文章编号：2095-2163(2022)10-0057-06 中图分类号：TP391.1 文献标志码：A

基于特征融合的中文分词研究

张 倩，高建瓴，丁 容
（贵州大学 大数据与信息工程学院，贵阳 550025）

摘 要：中文分词是自然语言处理中一项重要的基础任务。由于中文词汇存在多义词、同音字等特殊性，能够准确地完成分词任务是近年来中文分词研究面临的挑战之一。因此，本文提出了一种融合字符特征、拼音特征、五笔输入特征的共享 BiLSTM-CRF 模型，通过在训练过程中共享 LSTM-网络来有效地融合语言特征。经大量数据集实验表明，特征融合能显著提高标记的准确性。在没有利用任何外部词汇资源的情况下，AS 和 CityU 数据集中准确率可分别达到 96.9%和 97.3%。

关键词：中文分词；拼音；五笔输入；BiLSTM-CRF

Research on Chinese words segmentation based on feature fusion

ZHANG Qian, GAO Jianling, DING Rong
(College of Big Data and Information Engineering, Guizhou University, Guiyang 550025, China)

【Abstract】 Chinese Word Segmentation (CWS) is an important basic task in Natural Language Processing (NLP). Due to the particularity of polysemy and homonym in Chinese vocabulary, it is one of the challenges faced by Chinese word segmentation research in recent years to complete the task of word segmentation accurately. Therefore, this paper proposes a shared BiLSTM-CRF model which integrates character features, Pinyin features and Wubi input features, and effectively integrates language features by sharing LSTM network in the training process. Experiments on a large number of data sets show that feature fusion can significantly improve the accuracy of labeling. Without using any external vocabulary resources, the accuracy of AS and CityU data sets can reach 96.9% and 97.3% respectively.

【Key words】 Chinese word segmentation; Pinyin; Wubi input; BiLSTM-CRF

0 引 言

随着自然语言处理[1]以及人工智能[2]的快速发展，已陆续涌现出各类超大量级的文本信息。在大量的文本信息中，中文分词（CWS）、通常是作为中文自然语言处理的第一步，则被认为是一项基于字符的序列标记任务，而分词结果的好坏将直接影响着后续的应用。因此，中文分词就成为许多领域研究中至关重要的一部分。通过提高中文分词的准确性，则使得人力投入的减少也已成为可能[3]。

中文分词的任务是将整个句子在不改变语义的前提下切分成一个个单词。不同于英语、德语、法语等语言，中文词与词之间没有明确的空格分隔符，加上汉字文化博大精深，经常会遇到一个词语代表多个含义、一个词语以多种形式出现的情况，这种特性很容易造成切分歧义，例如，输入的句子为“重庆市长江东路”，只有被切分为“重庆市”和“长江东路”时，切分正确，对应的检索结果才是准确的。如果错误切分为“苏州市长”和“江东路”，检索结果就会出现偏差。通过上述的例子可知，作为自然语言处理中底层任务之一的分词，对于准确率有着较高要求。

近些年来，分词研究已经得到各方关注，且已掀起研究热潮，目前就提出多种算法、旨在提高分割准确率，但是也还未见到能够媲美人工分割精准度的算法，故其研究仍具有重要的应用价值。例如，文献[4]提出了一种利用膨胀卷积神经网络 DCNN 来进行中文分词的方法，解决了现阶段一些模型存在的计算速度慢、输入特征不足等问题。文献[5]提出使用长短时记忆神经网络（Long Short-Term Memory，LSTM）学习中文分词的字符表示，使用 CRF（Conditional Random Field）联合解码标签的方法。文献[6]提出了一种改进 BiLSTM-CRF 网络的分词方法，解决了原分词模型在编码过程中的记忆压缩问题。文献[7]提出了改进卷积神经网络（Convolutional Neural Networks，CNN）的中文分词模型，克服了模型过于依赖人工处理特征的缺点，简化了模型结构，从而提高了分词准确率。文献[8]提出一种基于样本迁移学习的中文分词方法，增强了分词模型的领域自适应能

作者简介：张 倩（1998-），女，硕士研究生，主要研究方向：深度学习、自然语言处理；高建瓴（1969-），女，副教授，硕士生导师，主要研究方向：数据库系统、数据挖掘；丁 容（1998-），女，硕士研究生，主要研究方向：深度学习、自然语言处理。
通讯作者：高建瓴 Email：454965711@ qq.com
收稿日期：2022-05-07

力,然而这些分词方法往往忽略了中文的本质特征。自 Mikolov 等人[9]提出 Word2Vec 技术以来,单词或字符的向量表示已成为神经网络解决不同语言的 NLP 任务的先决条件。现有的中文分词方法忽略了一个重要事实,即汉字同时包含语义和语音含义,目前存在各种各样的汉字表示法用于捕捉这些特征。最直观的是通过使用拼音来表达汉字。从学习汉字到推广普通话、从文本输入到信息沟通、从教育普及到国际交流,汉语拼音早已渗透进日常生活中的各个领域。但汉字中也还存在着不少多义词、同音词,这在中文分词任务中既常见、又关键。除了拼音之外,五笔输入也是汉字语义表达的另一种有效表征。因为汉语中有着大量较为丰富的象形文字,且五笔在嵌入结构方面更加有效,因此与偏旁相比[10-11],五笔包括了更加系统、全面的图形和结构信息,且这些信息与语义以及词语边界高度相关。

基于此,为了提高中文分词的准确性,本文提出一种结合拼音特征、五笔特征、字符特征的共享 Bi-LSTM-CRF 模型,可以有效地融合多种嵌入,并可共享有用的语言特征,并且通过在 Bakeoff2005 和 CTB6 语料库上进行评估实验证明,特征融合有助于在没有外部词汇资源的情况下得到高准确率的中文分词结果。

1 多重嵌入

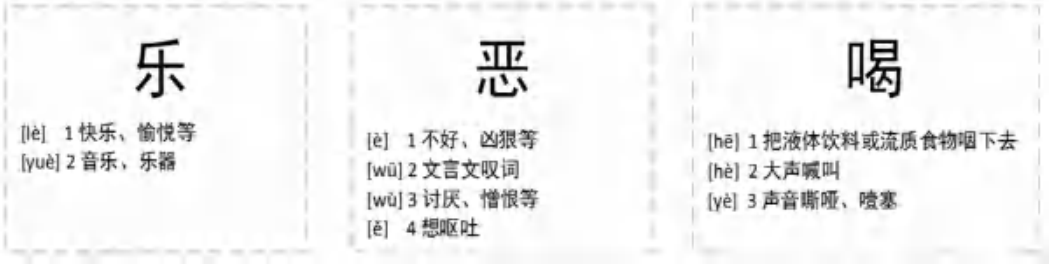
为了充分利用汉字的特征,本文将字符级嵌入分为 3 部分:文本特征的字符嵌入、语音特征的拼音嵌入和结构级特征的五笔输入嵌入。对此拟做研究阐释如下。

1.1 中文(汉字)特征

中文分词(CWS)通常被认为是一种基于字符的序列标记任务,主要作用是用{B,M,E,S}标记方案来标记每个字符。现有的大量研究表明,字符嵌入是神经网络最基本的输入[12]。然而,汉字往往包括语音、语义和象形文字三个方面,因此在本文中,将融合中文特征,以字符作为基本输入,并融入另外 2 种表示法、即拼音和五笔作为辅助。

1.2 拼音特征

汉语拼音是一种辅助汉字读音的工具,代表汉字的发音,其作用与英语中的音标无异。此外,拼音与语义有高度相关的联系,一个汉字可能存在不同的拼音、不同的语义,这种现象在中文文本中极为常见,被称为多音字、多义词。图 1 展示了多音字以及多义词的几个例子。例如,“乐”字在图 1(a)中有 2 种不同的发音。当发音为“lè”时,代表快乐、愉悦。然而,“yuè”的发音指的是音乐、乐器等意思。同样,“恶”字在图 1(b)中,甚至有 4 种含义,且有 4 种不同的拼音。通过拼音这一辅助工具,就能够在汉字和语义之间建立起直观联系。既然人类可以根据不同的发音来理解汉字的不同含义,那么神经网络也有可能自动学习语义和拼音之间的映射。显而易见的是,拼音可以提供中文分词所需的额外语音和语义信息,而且拼音是汉字计算机的主要输入方法,很容易用拼音作为补充输入来表示汉字。



(a) 示例一 (b) 示例二 (c) 示例三

图 1 多音词以及多义词示例

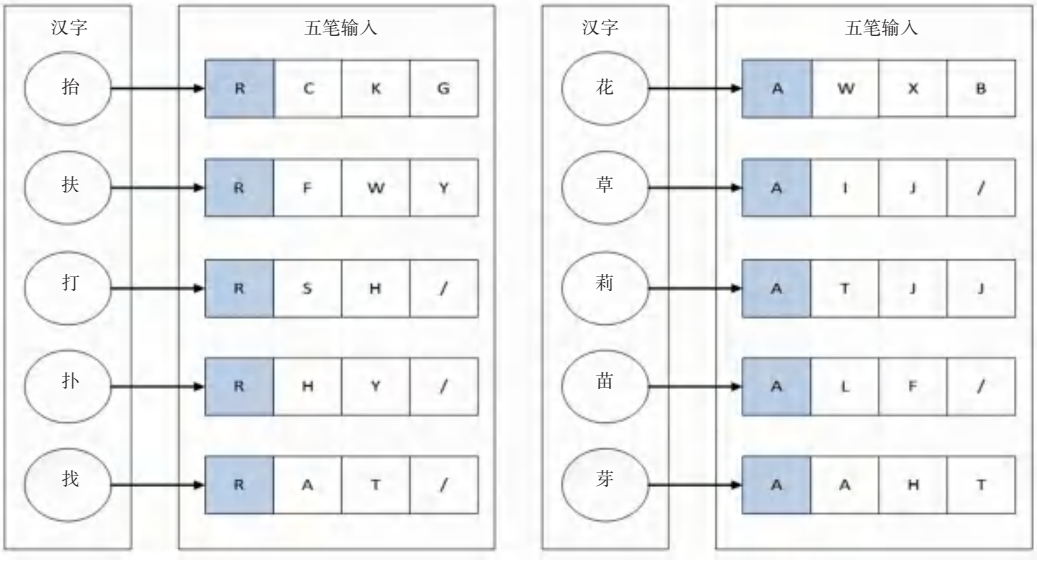
Fig. 1 Examples of polysyllabic words and polysemous words

1.3 五笔特征

五笔是五笔字型输入法的简称,五笔字型是按照汉字的笔画和字形特征来进行编码,属于典型的形码输入法。由于大量的汉字都是象形文字,因此使用五笔输入可以用来找出潜在的语义关系以及词语边界,且往往具有相似结构(例如部首)的汉字更有可能组成一个单词[13]。要了解其在结构描述中的有效性,必须遵循五笔输入法的规则。五笔是一种高效的编码系统,每个汉字最多使用 4 个英文字母来表示。具体来说,这些字母分为 5 个区域,每个区域代表一种笔划。

图 2 提供了一些汉字及其相对应的五笔码(4 个字母)的示例。例如图 2(a)中的“抬”、“扶”和“打”都是与手有关的动词,在中文文本中,这些汉

字都属于左右结构，并且具有相同的部首（在五笔码中为“R”）。也就是说，在语义上高度相关的汉字通常具有相似的结构，且这些结构可以被五笔完美地捕捉到。此外，结构相似的汉字一般更有可能组成一个单词。例如图 2（b）中的“花”、“草”和“苗”都是名词。而且也都是上下结构，并具有相同的部首（五笔码中的“A”）。而这些字通常可以组成词语，例如“花草”和“花苗”等。



（a）示例一　　（b）示例二

图 2　汉字的五笔输入

Fig. 2　Five-stroke input of Chinese characters

此外，五笔码中的顺序也是解释汉字关系的一种方法。在图 2 中，很容易找到一些有趣的编码规则。例如，研究后可以得出结论：

（1）五笔码的顺序意味着汉字结构的顺序，如：“IA”与“AI”和“IY”与“YI”。

（2）有些编码是具有实际意义的，如：“I”是指水。

因此，在本文提出的多特征模型中，五笔是一种有效的汉字编码，可以作为本文的特征来进行融合。

1.4　多特征嵌入

为了充分利用汉字的各种特征，本文引入了拼音嵌入和五笔嵌入作为 2 个补充字符级特征。本文首先按照 Lample 等人[14]的方法对字符进行预处理，获得基本字符嵌入。同时使用 Pypinyin 库注释拼音，并使用官方转换表将汉字转换为五笔编码。接着基于 Word2Vec 来得到拼音、五笔编码的向量映射。为了简易方便，该文将拼音编码和五笔码视为由 Word2Vec 处理得到的标准字符单元，这可能会丢弃掉一些语义相似性。但值得关注的是，考虑到中文结构是按字母顺序编码的，所以五笔编码中的序列顺序是一个有趣的特性（见下文第 2 节），对于这一点可以展开进一步研究。虽然本文生成拼音特征是依赖于外部资源，但五笔编码是在转换表下进行转换的，不会引入任何外部资源。

2　BiLSTM-CRF 模型

2.1　BiLSTM-CRF 模型概述

近年，各种各样的深度学习方法被广泛应用在中文分词任务中，通常采用 RNN 模型及其变体结构，在理论上来说，RNN 模型可以有效捕获远程上下文之间的关系，但其缺点是会因为梯度消失或是梯度爆炸而失效。因此，在实际应用中往往会选择 LSTM 来解决这类问题。在中文分词任务中，将会同时访问当下时刻的上下文，以便去预测当前的时刻，但是 LSTM 的隐藏状态 $h(t)$ 只会接受状态之前的信息。因此本文使用 Bi-LSTM 模型来获得每一个状态的上下文，且从左到右和从右到左使用 2 个相对独立的隐藏状态，用来同时捕获过去和未来的信息[14]。而用 CRF 模型实现中文分词，则可在具有良好学习性能的同时从一定程度上实现对生词的识别。

基于此，本文采用 BiLSTM-CRF 作为本次实验的基线模型，该基线模型不包含拼音嵌入和五笔嵌入，类似于 Lample 等人[14]提出的架构。为了获得多个功能的有效融合和共享机制，本文设计了一种基于特征融合的共享 Bi-LSTMs-CRF 模型架构如图 3 所示。接下来，本文将给出详细的解释和分析。

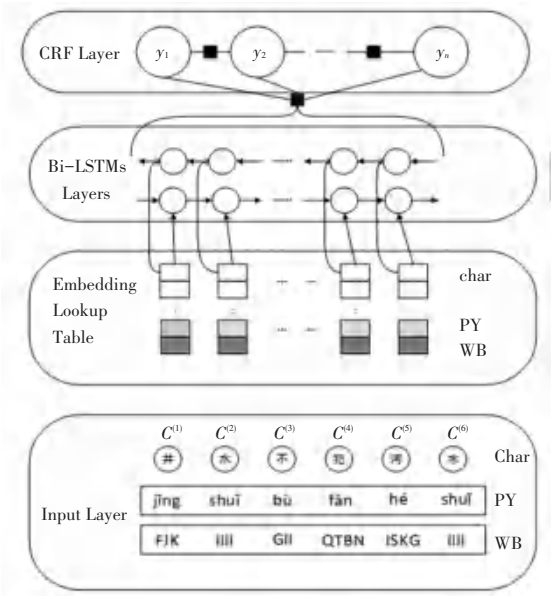


图 3 基于特征融合的 BiLSTM-CRF 网络结构

Fig. 3 BiLSTM-CRF network structure based on feature fusion

2.2 Shared Bi-LSTMs-CRF 模型

在该模型中,本文将字符、拼音和五笔嵌入顺序输入到具有相同参数的堆叠 Bi-LSTM 网络中,若将 3 个 LSTM 网络设置为互为独立的参数,这将忽略不同嵌入之间的相互作用,而且在训练期间会导致计算成本有较大增加。因此为了在保持训练效率的同时解决特征依赖性的问题,本文模型引入了一种共享机制,如图 4 所示。该模型不是让拼音和五笔使用不同的 Bi-LSTM 网络,而是通过字符嵌入来共享相同的 LSTM。研究中推出的数学公式如下:

$$\begin{bmatrix} h_{3,C}^{(t)} \\ h_{3,P}^{(t)} \\ h_{3,W}^{(t)} \end{bmatrix} = Bi-LSTMs\left(\begin{bmatrix} w_C^{(t)} \\ w_P^{(t)} \\ w_w^{(t)} \end{bmatrix}, \theta\right) \tag{1}$$

$$h^{(t)} = h_{3,C}^{(t)} + h_{3,P}^{(t)} + h_{3,W}^{(t)} \tag{2}$$

其中, θ 表示 Bi-LSTM 的共享参数,3 层 Bi-LSTM 的输出为 $h_{3,C}^{(t)}$、$h_{3,P}^{(t)}$ 和 $h_{3,W}^{(t)}$, 共同构成 CRF 层 h^t 的输入; $w_C^{(t)}$、$w_P^{(t)}$、$w_w^{(t)}$ 为其连通层的可训练参数。具体来说,在每个历元,3 个网络的参数都是基于统一的序列字符、拼音和五笔嵌入进行更新的。第二个 LSTM 网络将与第一个网络共享(或同步)参数,而后开始以拼音作为输入的训练过程。这样,第二个网络将在前一个相关嵌入的基础上,在细化参数方面仅需花费更少的时间。第三个网络也是如此(以五笔嵌入为输入)。

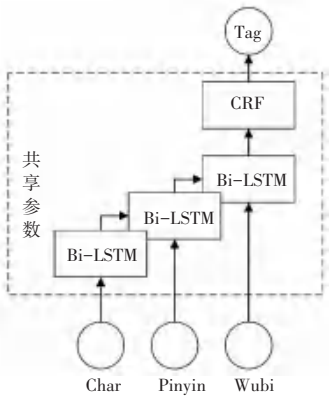


图 4 Shared Bi-LSTMs-CRF 模型

Fig. 4 Shared Bi-LSTMs-CRF model

3 实验评估

本节中,通过实验结果验证了多特征融合对中文分词的有效性。对此研究可知,本文提出模型 Shared Bi-LSTMs-CRF 可对数据集进行有效训练,虽然成本略高于基线模型,但能得到更高的准确率。

3.1 实验设置

为了使实验结果更具有可比性和说服力,本文在 SIGHAN 2005 (Emerson, 2005) 和中国树库 6.0 (CTB6)数据集上评估了本文模型,并将利用标准的 Word2Vec 工具来训练拼音、五笔等多重嵌入。实验中,本文根据 Yao 等人[15] 调整了嵌入大小,设置 *batch size* = 256, 并将 Bi-LSTM 层的数量设置为 3。

3.2 实验结果

本文全面分析了研究提出的模型架构。模型在 5 个数据集上的分词性能见表 1。由表 1 可知,与仅以字符嵌入为输入的基线模型相比,多特征融合模型得到了相当大的改进。且本文提出的 Shared Bi-LSTMs-CRF 即使在可训练参数较少的情况下也能获得更好的性能。

表 1 模型在 5 个数据集上的分词性能

Tab. 1 Model's word segmentation performance on five data sets

模型	CTB6			PKU			MSR			AS			CityU		
	P	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*
基线模型	94.1	94.0	94.2	95.9	95.7	95.8	95.2	95.3	95.5	95.5	95.4	95.5	95.8	95.7	96.0
Shared Bi-LSTMs-CRF	95.4	95.0	95.3	96.3	96.2	96.1	97.1	96.9	97.1	96.8	96.9	96.8	97.0	97.1	97.1

3.3 对比实验

为了证明融合拼音特征和五笔输入特征对中文分词的有效性，将该文提出的方法与近年其他的先进分词方法在 Bakeoff 2005 语料库上进行了比较，其结果见表 2。

表 2　与其他文献的模型在 Bakeoff 2005 四个数据集上的分词性能比较

Tab. 2　Comparison of word segmentation performance with other literature models on four data sets of Bakeoff 2005

模型	PKU	MSR	AS	CityU
文献[12]模型	**96.5**	97.4	/	/
文献[15]模型	96.1	97.4	96.2	97.2
文献[17]模型	96.0	97.9	96.1	96.9
文献[19]模型	94.4	**98.1**	96.4	96.9
Baseline	95.7	95.6	95.6	96.1
Baseline+PY	96.1	96.7	96.8	97.1
Baseline+WB	96.3	97.2	96.7	**97.4**
Baseline+PY+WB	96.2	97.1	**97.0**	97.2

由表 2 可以看出，该文提出的基于多特征融合的 Bi-LSTMs-CRF 模型的在 AS 和 CityU 数据集上 F 值分别达到了 97.0% 和 97.4%，且在没有利用外部资源的情况下（例如预先培训的字符或单词嵌入、额外的字典、有标签或无标签的语料库），也能在 PKU 和 MSR 数据集上取得具有竞争力的成绩。此外值得注意的是，经研究发现，AS 和 CityU 数据集的容量更大、词汇量更高，研究人员认为在这 2 个数据集上进行实验更具有真实性，这也再次验证了本文所提出的拼音和五笔嵌入能够降低大规模数据中的误切分率，且能够得到更准确的分词结果。

3.4 嵌入消融

本文通过在 CTB6 和 CityU 上进行嵌入消融实验来分别验证拼音和五笔嵌入的有效性。CTB6 和 CityU 上的嵌入消融见表 3。表 3 中，+PY 和+WB 表示在基线模型下分别注入拼音和五笔嵌入。与普通的单字符嵌入模型、即基线模型相比，嵌入拼音特征和五笔特征在 F_1 分数上能得到显著提高。此外，与嵌入拼音特征的模型相比，融入五笔特征得到的效果更为显著。

表 3　CTB6 和 CityU 上的嵌入消融

Tab. 3　Embedded ablation on CTB6 and CityU

模型	CTB6			CityU		
	P	*R*	*F*	*P*	*R*	*F*
基线模型	94.1	94.0	94.2	95.8	95.7	96.0
基线+PY	94.6	94.9	94.8	96.7	96.5	96.5
基线+WB	95.3	95.4	95.3	97.2	97.3	97.2

4　结束语

本文提出一种融合字符特征、拼音特征、五笔输入特征的中文分词模型 Shared Bi-LSTMs-CRF，该模型通过利用汉字的语音、结构和语义特征来达到提高分词准确性的目的。本文通过对比实验以及消融实验来验证了拼音和五笔嵌入在中文分词任务中起到了极大的作用，此外，本文提出了一个 Shared Bi-LSTMs-CRF 模型来融合多重嵌入，并在 5 个公共语料库中验证其可行性。经多个实验结果得到：本文所提出的共享的 Shared Bi-LSTMs-CRF 模型可以有效地进行训练，并在 AS 和 CityU 这 2 个语料库上能够得到最高准确率，而且证明了融入拼音嵌入以及五笔输入嵌入可以提高模型的性能。在未来，希望可以将中文语言特征继续应用于其他 NLP 任务（如实体识别和中文分类等）中。

参考文献

[1] 陈肇雄，高庆狮. 自然语言处理[J]. 计算机研究与发展，1989(11)：3-18.

[2] 林尧瑞，马少平. 人工智能导论[M]. 北京：清华大学出版社，1989.

[3] HEARST M A. Automatic acquisition of hyponyms from large text corpora [C]// Proceedings of the 14th Conference on Computational Linguistics (CoLing). Stroudsburg, PA, USA: ACL, 1992: 539-545.

[4] 王星，李超，陈吉. 基于膨胀卷积神经网络模型的中文分词方法[J]. 中文信息学报，2019，33(09)：24-30.

[5] PENG Nanyun, DREDZE M. Multi-task domain adaptation for sequence tagging[C]//Proceedings of the Second Workshop on Representation Learning for NLP. Vancouver, Canada: dblp, 2017: 91-100.

[6] 金宸. 基于改进的双向 LSTM-CRF 中文分词模型[D]. 昆明：云南大学，2018.

[7] 涂文博，袁贞明，俞凯. 无池化层卷积神经网络的中文分词方法[J]. 计算机工程与应用，2020，56(02)：120-126.

[8] 张艳娜. 基于样本迁移学习的中文分词领域自适应方法的研究[D]. 北京：北京交通大学，2019.

[9] MIKOLOV T, SUTSKEVER I, CHEN Kai, et al. Distributed representations of words and phrases and their compositionality[J]. CoRR, abs/1310.4546, 2013.

[10] SUN Yaming, LIN Lei, YANG Nan, et al. Radical-enhanced chinese character embedding[C]// International Conference on Neural Information Processing (ICONIP). Kuching, Malaysia: Springer, 2014: 279-286.

[11] SHAO Yan, HARDMEIER C, TIEDEMANN J, et al. Character-based joint segmentation and POS tagging for Chinese using bidirectional RNN-CRF[C]//International Joint Conference on Natural Language Processing(IJCNLP). Taiwan: dblp, 2017: 173-183.

[12] CHEN Xinchi, QIU Xipeng, ZHU Chenxi, et al. Long short-term memory neural networks for chinese word segmentation [C]//Conference on Empirical Methods in Natural Language Processing(EMNLP). Lisbon: ACL, 2015: 1197-1206.

(下转第 67 页)