

中文命名实体识别综述

赵 山, 罗 睿, 蔡志平⁺

国防科技大学 计算机学院, 长沙 410073

+ 通信作者 E-mail: zpcai@nudt.edu.cn

摘 要: 中文命名实体识别(NER)任务是信息抽取领域内的一个子任务,其任务目标是给定一段非结构文本后,从句子中寻找、识别和分类相关实体,例如人名、地名和机构名称。中文命名实体识别是一个自然语言处理(NLP)领域的基本任务,在许多下游NLP任务中,包括信息检索、关系抽取和问答系统中扮演着重要角色。全面回顾了现有的基于神经网络的单词-字符晶格结构的中文NER模型。首先介绍了中文NER相比英语NER难度更大,存在着中文文本相关实体边界难以确定和中文语法结构复杂等难点及挑战。然后调研了在不同神经网络架构下(RNN、CNN、GNN和Transformer)最具代表性的晶格结构的中文NER模型。由于单词序列信息可以给基于字符的序列学习更多边界信息,为了显式地利用每个字符所相关的词汇信息,过去的这些工作提出通过词-字符晶格结构将单词信息整合到字符序列中。这些在中文NER任务上基于神经网络的单词-字符晶格结构的性能要明显优于基于单词或基于字符的方法。最后介绍了中文NER的数据集及评价标准。

关键词: 命名实体识别(NER);晶格结构;神经网络

文献标志码: A **中图分类号:** TP391

Survey of Chinese Named Entity Recognition

ZHAO Shan, LUO Rui, CAI Zhiping⁺

College of Computer, National University of Defense Technology, Changsha 410073, China

Abstract: The Chinese named entity recognition (NER) task is a sub-task within the information extraction domain, where the task goal is to find, identify and classify relevant entities, such as names of people, places and organizations, from sentences given a piece of unstructured text. Chinese named entity recognition is a fundamental task in the field of natural language processing (NLP) and plays an important role in many downstream NLP tasks, including information retrieval, relationship extraction and question and answer systems. This paper provides a comprehensive review of existing neural network-based word-character lattice structures for Chinese NER models. Firstly, this paper introduces that Chinese NER is more difficult than English NER, and there are difficulties and challenges such as difficulty in determining the boundaries of Chinese text-related entities and complex Chinese grammatical structures. Secondly, this paper investigates the most representative lattice-structured Chinese NER models under different neural network architectures (RNN (recurrent neural network), CNN (convolutional neural network), GNN (graph neural network) and Transformer). Since word sequence information can capture more boundary information for character-based sequence learning, in order to explicitly exploit the lexical information associated with each character, some prior work has proposed integrating word information into character sequences via word-character lattice structures.

基金项目: 国家重点研发计划(2020YFC2003400)。

This work was supported by the National Key Research and Development Program of China (2020YFC2003400).

收稿日期: 2021-07-08 **修回日期:** 2021-09-22

These neural network-based word-character lattice structures perform significantly better than word-based or character-based approaches on the Chinese NER task. Finally, this paper introduces the dataset and evaluation criteria of Chinese NER.

Key words: named entity recognition (NER); lattice structure; neural network

随着新闻、公司档案、病历、政府文件、法院听证会和社交媒体等形式的信息激增,每个人都被信息过载所淹没。这些信息大部分是非结构化的,即自由文本,因此很难进行推理和解释。在自然语言处理领域,通过分析文本从而实现有价值信息抽取的过程,被称为信息抽取(information extraction, IE)^[1]。通过对结构不统一的非结构化文本进行信息抽取,能够得到简单明确的结构化数据,便于人们对数据进行高效检索及管理^[2]。实体抽取作为信息抽取技术的主要任务,在近年来受到学术界和产业界的极大关注。

命名实体识别(named entity recognition, NER)^[3]任务是信息抽取领域内的一个子任务,其任务目标是给定一段非结构文本后,从句子中寻找、识别和分类相关实体,例如人名、地名和机构名称。命名实体识别及其相关技术无论在学术研究上还是实际应用中都具有重大且深远的意义。近年来随着大规模有监督数据集的相继提出与深度学习技术的快速发展,命名实体识别相关技术在各大顶级学术会议和期刊上越来越受到关注。语义知识丰富度的提升给阅读理解、自动问答和机器翻译等任务带来了进一步的效果的提高^[4-6]。

与英文NER相比,中文NER涉及到分词问题。中文命名实体识别是一个自然语言处理(natural language processing, NLP)领域的基本任务,在许多下游NLP任务中,包括信息检索^[7]、关系抽取^[8]和问答系统^[9]中扮演着重要角色。与英语NER相比,中文命名实体难度更大。这主要是因为中文文本相关实体边界难以确定和中文语法结构复杂。在中文命名实体识别中实体的边界也即是单词的边界。分词过程必然导致错误传播问题,这也是目前中文命名实体识别的一大挑战。

传统上,中文NER的任务被分解为两个分离的子任务,即中文分词和单词序列标注^[10]。此方法的主要缺点是错误传播:中文词语边界分割错误会对实体识别产生影响^[11-12]。另一方面,基于字符的模型可以避免单词分割错误,从而胜过基于单词的方法。

由于单词序列信息可以给基于字符的序列学习更多边界信息,为了显式地利用每个字符所相关的词汇信息,过去的一些工作^[13]已经提出通过词-字符晶结构将单词信息整合到字符序列中。这些在中文NER任务上基于神经网络的单词-字符晶格结构的性能要明显优于基于单词或基于字符的方法。

1 研究难点及挑战

本章总结了中文命名实体识别面临的若干技术难点与挑战。中文命名实体识别任务是中文信息抽取领域内的一个子任务,其任务目标是在给定一段非结构文本后,从句子中寻找、识别和分类相关实体,例如人名、地名和机构名称。由于单词序列信息可以给基于字符的序列学习更多边界信息,在中文NER任务上基于神经网络的单词-字符晶格结构的性能要明显优于基于词或基于字符的方法。如图1所示为基于Lattice LSTM^[14]的中文命名实体识别模型结构图。该模型由Zhang等在ACL2018提出。他们提出了一种LSTM模型的变种(Lattice LSTM)模型。该模型利用一个句子中所有被单个字符匹配到的词语,把这些词语编码为一个有向无环图(directed acyclic graph, DAG)。得益于丰富词汇信息,Lattice LSTM模型已在各种数据集上取得了不错的结果,而且与单一的基于字符的模型或基于词的模型都有明显的模型效果的提升。然而,这个有向无环图结构有时无法选择正确的路径,这可能导致晶格模型退化为部分基于单词的模型。此外该模型由于是基于BiLSTM,不同字匹配的词之间不能信息交互。WC-LSTM(word-character LSTM)^[13]为改进这一问题,采取四种词嵌入策略:最短词汇信息、最长词汇信息(如图2所示)、average、self-attention。如果没有对应的词汇则用<PAD>代替,从而可以进行batch并行化。这四种策略的目的都是将每个字符为结尾的词汇信息进行固定编码表示。以最长词汇信息为例:选取当前字符匹配到的字符最长的词作为匹配词。但是WC-LSTM仍然存在信息损失问题,不能充分利用词汇信息。

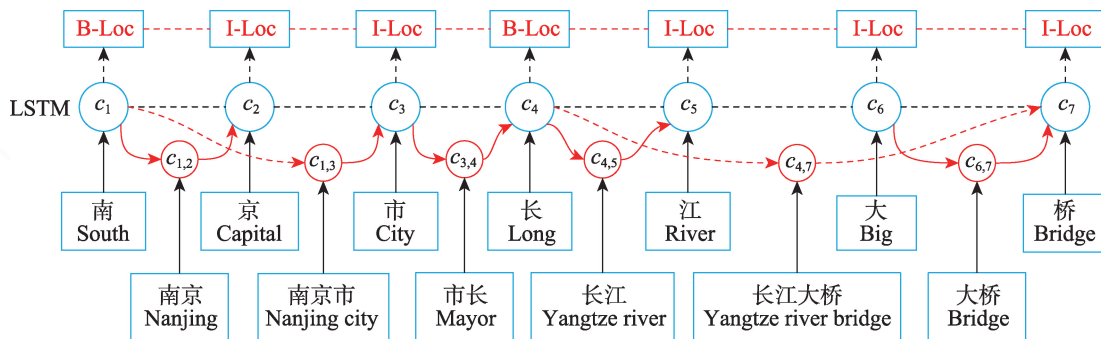


图1 原始 Lattice LSTM

Fig.1 Original Lattice LSTM

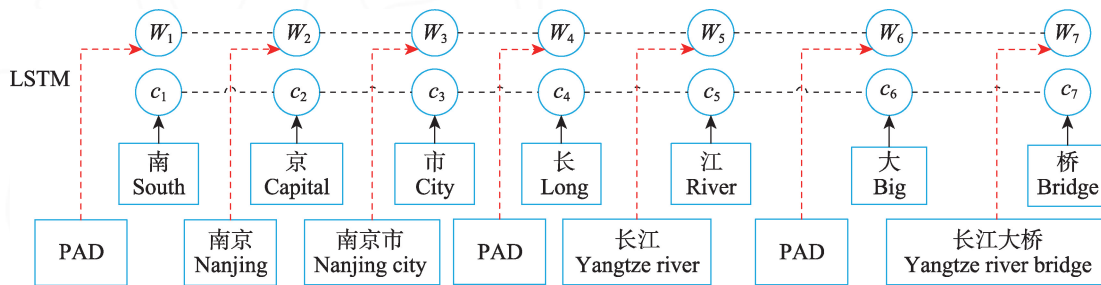


图2 最长策略

Fig.2 Longest strategy

2 研究现状及热点

本章将会介绍命名实体识别任务的研究现状以及一些当前研究人员关注的研究热点。

2.1 基于规则的方法

基于规则的NER系统依赖于人工制定的规则。可以基于特定领域的地名词典和句法词汇模式来设计规则。比较知名的系统包括LaSIE-II^[15]、NetOwl^[16]、Facile^[17]和SAR^[18]。当词汇库详尽无遗时,基于规则的系统是一种好的选择。但是在一些特定领域,由于其特定的规则和不完整的词汇库,经常会从此类系统中观察到高精度和低召回率,并且这些系统无法转移到其他领域。

2.2 基于无监督的方法

无监督学习的典型方法是聚类^[19]。基于聚类的NER系统抽取相关实体是通过上下文相似度的聚类实现的。Collins等人^[20]仅仅使用少量的种子标注数据和7个特征,包括拼写(比如大小写)、实体上下文、实体本身等,进行实体识别。Nadeau等人^[21]提出一种地名词典构建和命名实体歧义解析的无监督系统。该系统基于简单而高效的启发式方法,结合了实体提取和歧义消除。

2.3 基于有监督的方法

通过监督学习,NER可以被转换为多分类或序

列标注任务。给定带标注的数据样本,经过精心设计的特征可以用来表示每个训练示例。然后利用机器学习算法学习模型,从未知的数据中识别相似的模式。许多机器学习算法已在有监督的NER中应用,包括隐马尔可夫模型(hidden Markov model, HMM)^[22]、决策树^[23]、最大熵模型^[24]、支持向量机(support vector machine, SVM)^[25]和条件随机场(conditional random fields, CRF)^[26]。

2.4 基于深度学习的方法

近年来,基于深度学习的NER模型占主导地位并取得最新成果。与基于特征的方法相比,深度学习有助于自动发现隐藏的特征。对于英文NER,目前根据单词在句子中的表现形式,可以把基于神经网络的NER结构分为词级和字级。

(1)词级:在此体系结构中,句子的单词作为递归神经网络的输入,每个单词都由其单词嵌入表示。Collobert等人^[27]提出了第一个词级的神经网络模型,该模型使用了卷积层而不是BiLSTM层,并且卷积层的输出被提供给CRF层进行最终预测。通过引入地名词典和SENNA嵌入,该模型在English CoNLL 2003数据集上获得了89.59%的F1分数。Yadav等人^[28]提出了一个词级别LSTM结构,如图3所示。此外,他们的模型额外增加了一个CRF层用于提高模型性能。

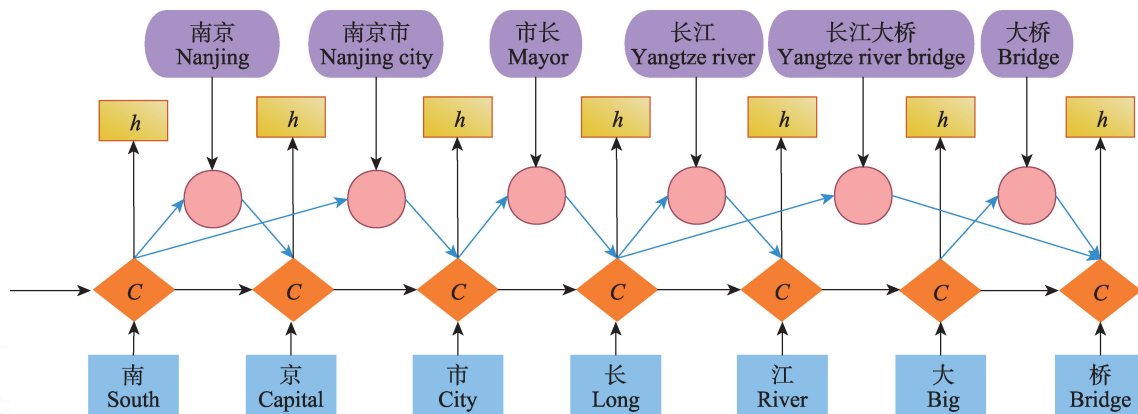


图3 晶格LSTM结构

Fig.3 Lattice-LSTM structure

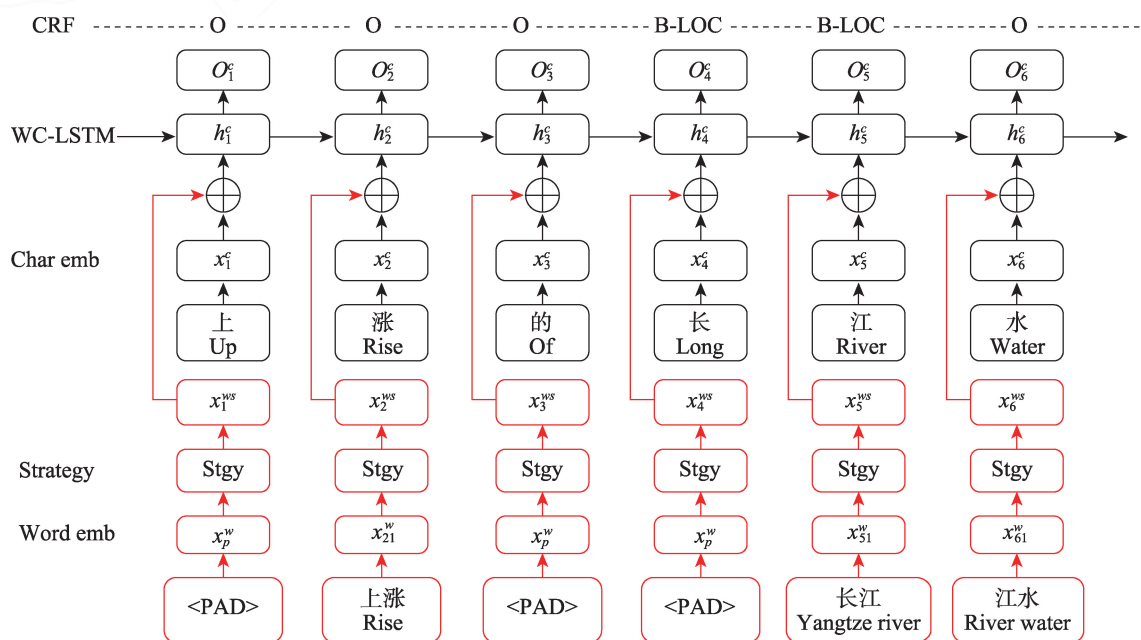


图4 WC-LSTM结构

Fig.4 WC-LSTM structure

在CoNLL 2003数据集上获得了84.26%的F1分数。

(2)字级别:在此模型中,句子被视为字符序列。该序列通过RNN(recurrent neural network)传递,预测每个字符的标签(图4所示)。Ma等人^[29]利用CNN(convolutional neural network)提取单词的字符级表示。早先大多数GNN(graph neural networks)的处理对象都是同质的图结构,通过模型学习这类图的节点表示。最近,Peters等人^[30]提出了ELMo(embeddings from language models)单词表示形式,该表示形式是在具有字符卷积的两层双向语言模型的基础上计算的。

2.5 基于晶格结构的方法

与英语命名实体相比,中文命名实体难度大。

主要是因为中文文本相关实体边界难以确定和中文语法结构复杂。传统上,中文NER的任务被分解为两个分离的子任务,即中文分词和单词序列标注^[10]。此方法的主要缺点是错误传播:中文词语边界分割错误会对实体识别产生影响^[11]。基于字符的NER的一个缺点是没有充分利用明确的单词和单词序列信息,哪怕这些信息可能是有用的。由于单词序列信息可以给基于字符的序列学习多边界信息,在中文NER任务上基于神经网络的词-字符晶格结构的性能要明显优于基于词或基于字符的方法。为了解决这个问题,Zhang和Yang^[14]首先提出了一种LSTM模型的变种(Lattice-LSTM)模型,如图3所示。该模型利用一个句子中所有被单个字符匹配到的词语,把这

些词语编码为一个有向无环图(directed acyclic graph, DAG)。该模型利用一个句子中所有被单个字符匹配到的词语,把这些词语编码为一个有向无环图。得益于丰富词汇信息,Lattice-LSTM模型已在各种数据集上取得了不错的结果。然而,这个有向无环图结构有时无法选择正确的路径,这可能导致晶格模型退化为部分基于单词的模型。

为了解决上述问题,Liu等人^[13]提出了一种新的单词-字符LSTM(WC-LSTM)来将单词信息集成到基于字符的模型中,如图4所示。该模型首先将一个中文句子表示为一系列字符-单词对,以将单词信息整合到每个字符中,从而保证模型不会退化为部分基于单词的模型。同时,Liu等人^[13]还设计了四种不同的词编码策略。这些策略可以将词信息编码为固定大小向量,从而使得模型能够批训练并适应各种应用场景。

Gui等人^[31]提出了一种基于CNN的NER模型(CNN-based Chinese NER with lexicon rethinking, R-

CNN),该模型对匹配的词语以不同的窗口大小进行编码(如图5所示)。该模型主要通过CNN并行处理整个句子以及所有潜在的单词,并且还应用了一种反思机制来处理词典中潜在单词之间的冲突。这种反思机制可以利用高级语义来细化嵌入单词的权重,并解决潜在单词之间的冲突。

此外,Gui等人^[32]提出了一个基于词典的图形神经网络(lexicon-based graph neural network for Chinese NER, LGN),把中文NER当作一个节点分类任务(如图6所示)。得益于图神经网络实现了字符和单词之间更好的连接,实现了字符和单词之间更好的交互效果。词汇知识将相关字符连接起来,以捕获本地特征。同时,设计了一个全局中继节点来捕获远程依赖和高级特征。LGN遵循邻域聚集方案,其中通过递归聚集其输入边和全局中继节点来计算节点表示。由于聚合的多次迭代,该模型可以使用全局上下文信息来重复比较歧义词,以获得更好的预测。

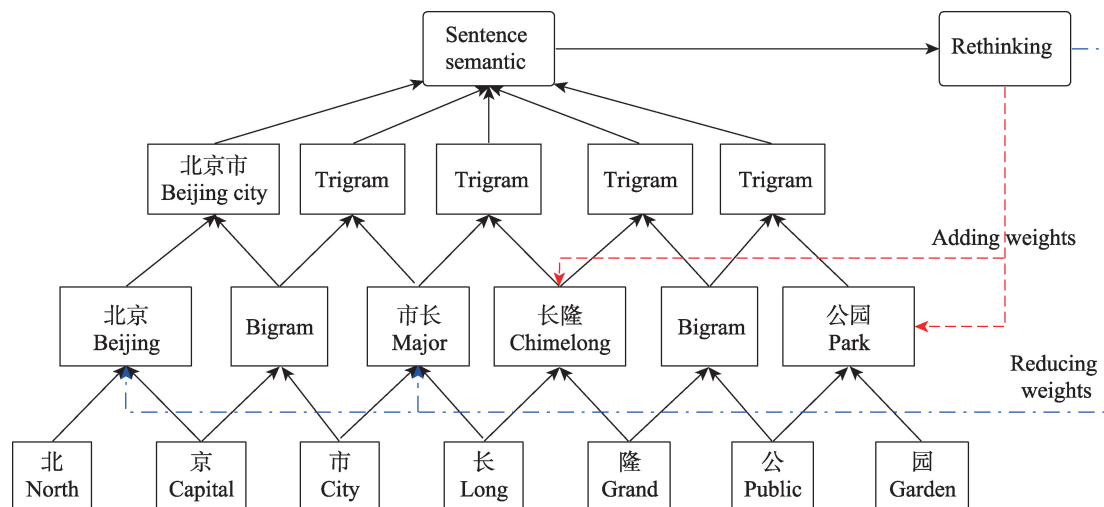


图5 R-CNN模型示意图

Fig.5 R-CNN schematic model

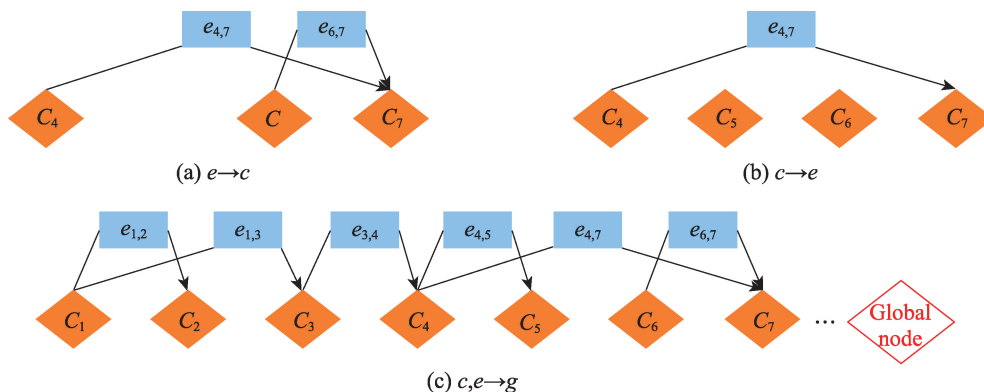


图6 LGN的聚合

Fig.6 LGN aggregation

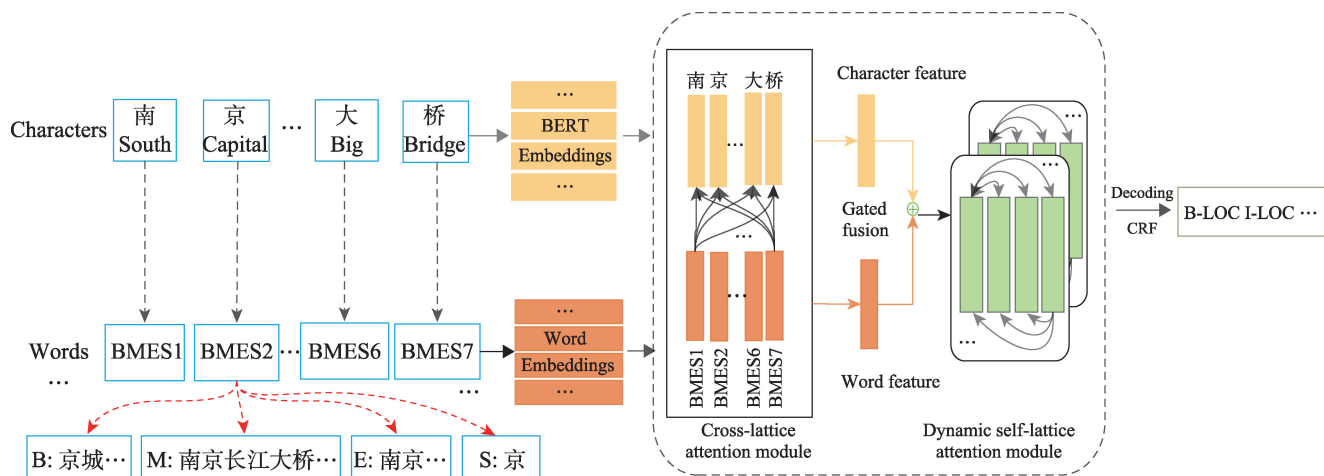


图7 跨自晶格模型注意网络方法

Fig.7 Cross- and self-lattice attention network

与之类似的, Sui等人^[33]针对词汇匹配问题, 提出了一种基于字符的协同图形网络, 包括编码层、图形层、融合层和解码层。Ding等人^[34]为了解决地名录的冲突匹配问题, 提出了一个多向图模型, 该模型能够通过学习上下文信息有效地解决冲突匹配问题。同时为了更好地训练模型, Ding等人^[34]还开源了一个新的电子商务领域中文数据集。

但是, 由于NER对句子结构非常敏感, 这些方法仍然需要使用LSTM作为骨干编码器, 这种混合编码会使得模型变得复杂。

最近, Yan等人^[35]在中文NER任务上提出了一种改进的编码器TENER(transformer encoder for named entity recognition)。Yan等人^[35]通过使用方向相对位置编码, 减少参数的数量, 并提高注意分布, 来提高基于Transformer的模型在NER任务中的性能。经过实验验证, 模型性能有了很大的提高, 甚至比基于BiLSTM的模型性能更好。

Ma等人^[36]提出了一个简单的方法来实现Lattice-LSTM的思想, 即把每个字符的所有匹配词合并到一个基于字符的NER模型中。该模型不仅在字符表示中编码词典信息, 同时还结合了新的编码方案。新的编码方案能够尽可能多地保留词典匹配结果。因此该模型能够实现快速的推理速度。与Lattice-LSTM相比, 该模型不需要复杂的模型结构, 更容易实现, 并且可以通过调整字符表示层快速适应任何合适的神经NER模型。

Li等人^[37]利用平面晶格结构, 以便Transformer可以通过位置编码来捕获词信息, 由此提出了FLAT(Chinese NER using flat-lattice transformer)模型。FLAT

采用全连接的自我关注来模拟序列中的长距离依赖关系。为了保持位置信息, Transformer为序列中的每个标记引入了位置表示。Transformer的自我关注机制使角色能够直接与任何潜在的词互动, 包括自我匹配的词。

Zhao等人^[38]提出了一种动态的跨自晶格注意网络方法。该方法的灵感来源于计算机视觉中的VQA(visual question answering)任务。如图7所示, 他们把字符和单词序列视为两种不同的模态。为了对词-字晶格结构上的信息交互进行建模, 首先设计了一个跨晶格注意力模块, 该模块旨在捕获两个输入特征空间之间细粒度的相关性。然后, 进一步构造一个动态的自晶格注意模块, 该模块能够动态融合单词特征并且在不管两个任意字符之间的距离如何而建立它们之间的直接联系。给定词和字向量表示和对齐的晶格结构, 模型首先利用跨晶格注意力模块以生成拥有词信息的字符特征。然后采用动态的自晶格注意力模块结合字符和单词特征, 最终获得自我关注的字符特征。以这种方式, 本文网络可以充分捕获词字晶格结构上的信息交互, 从而为中文NER预测提供了丰富的表示形式。

众所周知, 汉字是基于古代的象形文字发展而来的, 其结构往往反映了更多的文字信息。Wu等人^[39]为了利于汉字的结构信息, 提出了一种新颖的基于多元数据嵌入的Cross-Transformer模型。如图8所示, Wu等人^[39]在FLAT模型的基础上引入radical-stream模块, 该模块可以融合中文结构信息, 例如偏旁部首。该方法利用的中文结构信息对于同音字干扰中文分词也是一种有效的解决方法。

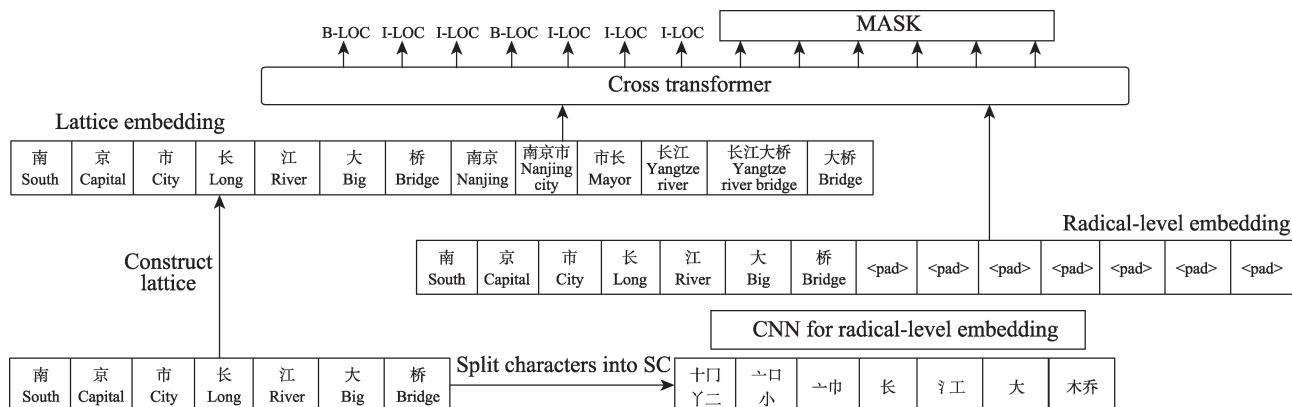


图8 多元数据嵌入的Cross-Transformer模型

Fig.8 Multi-meta data embedding based Cross-Transformer

表1 常见NER数据集

Table 1 Common NER datasets

数据集名称	年份	来源	实体类型数量	网址
OntoNotes	2007—2012	Magazine, news, Web 等	18	https://catalog.ldc.upenn.edu/LDC2013T19
Resume	2018	SinaFinance text	8	https://github.com/jiesutd/LatticeLSTM
Weibo	2015	social media	4	https://github.com/quincyliang/nlp-public-dataset/tree/master/ner-data/weibo
MSRA	2006	news	3	https://docs.qq.com/sheet/DVnpkTnF6VW9UeXdh?c=A1A0A0&tab=BB08J2
E-commerce	2019	e-commerce	2	https://github.com/PhantomGrapes/MultiDigraphNER

3 数据集及评价标准

本章将会介绍部分常用的中文NER数据集以及评价标准。

3.1 常见数据集

高质量标注数据集对于模型学习和评估都至关重要。本节总结了广泛使用的数据集。表1中列出了一些使用广泛的数据集,包括其数据源和实体类型(也称为标签类型)。

OntoNotes数据集语料来自很多领域,目前有5个版本1.0~5.0。数据集包含8种实体类型。Resume数据集来自新浪财经,其中包括多家中国股市上市公司高管简历。随机选择了1027份简历摘要,并用YEDDA手动标注了8种命名实体类型。MSRA和Weibo数据集分别来源于新闻利用和社交媒体(微博)。两个数据集包含的实体类型数量分别为3种和4种。E-commerce通过人工标注电商领域文本获得,包括两种实体类型。

3.2 评价标准

该任务通常采用F1值来评价模型的性能:

$$F1 = \frac{precision \times recall}{precision + recall} \quad (1)$$

其中, *precision* 表示模型输出正确实体所占的百分比, *recall* 表示模型在数据集中找到正确实体的百分比。在该任务中,普遍认为只要当实体的类型和边界都预测正确时,该实体才被认为识别正确。

4 总结

本文旨在回顾基于神经网络的单词-字符晶格结构的中文NER解决方案的最新研究,以帮助新研究人员对该领域的全面了解。本次调研包括了中文NER研究的背景、传统方法的简介、当前的研究现状和热点。首先,对比了中文NER与英文NER,相比之下,中文命名实体难度更大。这主要是因为中文文本相关实体边界难以确定和中文语法结构复杂。然后,调研了在不同神经网络架构下(RNN、CNN、GNN和Transformer)最具代表性的晶格结构的中文NER模型。最后,介绍了中文NER的数据集及评价标准。

参考文献:

- [1] 郭喜跃, 何婷婷. 信息抽取研究综述[J]. 计算机科学, 2015, 42(2): 14-17.
- GUO X Y, HE T T. Survey about research on information

- extraction[J]. Computer Science, 2015, 42(2): 14-17.
- [2] 邹晶. 基于深度学习的实体与关系联合提取方法研究[D]. 成都: 电子科技大学, 2019.
- ZOU J. Research on joint extraction method of entity and relation based on deep learning[D]. Chengdu: University of Electronic Science and Technology of China, 2019.
- [3] ZHAO S, CAI Z P, CHEN H W, et al. Adversarial training based lattice LSTM for Chinese clinical named entity recognition[J]. Journal of Biomedical Informatics, 2019, 99: 103290.
- [4] 石春丹, 秦岭. 基于BGRU-CRF的中文命名实体识别方法[J]. 计算机科学, 2019, 46(9): 237-242.
- SHI C D, QIN L. Chinese named entity recognition method based on BGRU-CRF[J]. Computer Science, 2019, 46(9): 237-242.
- [5] 卓玛措, 桑杰端珠, 才让加. 基于深度学习的古汉语命名实体识别研究[J]. 计算机科学与应用, 2020, 10(7): 1359-1366.
- ZHUO M C, SANGJIE D Z, CAI R J. Ancient Chinese named entity recognition based on deeping learning[J]. Computer Science and Application, 2020, 10(7): 1359-1366.
- [6] 孙长志. 基于深度学习的联合实体关系抽取[D]. 上海: 华东师范大学, 2019.
- SUN C Z. Joint entity relation extraction with deep learning[D]. Shanghai: East China Normal University, 2019.
- [7] CHEN Y B, XU L H, LIU K, et al. Event extraction via dynamic multi-pooling convolutional neural networks[C]// Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, Jul 26-31, 2015. Stroudsburg: ACL, 2015: 167-176.
- [8] BUNESCU R C, MOONEY R J. A shortest path dependency kernel for relation extraction[C]//Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, Vancouver, Oct 6-8, 2005. Stroudsburg: ACL, 2005: 724-731.
- [9] DIEFENBACH D, LÓPEZ V, SINGH K D, et al. Core techniques of question answering systems over knowledge bases: a survey[J]. Knowledge and Information Systems, 2018, 55(3): 529-569.
- [10] YANG J, TENG Z Y, ZHANG M S, et al. Combining discrete and neural features for sequence labeling[C]//LNCS 9623: Proceedings of the 17th International Conference on Computational Linguistics and Intelligent Text Processing, Konya, Apr 3-9, 2016. Cham: Springer, 2016: 140-154.
- [11] PENG N Y, DREDZE M. Named entity recognition for Chinese social media with jointly trained embeddings[C]// Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Sep 17-21, 2015. Stroudsburg: ACL, 2015: 548-554.
- [12] HE H F, SUN X. F-score driven max margin neural network for named entity recognition in Chinese social media[J]. arXiv:1611.04234, 2016.
- [13] LIU W, XU T G, XU Q H, et al. An encoding strategy based word-character LSTM for Chinese NER[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, Jun 2-7, 2019. Stroudsburg: ACL, 2019: 2379-2389.
- [14] ZHANG Y, YANG J. Chinese NER using lattice LSTM[C]// Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Jul 15-20, 2018. Stroudsburg: ACL, 2018: 1554-1564.
- [15] HUMPHREYS K, GAIZAUSKAS R, AZZAM S, et al. University of Sheffield: description of the LaSIE-II system as used for MUC-7[C]//Proceedings of the 7th Message Understanding Conference, Fairfax, Apr 29-May 1, 1998. Stroudsburg: ACL, 1998: 1-20.
- [16] KRUPKA G R, HAUSMAN K. IsoQuest Inc.: description of the NetOwl™ extractor system as used for MUC-7[C]// Proceedings of the 7th Message Understanding Conference, Fairfax, Apr 29-May 1, 1998. Stroudsburg: ACL, 1998: 1-10.
- [17] BLACK W J, RINALDI F, MOWATT D. FACILE: description of the NE system used for MUC-7[C]//Proceedings of the 7th Message Understanding Conference, Fairfax, Apr 29-May 1, 1998. Stroudsburg: ACL, 1998: 1-10.
- [18] AONE C, HALVERSON L, HAMPTON T, et al. SRA: description of the IE2 system used for MUC-7[C]//Proceedings of the 7th Message Understanding Conference, Fairfax, Apr 29-May 1, 1998. Stroudsburg: ACL, 1998: 1-14.
- [19] NADEAU D, SEKINE S. A survey of named entity recognition and classification[J]. Lingvisticae Investigationes, 2007, 30(1): 3-26.
- [20] COLLINS M, SINGER Y. Unsupervised models for named entity classification[C]//Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, College Park, Jun 21-22, 1999. Stroudsburg: ACL, 1999: 1-11.
- [21] NADEAU D, TURNEY P D, MATWIN S. Unsupervised named-entity recognition: generating gazetteers and resolving ambiguity[C]//LNCS 4013: Proceedings of the 19th Conference of the Canadian Society for Computational Studies of Intelligence, Québec City, Jun 7-9, 2006. Berlin, Heidelberg: Springer, 2006: 266-277.
- [22] EDDY S R. Hidden Markov models[J]. Current Opinion in

- Structural Biology, 1996, 6(3): 361-365.
- [23] QUINLAN J R. Induction of decision trees[J]. Machine Learning, 1986, 1(1): 81-106.
- [24] KAPUR J N. Maximum-entropy models in science and engineering[M]. New York: John Wiley & Sons, Inc., 1989.
- [25] SUTHAHARAN S. Support vector machine[M]. Cham: Springer, 2016.
- [26] LAFFERTY J D, MCCALLUM A, PEREIRA F C. Conditional random fields: probabilistic models for segmenting and labeling sequence data[C]//Proceedings of the 18th International Conference on Machine Learning, Williamstown, Jun 28-Jul 1, 2001. San Mateo: Morgan Kaufmann, 2001: 282-289.
- [27] COLLOBERT R, WESTON J, BOTTOU L, et al. Natural language processing (almost) from scratch[J]. Journal of Machine Learning Research, 2011, 12: 2493-2537.
- [28] YADAV V, BETHARD S. A survey on recent advances in named entity recognition from deep learning models[J]. arXiv:1910.11470, 2019.
- [29] MA X Z, HOVY E H. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF[J]. arXiv:1603.01354, 2016.
- [30] PETERS M E, NEUMANN M, IYYER M, et al. Deep contextualized word representations[J]. arXiv:1802.05365, 2018.
- [31] GUI T, MA R T, ZHANG Q, et al. CNN-based Chinese NER with lexicon rethinking[C]//Proceedings of the 28th International Joint Conference on Artificial Intelligence, Macao, China, Aug 10-16, 2019: 4982-4988.
- [32] GUI T, ZOU Y C, ZHANG Q, et al. A lexicon-based graph neural network for Chinese NER[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, Hong Kong, China, Nov 3-7, 2019. Stroudsburg: ACL, 2019: 1040-1050.
- [33] SUI D B, CHEN Y B, LIU K, et al. Leverage lexical knowledge for Chinese named entity recognition via collaborative graph network[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, Hong Kong, China, Nov 3-7, 2019. Stroudsburg: ACL, 2019: 3828-3838.
- [34] DING R X, XIE P J, ZHANG X Y, et al. A neural multi-digraph model for Chinese NER with gazetteers[C]//Proceedings of the 57th Conference of the Association for Computational Linguistics, Florence, Jul 28-Aug 2, 2019. Stroudsburg: ACL, 2019: 1462-1467.
- [35] YAN H, DENG B C, LI X N, et al. TENER: adapting transformer encoder for name entity recognition[J]. arXiv: 1911.04474, 2019.
- [36] MA R T, PENG M L, ZHANG Q, et al. Simplify the usage of lexicon in Chinese NER[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Jul 5-10, 2020. Stroudsburg: ACL, 2020: 5951-5960.
- [37] LI X N, YAN H, QIU X P, et al. FLAT: Chinese NER using flat-lattice transformer[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Jul 5-10, 2020. Stroudsburg: ACL, 2020: 6836-6842.
- [38] ZHAO S, HU M H, CAI Z P, et al. Dynamic modeling cross- and self-lattice attention network for Chinese NER [C]//Proceedings of the 35th AAAI Conference on Artificial Intelligence, 33rd Conference on Innovative Applications of Artificial Intelligence, the 11th Symposium on Educational Advances in Artificial Intelligence, Feb 2-9, 2021. Menlo Park: AAAI, 2021: 14515-14523.
- [39] WU S, SONG X N, FENG Z H. MECT: multi-metadata embedding based cross-transformer for Chinese named entity recognition[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Aug 1-6, 2021. Stroudsburg: ACL, 2021: 1529-1539.



赵山(1990—),男,安徽六安人,博士研究生,主要研究方向为自然语言处理。

ZHAO Shan, born in 1990, Ph.D. candidate. His research interest is natural language processing.



罗睿(1998—),男,江西高安人,硕士研究生,主要研究方向为自然语言处理。

LUO Rui, born in 1998, M.S. candidate. His research interest is natural language processing.



蔡志平(1975—),男,湖南益阳人,教授,博士生导师,CCF杰出会员,理论计算机科学专委会秘书长,主要研究方向为网络安全、大数据、分散计算。

CAI Zhiping, born in 1975, professor, Ph.D. supervisor, distinguished member of CCF. His research interests include network security, big data and dispersed computing.