doi:10.3969/j.issn.1001-4616.2022.01.018

基于组合深度模型的现代汉语数量名短语识别

施寒瑜[1],曲维光[1,2],魏庭新[2,3],周俊生[1],顾彦慧[1]

(1.南京师范大学计算机与电子信息学院/人工智能学院,江苏 南京 210023)
(2.南京师范大学文学院,江苏 南京 210097)
(3.南京师范大学国际文化教育学院,江苏 南京 210097)

[摘要] 数量名短语的识别是识别由数量短语修饰的名词短语左右边界的研究. 以往研究中,基于统计学习模型的数量短语识别方法依赖人工特征,需要通过专家知识构建知识库来实现对“数词+量词”短语的识别. 本文在以往研究基础上纳入“名词”形成“数词+量词+名词”等八类数量名短语,并采用深度学习方法解决这一边界识别任务. 通过 BERT 模型对原始文本进行上下文特征表示,利用 Lattice LSTM 模型字词结合的思想将标准分词作为软特征融入文本字符级的特征表示中,最后通过 CRF 全局约束识别数量名短语边界. 实验结果表明,本文方法在 AMR 语料上达到较优结果,精确率、召回率、*F*1 值分别为 80.83%,89.78%,85.07%.

[关键词] 数量名短语识别,BERT,Lattice LSTM;CRF
[中图分类号]TP391　[文献标志码]A　[文章编号]1001-4616(2022)01-0127-09

Quantity Noun Phrase Structure Recognition Based on Combined Deep Learning Model

Shi Hanyu[1],Qu Weiguang[1,2],Wei Tingxin[2,3],Zhou Junsheng[1],Gu Yanhui[1]

(1.School of Computer and Electronic Information/School of Artificial Intelligence,Nanjing Normal University,Nanjing 210023,China)
(2.School of Chinese Language and Literature,Nanjing Normal University,Nanjing 210097,China)
(3.International College for Chinese Studies,Nanjing Normal University,Nanjing 210097,China)

Abstract:The research on recognition of quantity noun phrases is the identity of the left and right boundaries of quantity noun phrases. In previous studies,this task focuses on the recognition of quantity phrase and relies on artifical features which are constructed by experts based on statistical learning models. In this paper,we aim at the recognition of quantity noun phrases which have 8 subtypes and propose a neural network model to address the issue. Firstly,BERT is used to represent the contextual features of the original text. Then,the standard word segmentation is incorporated into the feature representation of the text character level as a soft feature by using the idea of Lattice LSTM model. Finally,the left and right boundaries of the“quantity noun phrase”are identified by the CRF global constraint. The experimental results show that this method achieves the better results and the precision,recall and *F*1 value reaches 80.83%,89.78%,85.07% respectively in the corpus of CAMR.

Key words:the recognition of quantity noun phrases,BERT,lattice LSTM,CRF

数量短语是现代汉语中主要起计数作用的短语,包括统计数目多少或计算次序先后等. 数量短语作为现代汉语中常用的语法结构,一般用于修饰名词短语,并以“数词+量词+名词”的常规语序出现,然而在真实语境中,语言表达通常遵循经济原则,若名词中心语在上下文有提示,数量短语通常会以中心语省略的形式出现在日常语言中使用. 如“店家无奈,只好又给武松筛酒…武松前后共吃了十八碗.”这种数量短语的特殊用法黎锦熙将其称之为“替代中心词”[1],即数量词修饰的中心语名词被省略,由前面的数量词替代,省略条件为中心语名词在上下文至少出现一次;并且该类数量短语在句法上可独立充当主语、宾语等句子成分.

收稿日期:2020-12-26.
基金项目:国家自然科学基金项目(61772278、61472191)、国家社科基金项目(21&ZD288、18BYY127).
通讯作者:曲维光,博士,教授,研究方向:自然语言处理. E-mail:wgqu_nj@163.com

近年来,有不少关于数量短语识别的研究工作. 以往大部分研究是基于知识库和规则的方法实现“数词+量词”短语识别. 白晓革等[2]将数量短语的构成模式细分为数词短语、数量短语、模糊数词短语、模糊数量短语、序数数量短语、范围数量短语、特殊符合量词短语和指量短语,并加以概念层次网络(hierarchical network of concepts,HNC)世界知识库以及数量短语 3 大词库对语料进行分析和提取,正确率和召回率达到 90%. 张玲等[3]在探讨数量短语构成模型的基础上构建一个为数量短语识别提供词汇知识和短语结构知识的数据库,在 1 万字人民日报的新闻语料中正确率为 90.9%,召回率为 98.7%. 熊文等[4]在张玲等[2-3]的基础上又提出一种基于规则不依赖于分词的中文数量短语的识别,该方法在人民日报的未标注语料进行了识别,召回率达到 98.7%,精度为 90.9%. 以上研究都是针对“数词+量词”边界的识别. 然而仅仅识别数量短语在实际应用中是不够的,数量短语是名词中心语的修饰成分,识别数量短语只能实现 HNC 句类分析的前置处理. 本文将数量名短语作为一个整体进行识别,有利于数量信息的抽取,利于机器翻译、问答等相关工作的实现,同时有利于解决中文抽象语义表示(CAMR)模式中数量短语增补外部概念节点(名词性词汇语类缺省添加),补全数量短语省略的工作,有助于 CAMR 语义自动解析工作. 目前对于“数+量+名”短语的识别研究较少,且使用的是基于统计学习模型的识别方法. 方芳等[5]将数量名短语归纳为“基数词+量词+名词”“基数词+量词+修饰+名词”“序数词+量词+名词”“数词+名词”以及“指示代词+量词+名词”等 5 类,并基于规则库的方法在 240 万字的当代新闻小说语料上进行识别,调和平均值 $F1$ 达到 80%. 但该文只针对一种语序的数量名短语的识别,对于自然语言中出现的其他形式的数量名短语并未给出解决方案.

本文通过对语料的统计分析,发现数量名短语除了上述五类之外,还有 3 种情况:“名词+数词+量词(倒装)”“不定数词+(量词)+(修饰)+名词”和“数量名短语省略”这 3 种情况在自然语料中所占比例有 17.44%. 因此为扩大识别数量名短语形式的规模,涵盖更全面的具有计数意义的短语,提升模型泛化性,本文将其均纳入研究范围,实现对如上 8 种类型数量名短语的边界识别.

在自然语言处理中将此类短语结构化识别问题归于序列标注问题,相类似的有词性标注、命名实体识别等任务. 近几年,序列标注问题通过神经网络的训练方法表现出很好的性能. Collobert 等[6]首次实现将 CNN 模型与 CRF 结合应用于命名实体识别任务,在 CoNLL2003 的语料集上取得较好效果. 随后,Huang 等[7]采用一个人工设计语义特征的 BLSTM-CRF 模型在 CoNLL2003 语料上将 $F1$ 值提升到 88.83%. Chiu 和 Nichols[8]将 CNN 和 LSTM 结合,在 CoNLL2003 语料上将命名实体识别任务 $F1$ 值提升至 91.62%.

虽然神经网络在命名实体识别上取得优异的效果,但将这些模型迁移到现代汉语数量名短语的识别领域中还存在若干问题. 相比命名实体识别,数量名短语识别有以下几个难点:(1)数量名词短语有过多的干扰项,如年月日、度量衡等这些非数量名短语的专有数量短语;(2)现代汉语中数词千变万化,量词的种类繁多,数量名短语的组合方式多样;(3)数词或量词或名词的省略现象在现代汉语的使用中尤为常见,这导致识别过程中边界模糊问题成为难点.

本文通过深度学习方法,减弱对人工特征设计和专家知识的依赖,实现数量名短语的识别. 本文利用中文抽象语义表示(CAMR)的语义表示体系[9]精确辨别数量名短语的左右边界,解决手工标注时往往遇到边界模糊的问题. 本文语料采用中文抽象语义表示[10](CAMR)语料,在该语料上识别效果 $F1$ 值达到 85.07%.

1 深度学习模型

1.1 模型整体框架

本文 BERT-Lattice LSTM-CRF 模型一共有 3 部分组成,如图 1 所示:

(1)BERT 预训练模块:采用 BERT 模型进行预训练,对输入序列中的每个字符生成字符级特征表示,弥补数据集较少情况下,特征学习不充分的现象;

(2)Lattice LSTM 特征获取模块:采用 Lattice LSTM 模块进行特征表示,该模块融合正确分词的软特征. 模块将 BERT 模型的输出和名词向量融合,解决因分词错误而导致的错误传递现象并进一步丰富特征表示;

(3)CRF 解码模块:采用线性 CRF 模块,该模块将 Lattice LSTM 模型的输出的特征表示解码,获取一

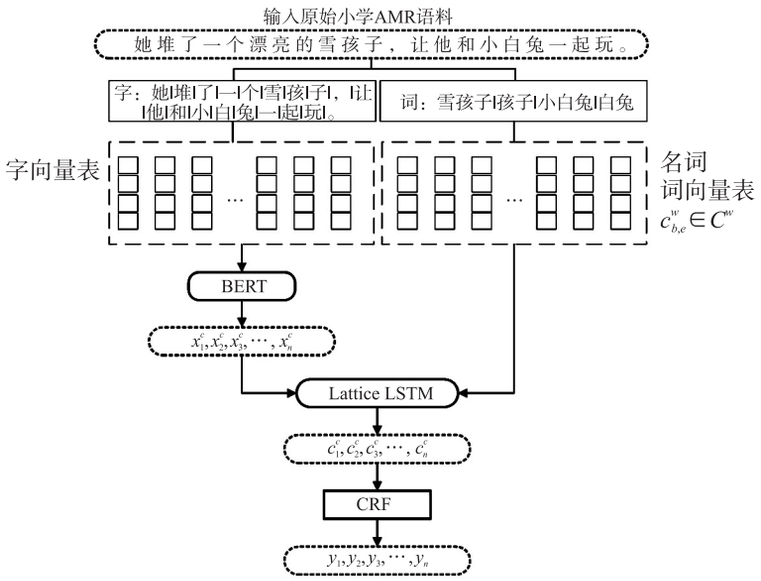

图 1　BERT-Lattice LSTM-CRF 模型框架

Fig. 1　BERT-Lattice LSTM-CRF model

个最优的标注序列.

1.2　BERT 预训练

在自然语言处理领域中为了更好地表示文本的特征,解决一词多义问题,通常使用 ELMo(Embedding from Language model)[11],一种基于语境的深度词表示模型. 而 2018 年由 Devlin 等[12]提出 BERT(Bidirectional Encoder Representation and Transformers)模型在 ELMo 模型基础上改进,通过超大数据、巨大模型、和极大的计算开销训练而成,并在 11 个自然语言处理任务中取得优异结果.

BERT 模型的子结构是 Transformer[13]双向编码器,它摒弃了 RNN 的循环网络结构,把 Transformer 编码器当作模型的主题结构,利用自注意力机制对句子建模. 本文将语料输入 BERT 模型预处理,可以充分学习语料的字符之间、词语之间以及句子与句子之间关系特征,为输入语料的每个字符生成基于当前语境上下文的动态字符级嵌入向量,解决传统词嵌入方法将不同语境中的同一单词映射到相同语义空间的问题,提升字符级嵌入向量的文本特征表示能力.

如图 2 所示,为获取语料预训练的字符级嵌入向量,将原始语料处理成 token 嵌入,segment 嵌入,position 嵌入 3 部分输入. BERT 层通过联合调节内部的双向 Transformer 编码器,利用自注意力机制学习上下文中其余字符对当前字符的贡献程度,从而增强上下文语义信息的获取. 最终,编码生成基于当前语境上下文的字符级嵌入向量.

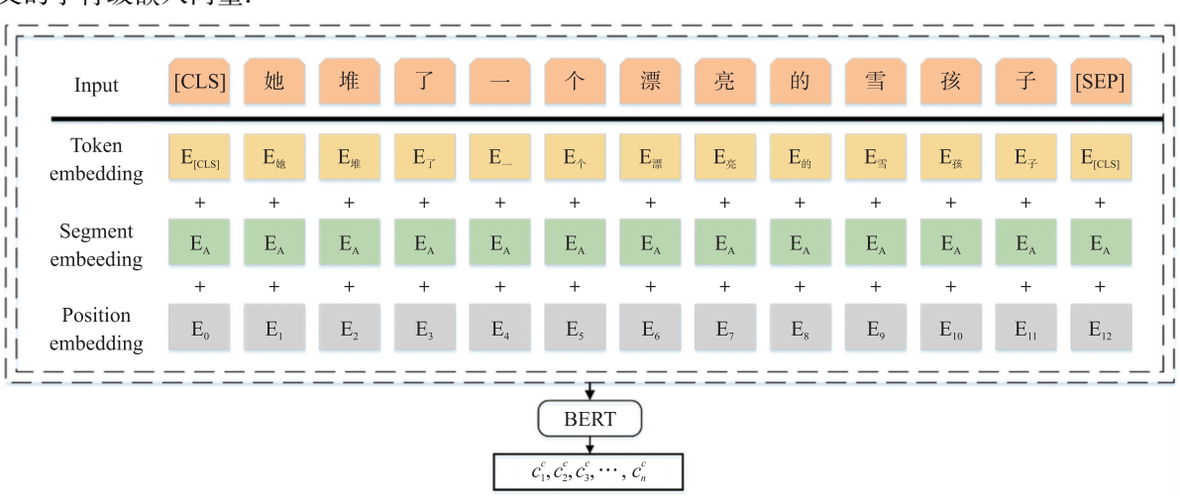

图 2　BERT 预处理字符向量

Fig. 2　Pre-trained character vectors by BERT

本文将 AMR 语料输入 BERT 模型,预训练生成融合上下文信息的字符级嵌入向量,并将该字符级嵌入向量输入 Lattice LSTM 模型,通过 Lattice LSTM 模型生成含有词语信息的字符级特征表示.

1.3 Lattice LSTM 模块

LSTM(Long Short-Term Memory)长短时记忆网络[14]是 RNN(Recurrent Neural Network)的一种. 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. 在基于字符级的 LSTM 模型中,每个字符 $\boldsymbol{c}_j$ 用字符输入向量来表示:

$$\boldsymbol{x}_j^c=\boldsymbol{e}^c(\boldsymbol{c}_j), \tag{1}$$

其中 $\boldsymbol{e}^c$ 是字符嵌入向量查询表. 基本的 LSTM 结构是由字符单元向量 $\boldsymbol{c}_j^c$ 和字符 $\boldsymbol{c}_j$ 对应的隐藏向量 $\boldsymbol{h}_j^c$ 构成,其中 $\boldsymbol{c}_j^c$ 用于记录从当前句子开始到字符 c_j^c 得到的信息流,而 $\boldsymbol{h}_j^c$ 则是作为 CRF 序列标注的输入. 基本的 LSTM 模型公式如下:

$$i_j^c=\sigma(W_i^c h_{j-1}^c+U_i^c x_j^c+b_i^c), \tag{2}$$

$$f_j^c=\sigma(W_f^c h_{j-1}^c+U_f^c x_j^c+b_f^c), \tag{3}$$

$$\tilde{c}_j^c=\tanh(W_c^c h_{j-1}^c+U_c^c x_j^c+U_c^c), \tag{4}$$

$$c_j^c=f_j^c\cdot c_{j-1}^c+i_j^c\cdot\tilde{c}_j^c, \tag{5}$$

$$o_j^c=\sigma(W_o^c h_{j-1}^c+U_o^c x_j^c+b_o^c), \tag{6}$$

$$h_j^c=o_j^c\cdot\tanh(c_j^c). \tag{7}$$

$\sigma(\ )$表示的是 sigmoid 激活函数,$\cdot$是点乘运算,tanh 表示双曲正切激活函数. $\boldsymbol{x}_j^c$ 是在时间 j 输入的字符向量,并且 h_j^c 是在时间 j 存储所有有用信息的隐藏状态向量. $\boldsymbol{U}_i^c,\boldsymbol{U}_f^c,\boldsymbol{U}_c^c,\boldsymbol{U}_o^c$ 表示输入 $\boldsymbol{x}_j^c$ 的不同门的权重矩阵,$W_i^c,W_f^c,W_c^c,\boldsymbol{W}_o^c$ 是隐藏状态 h_t 的权重矩阵. $\boldsymbol{b}_i^c,\boldsymbol{b}_f^c,\boldsymbol{b}_c^c,\boldsymbol{b}_o^c$ 表示偏置向量.

Lattice LSTM[15]利用显式单词信息与先前生成的字符信息结合,避免字符分词的错误传递从而提高数量名短语识别的正确性. 此时字符表示 $\boldsymbol{c}_j^c$ 的计算考虑句子中匹配的词典子序列 $\boldsymbol{w}_{b,e}^d$,每个子序列 $\boldsymbol{w}_{b,e}^d$用如下公式表示:

$$x_{b,e}^w=e^w(w_{b,e}^d). \tag{8}$$

每个单词单元 $\boldsymbol{c}_{b,e}^w$用于表示从句子的开始至当前状态 $x_{b,e}^w$. 单词单元 $\boldsymbol{c}_{b,e}^w$用如下方法计算得:

$$i_{b,e}^w=\sigma(W_i^w x_{b,e}^w+U_i^w h_b^c+b_i^w), \tag{9}$$

$$f_{b,e}^w=\sigma(W_f^w x_{b,e}^w+U_f^w h_b^c+b_f^w), \tag{10}$$

$$\tilde{c}_{b,e}^w=\tanh(W_c^w x_{b,e}^w+U_c^w h_b^c+b_c^w), \tag{11}$$

$$c_{b,e}^w=f_{b,e}^w\cdot c_b^c+i_{b,e}^w\cdot\tilde{c}_{b,e}^w. \tag{12}$$

其中 $\boldsymbol{i}_{b,e}^w$和 $\boldsymbol{f}_{b,e}^w$是一组输入门和遗忘门. 这里单词单元不再需要输出门因为最终的序列标注是针对字符级别.

通过 $\boldsymbol{c}_{b,e}^w$有许多信息汇入用于表示当前字符 j 的向量 $\boldsymbol{c}_j^c$ 中. 例如,$\boldsymbol{c}_{10}^c$中包含 $\boldsymbol{x}_8^c$(雪),$\boldsymbol{c}_{9,10}^w$(孩子)和 $\boldsymbol{c}_{8,10}^w$(雪孩子)的信息,如图 3 所示. 我们把所有可以与当前状态 e 组成词语的起始位 b 的词语信息 $\boldsymbol{c}_{b,e}^w$全部融入 $\boldsymbol{c}_e^c$ 单元中. 因此我们针对每一个子序列 $\boldsymbol{c}_{b,e}^w$采用额外的输入门 $\boldsymbol{i}_{b,e}^c$来控制输入 $\boldsymbol{c}_{b,e}^c$单元的信息:

$$i_{b,e}^c=\sigma(W_i^l x_e^c+U_i^l c_{b,e}^w+b_i^l). \tag{13}$$

最终输入 CRF 模块进行序列标注的文本特征表示 $\boldsymbol{c}_j^c$ 单元计算如下:

$$\boldsymbol{c}_j^c=\sum_{b\in\{b'|w_{b',j}^d\in\mathbb{D}\}}\alpha_{b,j}^c\cdot c_{b,j}^w+\alpha_j^c\cdot\tilde{c}_j^c. \tag{14}$$

其中 $\boldsymbol{\alpha}_{b,j}^c$和 $\boldsymbol{\alpha}_j^c$ 是基于 $\boldsymbol{i}_{b,j}^c$和 $\boldsymbol{i}_j^c$ 两个输入门归一化得到的概率分布:

$$\alpha_{b,j}^c=\frac{\exp(i_{b,j}^c)}{\exp(i_j^c)+\sum_{b'\in\{b''|w_{b'',j}^d\in\mathbb{D}\}}\exp(i_{b',j}^c)}, \tag{15}$$

$$\alpha_j^c=\frac{\exp(i_j^c)}{\exp(i_j^c)+\sum_{b'\in\{b''|w_{b'',j}^d\in\mathbb{D}\}}\exp(i_{b',j}^c)}. \tag{16}$$

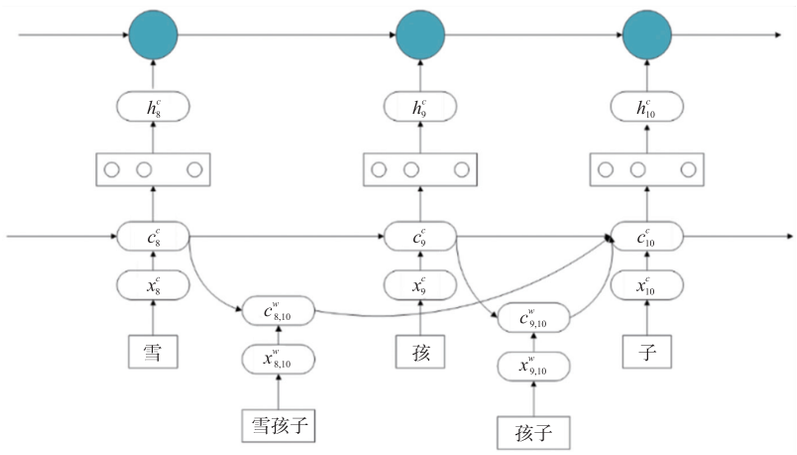

图 3　Lattice LSTM 字符和词的融合模型

Fig. 3　Lattice LSTM model with Character and word information fusion

1.4　CRF 模块

本文将条件随机场(conditional random field,CRF)[16]融合到 Lattice LSTM 模块中,对 Lattice LSTM 的输出进行处理,获得全局最优的标注序列,其中输入 CRF 的隐藏层向量 $\boldsymbol{h}_j^c$ 通过公式(7)计算得到. 对于一个句子 $\boldsymbol{S}=\{\boldsymbol{W}_1,\boldsymbol{W}_2,\boldsymbol{W}_3\cdots,\boldsymbol{W}_n\}$ 送入网络中训练,定义矩阵 $\boldsymbol{P}$ 是 Lattice LSTM 层的输出结果,其中 $\boldsymbol{P}$ 的大小 $n\times m$,n 是单词个数,m 是标签的种类. 定义 $\boldsymbol{P}_{ij}$ 代表句子中第 i 个单词的第 j 个标签的概率. 对于一个预测序列 $\boldsymbol{y}=\{\boldsymbol{y}_1,\boldsymbol{y}_2,\cdots,\boldsymbol{y}_n\}$,它的概率可以表示为:

$$K(X,y)=\sum_{i=0}^{n}A_{y_i,y_{i+1}}+\sum_{i=1}^{n}P_{i,y_i}. \tag{17}$$

式中,矩阵 $\boldsymbol{A}$ 是转移矩阵,例如 $\boldsymbol{A}_{ij}$ 表示由标签 i 转移到 j 的概率,y_0,y_n 则是预测句子起始和结束的标注,因此 $\boldsymbol{A}$ 是一个大小为 m+2 的方阵. 所以在原语句 $\boldsymbol{S}$ 的条件下产生标注序列 $\boldsymbol{y}$ 的概率为:

$$p(y\mid S)=\frac{\boldsymbol{e}^{K(X,y)}}{\sum_{\tilde{y}\in Y_X}K(X,\tilde{y})}. \tag{18}$$

式中,$\tilde{\boldsymbol{y}}$ 代表真实的标注值.

在训练过程中标注序列的似然函数:

$$\log(p(y\mid S))=K(X,y)-\log\left(\sum_{\tilde{y}\in Y_X}\boldsymbol{e}^{K(X,\tilde{y})}\right). \tag{19}$$

其中,$\boldsymbol{Y}_x$ 表示所有可能的标注集合,包括不符合 BIOES[17]标注规则的标注序列. 通过式(19)得到有效合理的输出序列. 预测时,由式(20)输出整体概率最大的一组序列:

$$y^*=\underset{\tilde{y}\in Y_X}{\operatorname{argmax}}K(X,\tilde{y}). \tag{20}$$

1.5　训练参数

训练过程中,使用具有批量大小 10 和动量 0.9 的小批量随机梯度下降(SGD)执行参数优化. 我们选择学习率为 0.015. 同时,通过实验发现在 LSTM 的输入和输出部分增加 Dropout 可以减轻模型过拟合的问题,Dropout[18]值选取了 0.5.

我们探索了其他更复杂的优化算法,如 AdaDelta[19],Adam[20]或 RMSProp[21],但它们都没有在我们的动量和梯度削减中改进 SGD.

2　实验结果

2.1　语料介绍

本文采用的语料是小学 1-6 年级语文教材(人教版)的抽象语义表示(AMR)语料[10],共 8 587 个中文句子,其中数量名短语总计有 6 142 个. 本文将语料以 7∶1 的比例分为训练集 7 587 句,数量名短语 5 320 个;测试集 1 000 句,数量名短语 822 个. 数量名短语分成如下 8 种类型:(1)基数词+量词+名词;(2)基数词+量词+修饰+名词;(3)序数词+量词+名词;(4)数词+名词;(5)指示代词+量词+名词;(6)名

词+数词+量词(倒装);(7)不定数词+(量词)+(修饰)+名词;(8)数量名短语省略. 如表 1 所示:

表 1　数量名短语形式短语分类

Table 1　Categories of quantity noun phrases

类型	样例	训练集	测试集	总计
基数词+量词+名词	一+口+井 三+个+学生 一+个+清晨	1 770	223	1 993
基数词+量词+修饰+名词	一+个+穿着裙子的+洋娃娃 一+丛丛+绿绿的+凤尾竹	659	136	795
序数词+量词+名词	第一+个+人 首+个+国家	96	35	131
数词+名词	三百+人 五万+同胞	1 032	162	1 194
指示代词+量词+名词	这+个+房屋 那+种+车辆 各+色+鲜艳的+花朵	837	121	958
名词+数词+量词	兵马俑+近八千+个 红丝+几万+条	67	5	72
不定数词+(量词)+(修饰)+名词	有的+女孩 许多+卡片 一些+漂亮的+礼物	525	102	627
数量名短语省略 数词+量词+(名词)、 数词+(名词)、 不定数词+(名词)	手里还拿着一把(火柴) 杀敌一千,自损八百(人) 有的(葡萄)运到阴房制成葡萄干	334	38	372
总计		5 320	822	6 142

在语料处理方面,为了能够清楚地表示语料中待识别的数量名短语,本文采用序列化标注任务中常用的 BIOES[16] 标注方式. 数量名短语的左边界标记为 B-NUM,右边界标记为 E-NUM,中间文本标记为 I-NUM,该方式能更清楚的划分数量名短语的左右边界. 本文为了消除分词错误造成的错误传递问题,在处理语料时以中文字符为单位进行标记. 该方法减少因中文分词歧义而造成边界模糊的问题,使得数量名短语左右边界更为精确便于深度学习模型的训练和测试.

2.2　实验结果及分析

本文通过对比实验的结果来分析各个模块在模型中起到的作用,实验结果如表 2,3 所示. 其中,BLSTM-CRF,CRF,Lattice LSTM-CRF 三种模型均采用基于 Baidu Encyclopedia 预训练的字向量;BERT-CRF,BERT-BLSTM-CRF,BERT-Lattice LSTM-CRF 3 种模型中 BERT 模型采用 BERT-Base,Chinese 字符级预训练模型.

表 2　模型结果对比

Table 2　Comparison of model results

模型	Test	Gold num	Pred num	Right num
BLSTM-CRF	1 000	822	811	573
CRF	1 000	822	776	599
BERT-CRF	1 000	822	877	668
Lattice LSTM-CRF	1 000	822	836	676
BERT-BLSTM-CRF	1 000	822	907	714
BERT-Lattice LSTM-CRF	1 000	822	913	738

表 3　模型精确率、召回率、*F*1 值

Table 3　Precision, recall and *F*1 results

模型	精确率/%	召回率/%	*F*1 值/%
BLSTM-CRF	70.65	69.79	70.22
CRF	77.19	72.96	75.02
BERT-CRF	76.16	81.26	78.62
Lattice LSTM-CRF	80.86	82.34	81.59
BERT-BLSTM-CRF	78.72	86.86	82.59
BERT-Lattice LSTM-CRF	80.83	89.78	85.07

本文的基线模型采用传统的机器学习方法 CRF 模型. 传统的机器学习方法在规则性较强的识别任务上可取得较好效果. 本文研究的数量名短语大部分左边界由具体的数词组成,数词的表达方式又明显有别于汉语其他词类,加之量词基本是一个封闭集合,所以 CRF 模型能够充分利用相邻标签关系实现边界识别. CRF 模型在本文数量名短语边界识别任务中 *F*1 值达到 75.02%.

在 CRF 模型基础上,本文加入 BERT 模型,利用 BERT 自身 Transformer 模块,通过自注意力机制使得

CRF 获取具有文本上下文信息的字符级特征表示,丰富 CRF 模型的特征获取. 实验证明,在加入预训练模型后,BERT-CRF 模型 $F1$ 值为 78.62%,比基线 CRF 模型的 $F1$ 值高出了 3.60%,BERT-CRF 模型的召回率也明显优于 CRF 模型.

BLSTM-CRF 深度学习模型效果远不如统计学习 CRF 模型,原因在于该模型训练数量名短语此类左边界相对封闭,右边界开放的短语结构时,从右往左读取的右边界信息的无规律性,导致右边界识别效果不佳;同时 BLSTM-CRF 模型受到训练数据集较少的约束而不能充分表示每个字符在文本中的特征信息,从而导致识别效果远不如统计学习 CRF 模型. 为弥补小数据集带来的缺陷,本文融合预训练模型 BERT,实验表明经过预训练处理后识别效果有显著提升. 相较 BERT-CRF 模型,BERT-BLSTM-CRF 模型中的双向 LSTM 模块能将 BERT 预训练获得的上下文字符向量训练出更符合上下文的特征表示,并将 $F1$ 值提升 3.97%. 由此可见深度学习方法在数量名短语边界识别的任务中能发挥其优势. 相比仅使用字符信息的 BLSTM 模型,本文认为结合了正确切分的名词信息的 Lattice LSTM 模型更适合处理中文语料. 实验结果表明,Lattice LSTM-CRF 比 BLSTM-CRF 模型的 $F1$ 值提升 11.37%. Lattice LSTM 能够融入正确分词软特征,丰富字符向量的上下文特征,并且能够减少模型因分词错误而造成的错误传递问题,从而提升模型效果. 本文最优模型 BERT-Lattice LSTM-CRF 将 BERT 预训练、Lattice LSTM 字词融合模型以及线性 CRF 模型相结合,将数量名短语的边界识别的 $F1$ 值提升至 85.07%.

BERT-Lattice LSTM-CRF 模型在本文 AMR 小学语料 1000 句测试集的 822 个数量名短语中,按照数量名短语不同类别,统计不同类别数量名短语的识别效果,如表 4 所示.

表 4　不同类别数量名短语精确率、召回率及 $F1$ 值

Table 4　Precision, recall and $F1$ results of different quantity noun phrases categories

数量名短语分类	精确率/%	召回率/%	$F1$ 值/%
基数词+量词+名词	89.24	97.55	93.21
基数词+量词+修饰+名词	83.77	89.58	86.57
序数词+量词+名词	85.71	85.71	85.71
数词+名词	67.91	88.48	76.84
指示代词+量词+名词	84.91	91.84	88.23
名词+数词+量词(倒装)	66.67	80.00	72.73
不定数词+(量词)+(修饰)+名词	78.01	82.27	80.29
数量名短语省略	90.90	78.94	84.51
总计	80.83	89.78	85.07

在数量名短语中,“基数词+量词+名词”是数量名短语中短语形式最为常规的一类,也是最为常见的一类,在测试集中占 24.82%,其识别效果 $F1$ 值为 93.21%. “基数词+量词+名词”“基数词+量词+修饰+名词”“指示代词+量词+名词”和“不定数词+(量词)+(修饰)+名词”这四种类别在现代汉语中出现频率较高,因而深度学习模型能够充分这四种类别的短语特征,取得相对较好的识别效果. “数词+名词”类别的 $F1$ 值为 76.84%,在所有数量名短语中识别效果较差,原因在于中文常见“数+名+数+名”短语的词语,如“一文一武”“千言万语”等词语识别成“一文”“一武”“千言”和“万语”等非数量名短语从而导致该类数量名短语识别的精确率和 $F1$ 值偏低. “名词+数词+量词(倒装)”类别的 $F1$ 值为 72.73%,在数量名短语识别中识别结果最低,其主要问题在于该类型数量名短语在训练语料中出现频率低,模型无法充分学习该类数量名短语的特征. 在以往数量短语识别中,仅仅识别未缺省的数量名短语,无识别“省略数量名短语”的相关工作. 本文将“数量名短语省略”类别纳入识别范畴,填补该类短语识别工作的空白,模型取得较优越效果,$F1$ 值达到 84.51%. 同时,该结果对于后期“省略数量名短语”的补全工作具有先导意义.

以往工作仅仅针对“基数词+量词+名词”“基数词+量词+修饰+名词”“序数词+量词+名词”“数词+名词”和“指示代词+量词+名词”这五种数量名短语识别(以符号 JLM 表示). 本文模型针对这五类数量名短语的识别效果远超方芳[5]等 80%的调和平均值,如表 5 所示,$F1$ 值达到 86.15%. 本文在 JLM 数量名短语的基础上将“名词+数词+量词(倒装)”“不定数词+(量词)+(修饰)+名词”和“数量名短语省略”三种数量名短语纳入识别范围,并实现 $F1$ 值 85.07%的

表 5　JLM 数量名短语精确率、召回率和 $F1$ 值

Table 5　Precision, recall and $F1$ results of JLM quantity nonn phrase

	精确率/%	召回率/%	$F1$ 值/%
JLM	81.03	91.95	86.15

识别效果.

综上,本文工作扩大数量名短语的识别范畴,填补“省略数量名短语”识别工作的空白,并通过 BERT-Lattice LSTM-CRF 组合深度模型实现目前最优的数量名短语识别效果.

3 结论

本文针对现代汉语数量名短语识别任务,通过 BERT 预训练模型获得具有上下文信息的字符向量表示,并通过 Lattice LSTM 网络将字和词的信息融合,丰富字符级网络的词语信息,减少分词错误造成的错误传递,最后通过 CRF 全局约束完成数量名短语识别工作,在 AMR 小学语料中取得较好性能. 主要结论如下:

(1)在现代汉语数量名短语识别任务中,人工特征和知识库对于结果的影响很大,但构建合适的人工特征需要大量的特征提取实验,导致了系统的成本提升、泛化能力下降. 本文采用的深度学习方法可以自动获取数量名短语的结构特征,大大降低人工获取特征的工作量并提升模型泛化能力.

(2)本文在 JLM 数量名短语的基础上扩大识别范畴,并采用深度学习模型,构建 BERT-Lattice LSTM-CRF 组合模型,在不使用任何人工特征的情况下,实现数量名短语边界识别 $F1$ 值达到 85.07%.

(3)实验表明,BERT 模型预训练字符向量的加入,使得模型实现对于含有不常用数词和量词这类数量名短语更有效地识别;通过 Lattice LSTM 网络把正确分词的信息融入字符向量的表示,减少因分词错误导致的边界偏移问题;最后采用线性 CRF 算法提升对含有多修饰词、边界模糊的现代汉语数量名短语的识别能力.

综上,本文研究基于组合神经网络的数量名短语识别方法. 通过 BERT 预训练模块学习文本字符级上下文特征表示,利用 Lattice LSTM 字词融合思想,结合线性 CRF 全局约束完成识别任务. 在下一步的研究中,我们将进一步扩大语料规模,使得模型学习更多更丰富的特征信息以取得更好效果. 随后,我们在识别数量名短语的基础上,解决中文抽象语义表示(CAMR)模式中数量短语增补外部概念节点(名词性词汇语类缺省添加),补全数量短语省略的工作并尝试将补全的省略数量名短语用于 CAMR 语义自动解析等工作中.

[参考文献]

[1] 黎锦熙. 论现代汉语中的量词[M]. 北京:商务印书馆,1978.

[2] 白晓革,李义杰. 数量短语的构成模式及其识别[C]//第三届 HNC 与语言学研究学术研讨会论文集,北京,2005:171-178.

[3] 张玲,熊文,李义杰,等. 基于知识库的现代汉语数量短语的识别[C]//第七届中文信息处理国际会议论文集,武汉,2007:295-299.

[4] 熊文,张玲. 一种基于规则不依赖于分词的中文数量短语的识别[C]//第七届中文信息处理国际会议论文集,武汉,2007:36-40.

[5] 方芳,李斌. 基于语料库的数量名短语识别[C]//第三届学生计算语言学研讨会论文集,沈阳,2006:331-337.

[6] PINHEIRO P H O,COLLOBERT R. Recurrent convolutional neural networks for scene parsing[EB/OL]. (2013-06-12)[2019-11-4]. //https://arxiv.org/abs/1306.2795.

[7] HUANG Z H,XU W,YU K. Bidirectional LSTM-CRF models for sequence tagging[EB/OL]. (2015-08-09)[2019-11-4]. https://arxiv.org/abs/1508.01991.

[8] CHIU J P C,NICHOLS E. Named entity recognition with bidirectional LSTM-CNNs[J]. Transaction of the association of computational linguistics,2016(4):357-370.

[9] 曲维光,周俊生,吴晓东,等. 自然语言句子抽象语义表示 AMR 研究综述[J]. 数据采集与处理,2017,32(1):26-36.

[10] 李斌,闻媛,宋丽,等. 融合概念对齐信息的中文 AMR 语料库的构建[J]. 中文信息学报,2017,31(6):93-102.

[11] PETERS M E,NEUMANN M,LYYER M,et al. Deep contextualized word representations[C]//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics:Human Language Technologies,New Orleans,Louisiana,United States of America. 2018:2227-2237.

[12] DEVLIN J,CHANG M W,LEE K,et al. BERT:Pre-training of deep bidirectional transformers for language understanding[C]//

Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019(1): 4171-4186.

[13] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems, 2017: 5998-6008.

[14] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural computation, 1997, 9(8): 1735-1780.

[15] ZHANG Y, YANG J. Chinese NER using lattice LSTM[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 2018: 1554-1564.

[16] LAFFERTY J D, MCCALLUM A, PEREIRA F C N. Conditional random fields: probabilistic models for segmenting and labeling sequence data[C]//Proceedings of the Eighteenth International Conference on Machine Learning, 2001: 282-289.

[17] RATINOV L, ROTH D. Design challenges and misconceptions in named entity recognition[C]//Proceedings of the Thirteenth Conference on Computational Natural Language Learning, 2009: 147-155.

[18] SRIVASTAVA N, HINTON G E, KRIZHEVSKY A, et al. Dropout: a simple way to prevent neural networks from overfitting[J]. Journay machine learning research, 2014, 15(1): 1929-1958.

[19] ZEILER M D. ADADELTA: an adaptive learning rate method[EB/OL]. (2012-12-22)[2019-11-4]. //https://arxiv.org/abs/1212.5701.

[20] KINGMA D P, BA J. Adam: a method for stochastic optimization[C]//3rd International Conference on Learning Representations, 2015: arXiv: 1412.6980.

[21] DAUPHIN Y N, VRIES H D, BENGIO Y. Equilibrated adaptive learning rates for non-convex optimization[C]//Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems, 2015: 1504-1512.

[责任编辑:顾晓天]