named Entity Recognition. *Proceedings of the* 13*th Conference on Computational Natural Language Learning*, Boulder, 4-5 June 2009, 147-155. https://doi.org/10.3115/1596374.1596399

[23] 冯元勇, 孙乐, 李文波, 等. 基于单字提示特征的中文命名实体识别快速算法[J]. 中文信息学报, 2008, 22(1): 105-110.

[24] 郑逢强, 林磊, 刘秉权, 等. 《知网》在命名实体识别中的应用研究[J]. 中文信息学报, 2008, 22(5): 97-101.

[25] Chen, X., Qiu, X., Zhu, C., Liu, P. and Huang, X. (2015) Long Short-Term Memory Neural Networks for Chinese Word Segmentation. *Proceedings of the* 2015 *Conference on Empirical Methods in Natural Language Processing*, Lisbon, September 2015, 1197-1206. http://aclweb.org/anthology/D15-1141 https://doi.org/10.18653/v1/D15-1141

[26] Limsopatham, N. and Collier, N. (2016) Bidirectional LSTM for Named Entity Recognition in Twitter Messages. *Proceedings of the* 2*nd Workshop on Noisy User-generated Text* (*WNUT*), Osaka, 11 December 2016, 145-152.

[27] Dong, C., Zhang, J., Zong, C., Hattori, M. and Di, H. (2016) Character-Based LSTM-CRF with Radical-Level Features for Chinese Named Entity Recognition. In: Lin, C.Y., Xue, N., Zhao, D., Huang, X. and Feng, Y., Eds., *Natural Language Understanding and Intelligent Applications*, Springer, Cham, 239-250. https://doi.org/10.1007/978-3-319-50496-4_20

[28] Zhang, Y. and Yang, J. (2018) Chinese NER Using Lattice LSTM. *Proceedings of the* 56*th Annual Meeting of the Association for Computational Linguistics*, **1**, 1554-1564. https://doi.org/10.18653/v1/P18-1144

[29] Ee, S. and Xiang, Y. (2017) Chinese Named Entity Recognition with Character-Word Mixed Embedding. *Proceedings of the* 2017 *ACM on Conference on Information and Knowledge Management*, ACM, 2055-2058.