文章编号：2095-2163(2019)01-0169-05 中图分类号：TP391.41 文献标志码：A

融合手工特征与双向 LSTM 结构的中文分词方法研究

徐 伟，车万翔，刘 挺
（哈尔滨工业大学 计算机科学与技术学院，哈尔滨 150001）

摘 要：分词作为中文自然语言处理的基础任务，既是学术界长期的研究重点，也是工业界的刚性需求。近年来，采用深度神经网络自动抽取特征并完成特征组合的方法取代传统的基于手工特征的方法，成为研究热点。不过，采用深度神经网络自动学习特征的方法在中文分词上效果并不突出。本文通过将手工特征与双向 LSTM 结构相结合，既融入了人类知识，又充分利用了深度神经网络的特征组合能力。实验结果表明，该种方法带来的分词效果提升非常明显。

关键词：中文分词；深度学习；方法融合

Chinese word segmentation by integration of handcraft-feature and Bi-LSTM

XU Wei, CHE Wanxiang, LIU Ting
(School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)

【Abstract】 As the foundational task of Chinese Natural Language Processing, Chinese word segmentation is not a long-term research focus of academia, but also a rigid demand of industry. In recent years, the method becomes popular which uses Deep Neural Networks to extract features automatically and complete feature combinations, replacing the traditional method based on handcraft features. However, the result on Chinese word segmentation is not outstanding through automatically learning features by Deep Neural Networks. This paper integrates the human knowledge and feature combination ability of Deep Neural Networks by joining handcraft feature and bidirectional LSTM structure. Experimental results show that the improvements brought by the proposed model are impressive.

【Key words】 Chinese word segmentation; deep learning; method integration

0 引 言

词作为“最小的能独立运用的语言单位”[1]，通常是上层自然语言处理任务的基础输入。分词结果的好坏，将直接影响到上层应用的效果。

考虑到中文词语之间没有明显的分隔符，因此想要获得较好的分词效果则并非易事。在 2002 年之前，学术界普遍使用基于规则或者基于统计的词典匹配方法，典型的如正向最大匹配、逆向最大匹配等[2]。2002 年，Xue 等人[3]首次提出了基于字标注的方法，次年，又使用最大熵模型实现的系统参加 Backoff-2003 评测[4]，取得优异成绩，从此，基于字符标注的中文分词方法即已迅速吸引了学界的广泛关注。基于字标注的方法首先将分词结果（词序列）转变为标签序列，然后通过序列标注模型学习字符序列与标签序列的关系来完成分词。词序列转换为标签序列依据的是每个字符在词语中出现的位置（词位）。目前常使用词位标签集为{B，M，E，S}，其中，B 表示字出现在词语的开始位置（Begin），相应的 M、E 分别表示字出现在词语的中间（Middle）和结尾（End），标签 S 表示单字成词（Single）。字符序列、词序列和标签序列的关系如图 1 所示。在基于字标注的方法成为主流后，学术界即已开始着重研究特征工程和序列标注模型改进。常见的特征包括 n-gram 特征、词典特征、字符类别特征和字符重叠信息等[5]，模型一般为 CRFs[6]或结构化感知器[7]等。2011 年，Collobert 等人[8]提出了一套针对词性标注、命名实体识别和语义角色标注的通用网络结构和学习算法。以此为起点，基于深度神经网络的中文分词（序列标注模型）研究即已陆续涌现，并获得蓬勃发展。所使用的网络结构包括多层感知器（Multi-Layer Perceptron，MLP）[9]、最大间隔张量网络（Max-Margin Tensor Neural Network，MMTNN）[10]、GRNN（Gated Recursive Neural Network）[11]、长短时记忆网络（Long-Short Term Memory，LSTM）[12]以及双向 LSTM（Bidirectional LSTM，Bi-LSTM）[13]、双向 LSTM-CRFs[14]等。在深度学习

作者简介：徐 伟（1993-），男，硕士研究生，主要研究方向：自然语言处理；车万翔（1980-），男，博士，副教授，博士生导师，主要研究方向：自然语言处理；刘 挺（1972-），男，博士，教授，博士生导师，主要研究方向：自然语言处理、文本挖掘、文本检索等。

收稿日期：2017-06-13

的浪潮下,研究者们普遍忽视了手工特征,大多数模型仅使用当前位置的字符信息(unigram)作为输入。但仍有部分学者的研究表明,在 MLP 或 MMTNN 网络中加入 bigram 特征可以显著提升模型的效果[9-10]。

字符序列	有	幸	来	到	哈	工	大	!
词序列	有幸		来到		哈工大			!
标签序列	B	E	B	E	B	M	E	S

图 1 字符序列、词序列和标签序列示例

Fig. 1 Examples of character sequence, word sequence and label sequence

研究认为,手工特征作为人类知识的体现,对模型而言是非常有价值的。本文将当下热门的双向 LSTM 结构与手工特征相融合,试图说明结合手工特征和深度神经网络的模型,相比传统中文分词方法以及不使用额外特征的深度神经网络模型效果更加优异。

1 融合手工特征与双向 LSTM 结构的中文分词方法

研究中,首先展示模型整体结构,随后依次探讨了手工特征融入方法、双向 LSTM 结构以及标签预测方法,最后给出本次研究在模型中所使用的手工特征。研究可得剖析论述如下。

1.1 模型整体结构

从宏观上,本文的模型结构可以分为 3 个层次。第一层将手工特征转变为连续值向量,将该层称为输入层;第二层为表示学习层,通过双向 LSTM 结构设计得出各位置间输入向量的特征组合;第三层为标签预测层,完成标签预测,模型的整体结构如图 2 所示。

1.2 手工特征融入神经网络模型的方法

手工特征一般可分为离散特征和连续值特征 2 类。对于深度神经网络,其输入一般是连续实值向量。因此离散特征往往需要转换为实值向量才能输入给神经网络。而对于连续值特征,可以有 2 种处理方式。一种是将连续值直接输入给网络,另一种是将连续值离散化,转换为离散特征进行处理。通常而言,将连续值离散化能够解决输入稀疏的问题,有利于模型泛化。文中就采用了离散化连续值特征的策略,因此手工特征都成为了离散特征的形式。

将离散特征转换为连续实值向量,一般是通过映射表的形式完成的。以 unigram 特征为例,记所有 unigram 构成字典 D,$n = |D|$ 为字典大小,即 unigram 个数。首先对 D 中每个 unigram 编号为 0,1,…,$n-1$,设某个 unigram 为 u,则 $i = D_u$ 即为 u 对应的编号;接着建立一个编号到值向量的映射表 M,M_i 就表示编号为 i 的 unigram 对应的实值向量。

由于选择使用了多种类型的手工特征,在每种特征均已生成了特征值到实值向量的映射后,还需要将这些向量组合起来,研究中采取拼接的方法将所有类型的特征向量组合成为一个输入向量。

整个将手工特征(离散特征)转换为连续实值向量的过程如图 3 所示。转换流程过后,就达到了将手工特征融入神经网络的目的,而这也是整个输入层面临的工作任务。

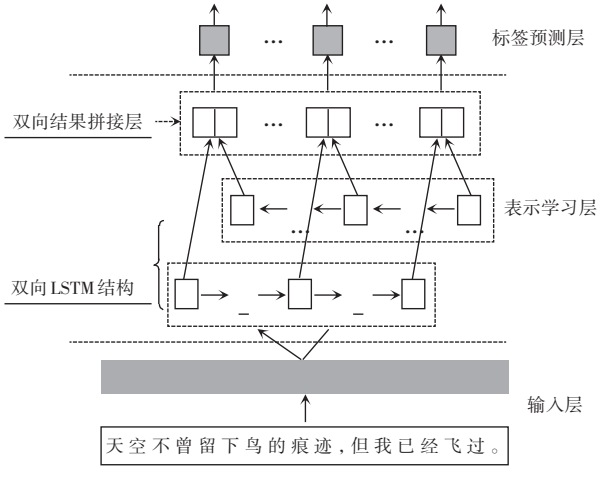

图 2 模型整体结构图

Fig. 2 The structure diagram of the model

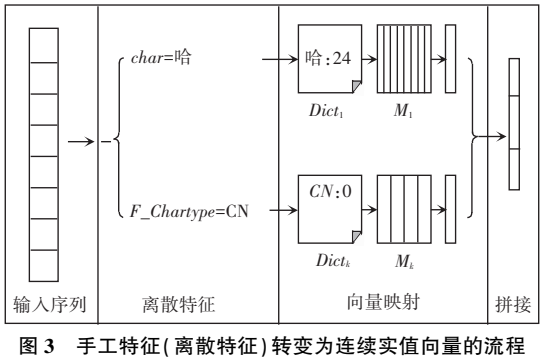

图 3 手工特征(离散特征)转变为连续实值向量的流程

Fig. 3 The process of manul features turning to continous real-value vector

1.3 双向 LSTM 结构

双向 LSTM 结构更准确的表述是在双向循环神经网络(Recurrent Neural Network,RNN)中以 LSTM 作为 RNN 单元的结构。

循环神经网络是一种处理时序输入的网络结构。RNN 结构理论上只包含一个 RNN 单元,该单元将在时间维度上反复循环地处理输入序列,并由

此而得名。RNN 单元接口和在时间序列上的展开效果如图 4 所示。RNN 考虑了前一个时刻的输出，因此被认为能够记录输入序列的历史信息。

RNN 单元有多种类型。简易的 RNN 单元只是将输入向量 x 和前一个时刻的状态 h_{t-1} 进行非线性组合，在训练较长的输入序列时容易出现梯度消失或者梯度爆炸的问题，难以训练模型。针对此问题，Hochreiter 等人[15]专门提出了 LSTM 单元，在 LSTM 单元内部引入了控制门和记忆单元，较大程度上解决了训练过程中可能出现的梯度问题。此外，其特征组合的能力也因其内部结构的复杂而更显强大。

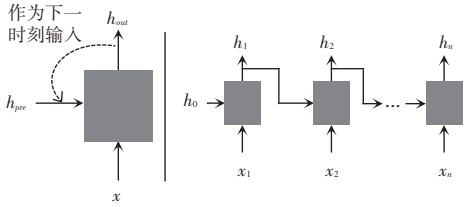

图 4　RNN 单元接口(左)以及在时间序列上的展开效果(右)

Fig. 4　RNN cell interface (left) and the expanding effect on time sequence (right)

在 RNN 的基础上，Schuster 等人[16]扩展出了双向 RNN 结构，图 5 就提供了双向 RNN 结构在长度为 3 的时间序列上的展开效果。由于前向单元能够编码历史信息，而后向单元可以融合未来的信息，因此理论上双向 RNN 结构在每个位置上都能够看到全局的信息。显而易见，这个特性对中文分词任务是非常重要的。

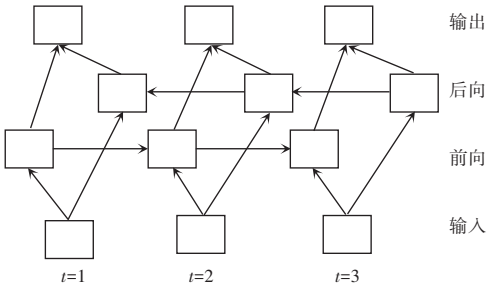

图 5　双向 RNN 的展开结构

Fig. 5　The expanding structure of Bi-RNN

双向 LSTM 结构搭建了本次研究模型中的表示学习层。该层用于设计生成各位置间输入特征向量的组合，并在每个位置上输出特征组合结果。

1.4　标签预测方法

在标签预测层，研究将在每个位置独立预测标签。就是说，对每个位置上表示学习层的输出向量均需经过 Softmax 操作才可运算得到每个标签的概率。需要注意的是，研究在每个位置上选取标签是在满足标签限制关系下概率最大的那个。举例来说，假设前一个位置的标签为 S，那么依据词位标签的限制关系，当前位置的合法标签候选集就是 $\{B, S\}$，研究会从该集合中选取概率最大的作为预测结果。

1.5　手工特征选择

研究在模型中使用的手工特征包括 unigram、bigram、词典特征和字符类别特征。这里，针对其研究要点可逐一阐释分述如下。

(1) unigram 特征。即是当前位置的字符特征，代表着原始的输入信息。

(2) bigram 特征。为当前位置字符和下一个位置字符构成的 bigram 表示。特别指出，研究中可令尾后字符(最后一个位置的后一个位置字符，在原始输入中并不存在)的表示为 EOS，用来与输入序列中最后一个位置的字符构成 bigram。bigram 表示对于中文分词尤为重要，因其不仅蕴含着语言模型的信息，同时有助于扩大输入空间，令模型在标签预测时更加容易。LSTM 单元由于兼顾了上一个位置的状态输出，因此理论上能够自动捕获一定量的 bigram 信息，这也是目前研究者们试图通过复杂化网络结构来使得模型仅根据原始输入序列(unigram 特征)就能做出良好预测的理由。但是，和把学习 bigram 特征的工作交给模型相比，直接输入 bigram 信息显然可使模型预测更趋便捷。

(3) 词典特征。使得模型具有了融合词典信息的能力。在抽取词典特征前，首先需要构建一个词典，这既可以从训练集中统计得到，也可以由外部指定。特别地，通过外部指定特定领域的词典，模型将具有一定的领域适应能力。在词表构建成功后，接下来就在输入句子的每个位置上对构建的词表做最大正向匹配，得到在每个位置上以此位置字符开始的词的最大长度 L_s、经过(不包含词首、尾)此位置的词的最大长度 L_p 和以此位置字符结尾的词的最大长度 L_e，接着将对这 3 个长度值进行离散化，设计操作如下：

$$L' = \begin{cases} L & L < 5 \\ 5 & L \geqslant 5 \end{cases} \tag{1}$$

研究中，则将离散化后的 $\{L_s', L_p', L_e'\}$ 作为当前位置的词典特征。

(4) 字符类别特征。研究将会判断当前位置的字符是否是{数字类，标点类，字母类}中的某一种。如果是，则取对应的字符类别，否则取值为“其它类”。该特征从字符类别的角度为字符提供了泛化

表示,使模型的泛化能力更强。

2 实验结果与分析

在本次实验中采用的数据集为人民日报 1998 年上半年数据(约 30 万行、七百万词)和微博数据(约五万七千行、一百万词)的合并集,具体训练集、开发集和测试集信息可见表 1。该数据集规模较大,能够充分发挥深度神经网络的能力。

表 1 中文分词数据集详情

Tab. 1 The detail of Chinese word segmentation data set

数据集	行数	词语数
训练集	337 638	7 797 529
开发集	8 000	178 792
测试集	12 500	275 440

为了直观比较模型效果,研究中使用了 2 个基准线模型。第一个模型为哈尔滨工业大学社会计算与信息检索研究中心发布的 LTP 工具[17]。与中文模型相比,LTP 使用相似的手工特征,但通过结构化感知器来拟合数据,属于线性模型。第二个基准线模型为仅使用 unigram 特征的双向 LSTM 模型,除输入特征不同外,其余部分与本文的模型结构完全一致。为了叙述方便,研究中将 unigram 的双向 LSTM 记为 Uni-Bi-LSTM,将本文的的模型记为 All-Bi-LSTM。

实验参数上,对于 LTP,设定使用默认参数完成训练;Uni-Bi-LSTM 和 All-Bi-LSTM 的参数设置可见表 2。

研究中选择使用 $F1$ 值作为评价指标,实验结果详见表 3。根据实验结果,融合手工特征和双向 LSTM 结构的方法取得了最优的效果,且相比其余 2 种方法提升明显。LTP 作为传统中文分词方法的代表,在开发集和测试集上均取得了不错的效果,但是其在测试集上的 $F1$ 值相比在开发集上低 0.13 个百分点,高于 Uni-Bi-LSTM 的 0.11 和 All-Bi-LSTM 的 0.07,这表明在此数据集上基于 Bi-LSTM 的方法泛化能力更强。仅使用 unigram 特征的 Uni-Bi-LSTM 方法效果最差,相比 All-Bi-LSTM 在开发集上低 1.65 个百分点,测试集上低 1.69 个百分点。这表明仅是通过双向 LSTM 结构去自动学习输入中的特征还是不够的,引入手工特征能够显著提升模型效果。最后,同样基于手工特征,使用 Bi-LSTM 结构的 All-Bi-LSTM 方法比使用结构化感知器的 LTP 效果优异,在开发集和测试集上分别高 0.28 和 0.34 个百分点,这说明 Bi-LSTM 结构的特征组合能力更加强大。

表 2 Uni-Bi-LSTM 和 All-Bi-LSTM 参数设计

Tab. 2 Parameter details of Uni-Bi-LSTM and All-Bi-LSTM

参数项	Uni-Bi-LSTM	All-Bi-LSTM
unigram 特征向量维度	50	25
bigram 特征向量维度	-	50
词典特征维度	-	5
字符类型特征维度	-	5
LSTM 单元隐层维度	100	100
LSTM 单元层数	1	1
参数优化方法	SGD	SGD
学习率	0.01	0.1
迭代轮次	15	15

注:学习率在{0.01, 0.05, 0.1}三个候选值中搜索,取在开发集上取得最优效果的值作为最终参数值。

表 3 实验结果

Tab. 3 Experimental results %

方法	开发集 $F1$ 值	测试集 $F1$ 值
LTP	96.69	96.56
Uni-Bi-LSTM	95.32	95.21
All-Bi-LSTM	**96.97**	**96.90**

3 结束语

本文探讨了将手工特征和双向 LSTM 结构相融合的中文分词方法,该方法既融入了人类的先验知识,又引入了深度神经网络模型对于输入特征的组合能力。实验结果表明,本文的方法相比传统基于手工特征的方法有较大提升,相比不使用额外手工特征的双向 LSTM 模型则有长足可观的大幅提升。这也进一步验证了本文提出的融合手工特征和双向 LSTM 结构的中文分词方法的有效性。

参考文献

[1] GB/T 13715-92. 信息处理用现代汉语分词规范[S]. 北京: 中国标准出版社, 1993.

[2] 黄昌宁, 赵海. 中文分词十年回顾[J]. 中文信息学报, 2007, 21(3): 8-19.

[3] XUE Nianwen, CONVERSE S P. Combining classifiers for Chinese word segmentation [C]//Proceedings of the first SIGHAN workshop on Chinese language processing - Volume 18. Stroudsburg, PA, USA: Association for Computational Linguistics, 2002: 57-63.

[4] XUE Nianwen, SHEN Libin. Chinese word segmentation as LMR tagging[C]//Proceedings of the second SIGHAN workshop on Chinese language processing - Volume 17. Sapporo, Japan: Association for Computational Linguistics, 2003: 176-179.

(下转第 177 页)