基于 LSTM 网络的序列标注中文分词法*

任智慧[1,2]，徐浩煜[2,3†]，封松林[2,3]，周 晗[2]，施 俊[1]
(1. 上海大学 通信与信息工程学院，上海 200444；2. 中国科学院上海高等研究院，上海 201210；3. 中国科学院大学，北京 100049)

摘 要： 当前主流的中文分词方法是基于字标注的传统机器学习方法，但传统机器学习方法需要人为地从中文文本中配置并提取特征，存在词库维度高且利用 CPU 训练模型时间长的缺点。针对以上问题进行了研究，提出基于 LSTM(long short-term memory)网络模型的改进方法，采用不同词位标注集并加入预先训练的字嵌入向量(character embedding)进行中文分词。在中文分词评测常用的语料上进行实验对比结果表明，基于 LSTM 网络模型的方法能得到比当前传统机器学习方法更好的性能；采用六词位标注并加入预先训练的字嵌入向量能够取得相对最好的分词性能；而且利用 GPU 可以大大缩短深度神经网络模型的训练时间；LSTM 网络模型的方法也更容易推广并应用到其他自然语言处理中序列标注的任务。

关键词： 中文分词；LSTM；字嵌入；自然语言处理
中图分类号： TP391.1 **文献标志码：** A **文章编号：** 1001-3695(2017)05-1321-04
doi:10.3969/j.issn.1001-3695.2017.05.009

Sequence labeling Chinese word segmentation method based on LSTM networks

Ren Zhihui[1,2], Xu Haoyu[2,3†], Feng Songlin[2,3], Zhou Han[2], Shi Jun[1]
(1. *School of Communication & Information Engineering, Shanghai University, Shanghai* 200444, *China*; 2. *Shanghai Advanced Research Institute, Chinese Academy of Sciences, Shanghai* 201210, *China*; 3. *University of Chinese Academy of Sciences, Beijing* 100049, *China*)

Abstract: Currently, the dominant state-of-the-art methods for Chinese word segmentation are based on character tagging methods by using traditional machine learning technology. However, there are some disadvantages in the traditional machine learning methods: artificially configuring and extracting features from Chinese texts, high dimension of the dictionary, long training time by just exploiting CPUs. This paper proposed an improved method based on long short-term memory (LSTM) network model. It used different tag set and added pre-trained character embeddings to perform Chinese word segmentation. Compared with the best result in Bakeoff and state-of-the-art methods, this paper conducted the experiments on common used corpuses. The results demonstrate that traditional machine learning methods are exceeded by the methods based on LSTM network. By using six-tag-set and adding pre-trained character embedding, the proposed method can reach the relatively highest performance on Chinese word segmentation. Then, it can greatly reduce the training time of deep neural network model by using GPUs. Moreover, the methods based on LSTM network can easily applied to other sequence labeling tasks in natural language processing(NLP).

Key words: Chinese word segmentation; LSTM; character embedding; NLP

0 引言

在基于中文的自然语言处理中，由于不像英文中词与词之间有固定的自然分界符，所以对中文进行分词通常是处理中文信息任务的基础；同时由于中文自身的复杂性，中文分词也一直是中文信息处理任务的难题。中文分词是进行词性标注、命名实体识别、关键词提取和文本聚类等后续自然语言处理任务的基础，也是语义分析等深层次文本理解任务的基础，在中文搜索引擎、输入法、机器翻译和智能问答系统等应用中起到了支撑作用。中文分词作为中文自然语言处理领域的重要基础研究，近些年来很多专家学者致力于该领域的研究，研究方法主要分为三种：a) 基于规则的方法；b) 基于传统机器学习模型的方法；c) 基于深度神经网络模型的方法。

基于规则的方法利用构词原理结合标注的词性等信息，构建基于句法—语义规则的分析系统，配合语法信息字典，并补充了大量消除歧义的信息。文献[1,2]均利用语法规则，其中文献[2]在语法规则的基础上增加了领域特征进行中文分词的处理。基于规则的方法的优点是具有针对性和暂时较高的准确率，但由于句法构造的领域相关性，适应性较差，词典与歧义消解处理难维护。

随着 SIGHAN 国际中文分词评测 Bakeoff 的展开，将中文分词任务视做序列标注问题来解决逐渐成为主流。基于传统机器学习模型的方法主要为基于字标注的机器学习模型方法，即字在字串的标注问题，该方法能平等地看待词典词和未登录

收稿日期：2016-03-25；修回日期：2016-05-25 基金项目：国家自然科学基金资助项目(61471231)；中国科学院先导资助项目(XDA06010301)

作者简介：任智慧(1992-)，男，安徽芜湖人，硕士研究生，主要研究方向为自然语言处理、数据分析等；徐浩煜(1978-)，男(通信作者)，上海人，研究员，主要研究方向为大数据挖掘、自然语言处理等(xuhy@sari.ac.cn)；封松林(1964-)，研究员，主要研究方向为传感技术；周晗(1986-)，男，浙江温州人，助理研究员，主要研究方向为数据分析、自然语言处理等；施俊(1977-)，男，江苏人，教授，主要研究方向为机器学习、生物医学信号等.

词的识别。在 Bakeoff 展开的初期，基于字标注的中文分词方法广泛应用，在评测中取得性能领先的系统均应用了此类思想[3,4]。Xue 等人[5,6]采用基于最大熵模型的四词位标注集(B,M,E,S)的方法进行中文分词。Peng 等人[7]提出使用链式条件随机场(conditional random fields，CRF)模型应用于中文分词。于江德等人[8,9]使用四词位标注集基于链式 CRF 模型，选择多种特征模板在 Bakeoff 语料上进行实验，研究得出 TMPT-10′的特征模板取得了更好的结果。在文献[10,11]中，采用链式 CRF 模型，使用六词位标注集(B,B_2,B_3,M,E,S)和 TMPT-6(六特征模板)实现的分词系统取得了很好的分词效果。在此基础上，文献[12～14]提出了基于子串标注方法；徐浩煜等人[15]提出了基于链式 CRFs 模型的改进方法，在分词性能相当的情况下能得到更好的未登录词召回率。以上基于传统机器学习模型的方法的性能受限于特征的选择和提取，模型的训练是基于提取出的人为设定的特征。

正是由于尽可能避免特征工程的影响，深度学习网络模型逐渐应用到中文分词等自然语言处理任务中。Zheng 等人[16]首先将深层神经网络[17,18]应用到中文分词任务，同时还提出了一种感知器算法，在几乎不损失性能的前提下加速训练过程。Pei 等人[19]在文献[16]的基础上，通过利用标签嵌入和基于张量的转换提出了 MMTNN 模型的方法用于中文分词任务。Chen 等人[20]将基于 recurrent neural network(RNN)模型改进了记忆单元的 LSTM 方法用于中文分词任务，并取得了与当前传统机器学习方法相当的分词性能；在文献[21]中利用改进的 gated recursive neural network 模型进行中文分词。当前深度神经网络模型的方法均是基于四词位标注的方式，未能充分表达字在词语中的词位信息，且缺少预训练的字嵌入分布式向量。

为了解决以上问题，本文提出基于 LSTM 神经网络，采用六词位标注，并加入预先训练的字嵌入分布式向量的方法。由于 LSTM 神经网络模型可以有效地保持较长时间的记忆，可以充分利用文本中有用的远距离的信息特性，与中文分词的任务非常契合。本文在中文分词评测经常使用的评测语料(MSRA，PKU 和 CTB6 corpus)上，通过实验研究表明：a) 采用六词位标注的方法在分词性能上优于四词位标注的方法；b) 在采用同一种词位标注的方法中，加入了预先训练的字嵌入向量，能得到更好的分词性能；c) 采用六词位标注并加入预先训练的字嵌入向量，能得到相对最好的分词结果。采用 LSTM 模型的方法避免了传统机器学习模型中的特征工程，能取得与当前基于传统机器学习模型方法相当的性能，甚至更好；LSTM 等深度学习模型可借助 GPU 来训练，大大缩短了训练时间；相比传统机器学习模型为特定 NLP 任务准备的特征工程，LSTM 网络模型的方法也更容易推广应用到其他 NLP 中序列标注的任务，具有一定的通用性。

1 基于 LSTM 模型的分词方法

中文分词可视为字级别的序列标注问题，将分词过程转换为每个字在文本序列中标注的过程[11]。在中文文本中，由于一个词语中每个字都占一个确定的词位，所以可以将分词过程视为学习这个字的词位信息的机器学习的过程。

标注集，即每个字词位信息熵预定义的标注。已有的研究工作中常用的标注集有三种：二词位、四词位和六词位标注集，各词位标注集的定义参考表 1。四词位的标注集主要用于最大熵模型的分词系统[6]；二词位的标注集大多用于早期基于字标注的 CRF 分词系统[22]，黄昌宁[10]在 Bakeoff-2 中首次使用了六词位的标注集；已有的深度神经网络模型的方法均采用四词位标注集。

相比于四词位和二词位标注集，六词位标注集更能有效地表现字在词语中的词位信息，表达能力更强。

表 1 三类词位标注集的定义

标注集	标 记	定义
二词位	S, N	开始，后续
四词位	B, M, E, S	开始，中间，结束，单字
六词位	B, B_2, B_3, M, E, S	开始，第二字符，第三字符，中间，结束，单字

基于深度神经网络的中文分词通用模块主要由三部分组成[16]，即文本向量化、一系列典型的神经网络层、标签推理层。通用框架如图 1 所示。

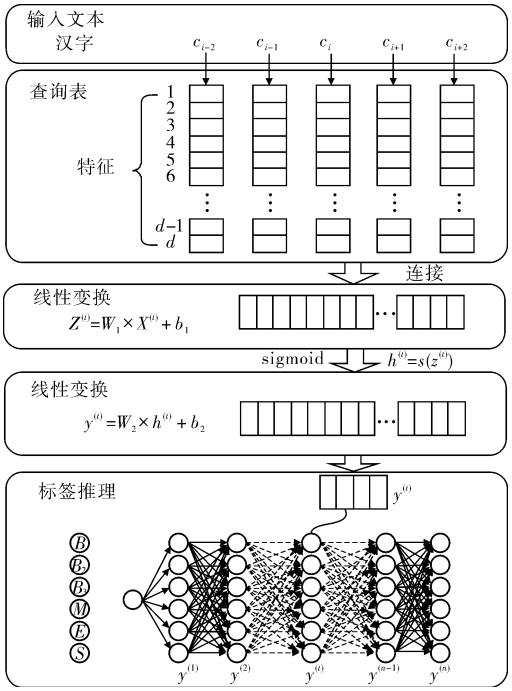

图 1 基于深度神经网络中文分词通用框架

基于字标注的分词方法是基于一个局部滑动窗口，假设一个字的标签极大地依赖于其相邻位置的字。给定长度为 n 的文本序列 $c^{(1:n)}$，大小为 k 的窗口从文本序列的第一个字 $c^{(1)}$ 滑动至最后一个字 $c^{(n)}$。如图 1 所示，对于序列中每个字 $c^{(t)}$，当窗口大小为 5 时，上下文信息($c^{(t-2)}$，$c^{(t-1)}$，$c^{(t)}$，$c^{(t+1)}$，$c^{(t+2)}$)将被送入查询表中，当字的范围超过了序列边界时，将以诸如“start”和“end”等特殊标记来补充；然后，将查询表中提取的字向量连接成一个向量 $X^{(t)}$。在神经网络下一层中，$X^{(t)}$ 经过线性变换后经由 Sigmoid 函数 $\sigma(x)=(1+E^{-x})^{-1}$ 或 tanh 函数激活。

$$h^{(t)}=\sigma(W_1X^{(t)}+b_1) \tag{1}$$

根据给定的标注集，将经过一个相似的线性变换，不同之处在于没有非线性函数，得到的 $y(t)$ 是每个可能标签的得分向量。本文选定的是更能充分表达词位信息的六词位标注集 $\{B,B_2,B_3,M,E,S\}$。

$$y^{(t)}=W_2h^{(t)}+b_2 \tag{2}$$

为了建模标签间依赖，引入了转移得分向量 A_{ij}，用于衡量从标签 i 跳转到标签 j 的概率。以往的研究表明，引入转移得

分向量非常适用于中文分词等序列标注的任务，但它仅利用了长度有限的窗口信息。下面将主要从文本向量化、LSTM 网络和标签推理等方面来介绍基于 LSTM 神经网络的中文分词方法。

1.1　文本向量化

使用诸如 LSTM 等神经网络来处理数据，需要先将文本向量化。文本向量化的方式主要有两种：集中式表示（one-hot representation）和分布式表示（distributed representation）[18]。集中式表示将每个词表示为一个很长的向量，向量的维度是词表大小，通常非常稀疏，且任意两个字间不存在联系，均是孤立的；分布式表示的方式以低维度的向量来表示，让相关的字在语义上更加接近，通常又称为 embedding（嵌入）。

以往的研究表明，加入预先训练的字嵌入向量可以提升自然语言处理任务的性能，本文在后面进行的部分实验中加入了预先训练的字嵌入向量。

1.2　LSTM 网络

RNN 具有循环的网络结构，具备保持信息的能力，其网络结构如图 2 所示。RNN 中的循环网络模块将信息从网络的上一层传输到下一层，网络模块的隐含层每个时刻的输出都依赖于以往时刻的信息。RNN 的链式属性表明其与序列标注问题存在着紧密的联系，目前已被应用到文本分类和机器翻译等 NLP 任务中。在 RNN 的训练中，存在梯度爆炸和消失的问题，且 RNN 难以保持较长时间的记忆。

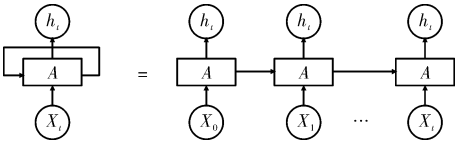

图 2　RNNs 网络结构

LSTM 网络是 RNN 的扩展，专门设计用来避免长期依赖问题。LSTM 的重复神经网络模块具有不同的结构，这与朴素 RNN 不同，存在四个以特殊方式相互影响的神经网络层，网络模块示意图如图 3 所示。LSTM 网络的关键在于细胞状态，有点类似于传送带。在 LSTM 中，通过门（gates）结构来对细胞状态增加或删除信息，而门结构是选择性地让信息通过的方式，通常由一个 Sigmoid 神经网络层和逐点乘积操作组成（Sigmoid 层的输出在 0 ~ 1，定义了信息通过的程度，0 表示什么都不让过，1 表示所有都让过）。

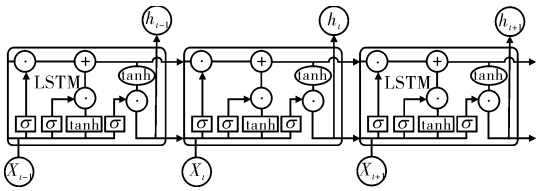

图 3　LSTM 神经网络结构

LSTM 网络具有输入门（input gates）、忘记门（forget gates）和输出门（output gates）等三种门结构，用以保持和更新细胞状态，以下公式中 i_t、f_t、o_t 和 C_t 分别表示 t 时刻对应的三种门结构和细胞状态。

a）从细胞状态中忘记信息，由忘记门的 Sigmoid 层决定，以当前层的输入 x_t 和上一层的输出 h_{t-1} 作为输入，在 $t-1$ 时刻的细胞状态输出为

$$f_t = \sigma(W_f \cdot (h_{t-1}, x_t) + b_f) \tag{3}$$

b）在细胞状态中存储信息，主要由两部分组成：（a）输入门的 Sigmoid 层的结果 i_t 作为将更新的信息；（b）由 tanh 层新创建的向量 C_t，将添加在细胞状态中。将旧的细胞状态 C_{t-1} 乘以 f_t，用以遗忘信息，与新的候选信息 $i_t \cdot \tilde{C}_t$ 的和，生成细胞状态的更新。

$$i_t = \sigma(W_i \cdot (h_{t-1}, x_t) + b_i) \tag{4}$$

$$\tilde{C}_t = \tanh(W_C \cdot (h_{t-1}, x_t) + b_C) \tag{5}$$

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t \tag{6}$$

c）输出信息，由输出门决定。先使用 Sigmoid 层来决定要输出细胞状态的部分信息，接着用 tanh 处理细胞状态，两部分信息的乘积得到输出的值。

$$o_t = \sigma(W_o \cdot (h \cdot, x \cdot) + b_o) \tag{7}$$

$$h_t = o_t \cdot \tanh(C_t) \tag{8}$$

LSTM 网络模型已被成功地应用到诸如文本/情感分类[23,24]、机器翻译[25]、智能问答[26,27]和看图说话[28]等自然语言处理任务中。由于 LSTM 网络通过记忆单元去学习从细胞状态中忘记信息、去更新细胞状态的信息，而且具有学习文本序列中远距离依赖的特性，很自然地想到可以使用 LSTM 网络模型进行中文分词的任务。

在中文分词任务中，LSTM 记忆单元的输入来自上下文窗口的汉字。对每个汉字 $c^{(t)}$，LSTM 记忆单元的输入为 $X^{(t)}$，由上下文字嵌入（$c^{(t-k)}$，…，$c^{(t)}$，…，$c^{(t+k)}$）连接而成，其中 k 代表与当前字的距离。LSTM 单元的输出在经过线性变换后用于标签推理函数，推理出汉字对应的标签。

$$X^{(t)} = V_c^{(t-k)} \oplus \cdots \oplus V_c^{(t+k)} \tag{9}$$

1.3　标签推理

为了建模标签间依赖，在以往的神经网络模型方法中引入了转移得分向量 A_{ij}，用于衡量从标签 i 跳转到标签 j 的概率。对于输入文本序列 $c^{(1:n)}$，其标注的标签序列为 $y^{(1:n)}$，序列级的得分是标签转移得分和网络标注得分的总和。

$$s(c^{(1:n)}, y^{(1:n)}) = \sum_{t=1}^{n} (A_{y(t-1)y(t)} + y_{y(t)}^{(t)}) \tag{10}$$

2　实验

2.1　实验环境、数据集和评测指标

本文实验的环境为 PowerLeader PR1792GH 服务器，主要参数 CPU：2 × Intel Xeon CPU E5-2620 v2 @ 2.10 GHz，GPU 卡：2 × Nvidia Tesla K20，内存为 96 GB，操作系统为 Ubuntu 14.04 64 bit。使用 Word2Vec[29] 进行 character embedding 的训练，MXNet[30] 提供的 LSTM 等神经网络模型。

本文实验采用了当前学术论文中经常使用到的训练语料和测试语料，分别是 MSRA corpus、PKU corpus 和 CTB6 corpus，其中 MSRA 和 PKU corpus 是由 SIGHAN 举办的第二届国际中文分词评测 Bakeoff 所提供的封闭语料，而 CTB6 corpus 是由 Linguistic Data consortium 提供的 Chinese TreeBank 6.0 数据集（LDC2007T36）。

在中文分词性能评估中，采用了分词评测常用的 R（召回率）、P（准确率）和 F（综合指标 F 值）等评测指标，以 F 值为主要评测指标。

2.2　实验设计及结果分析

本文设计了三组实验，分别为基于六词位标注的 LSTM 网络模型的方法，以及预训练好字嵌入分别基于四词位和六词位标注的方法，实验项分别记为 LSTM（6tag）、LSTM（4tag + ce）和

LSTM(6tag + ce) 。通过不断地调整超参数训练网络模型，直至在封闭测试集上得到较好的实验结果，与文献 [20] 的结果(记为 FDU-LSTM(4tag)) 进行比较，MSRA 和 PKU corpus 的结果还将与当年度 Bakeoff 评测最好的结果(记为 Bakeoff-Best) [3] ，基于链式条件随机场模型，采用六词位标注集结合 TMPT-10′特征模板的传统机器学习方法[15] (记为 6tag-tmpt10′) 进行对比，实验结果如表 2 所示。

表 2 在 MSRA、PKU 和 CTB6 语料上结果对比

Models	MSRA corpus			PKU corpus			CTB6 corpus		
	P	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*
Bakeoff-Best	0.966	0.962	0.964	0.946	0.953	0.95	–	–	–
6tag-tmpt10′	0.972	0.97	0.97	0.937	0.928	0.932	–	–	–
FDU-LSTM(4tag)	0.967	0.962	0.964	0.958	0.955	0.957	0.95	0.948	0.949
LSTM(6tag)	0.968	0.964	0.966	0.961	0.958	0.959	0.955	0.952	0.954
LSTM(4tag + ce)	0.972	0.97	0.971	0.963	0.962	0.963	0.96	0.956	0.958
LSTM(6tag + ce)	0.978	0.976	0.977	0.968	0.967	0.967	0.966	0.962	0.964

观察表 2 的结果可知，在 MSRA 和 PKU corpus 上，本文设计的深度学习网络模型在各评测指标都优于当年度 Bakeoff 评测中的最好结果，基于链式条件随机场模型改进的方法在 MSRA corpus 上各性能指标也优于 Bakeoff 评测中的最好结果。在选择的三类数据集上，与文献 [20] 基于四词位标注的 LSTM 网络模型的方法相比，在保持网络模型结构不变的情况下，基于六词位标注的 LSTM 网络模型的方法能取得更好的分词性能; 在同样保持网络模型和标注方式的情况下，加入了预先训练的字嵌入向量，同样能得到更好的分词性能; 在保持网络模型结果不变的情况下，基于六词位标注且加入预先训练好的字嵌入，能取得最优的分词性能，在 MSRA corpus 上准确率提升了 1.13%，*F* 值提升了 1.35%，而在 PKU corpus 上准确率和 *F* 值均提升了 1.04%，在 CTB6 corpus 上准确率提升 1.68%，*F* 值提升了 1.58%。

本次实验的 LSTM 网络模型中，不论有没有加入预先训练的字嵌入向量，采用六词位标注的方法在分词性能上均优于四词位标注的方法; 在采用同一种词位标注的方法中，加入了预先训练的字嵌入，能得到更好的分词性能。通常情况下，采用六词位标注并加入预先训练的字嵌入向量，能得到相对最好的分词结果。

采用 LSTM 模型的方法能取得与当前基于传统机器学习模型方法相当的性能，甚至更好。相比于传统机器学习模型的方法，基于 LSTM 模型的方法，无须人为地去提取中文文本中蕴涵的特征，而且传统机器学习方法的文本特征的配置和提取困难; 降低了维度，传统机器学习方法词库的维度可达几十万; 缩短了分词模型的训练时间，同样 CPU 和内存配置的情况下，LSTM 等深度学习模型可借助 GPU 来训练，大大缩短了训练时间。另外，LSTM 网络模型的方法也更容易推广应用到其他 NLP 中序列标注的任务。

3 结束语

本文的工作基于 LSTM 网络模型，采用六词位标注并加入预先训练的字嵌入向量的方法进行中文分词，并在常用语料上进行实验，与 Bakeoff 评测当年度最好结果、典型传统机器学习模型的方法和 FDU LSTM 的方法进行对比，结果表明: 不论有没有加入预先训练的字嵌入向量，采用六词位标注的方法在分词性能上优于四词位标注的方法; 在采用同一种词位标注的方法中，加入了预先训练的字嵌入向量，能得到更好的分词性能; 且通常情况下，采用六词位标注并加入预先训练的字嵌入，能得到相对最好的分词结果。而且，采用 LSTM 模型的方法避免了传统机器学习模型中的特征工程，能取得与当前基于传统机器学习模型方法相当的性能，甚至更好; 利用 GPU 可以大大缩短深度神经网络模型的训练时间; 基于 LSTM 网络模型的方法也更容易推广应用到其他 NLP 中序列标注的任务，具有一定的通用性。

参考文献:

[1] Wu Andi，Jiang Z. Word segmentation in sentence analysis [C] //Proc of International Conference on Chinese Information Processing. 1998: 169-180.

[2] Sui Z，Chen Y. The research on the automatic term extraction in the domain of information science and technology [C] //Proc of the 5th East Asia Forum of the Terminology. 2002.

[3] Emerson T. The second international Chinese word segmentation bakeoff [EB/OL]. [2010-06-14]. http://www.aclweb.org/anthology/105-3017.pdf.

[4] Levow G A. The third international Chinese language processing bakeoff: word segmentation and named entity recognition [C] //Proc of the 5th SIGHAN Workshop on Chinese Language Processing. 2006: 108-117.

[5] Xue Nianwen，Converse S P. Combining classifiers for Chinese word segmentation [C] //Proc of the 1st SIGHAN Workshop on Chinese Language Processing. Stroudsburg: Association for Computational Linguistics，2002: 1-7.

[6] Xue Nianwen，Shen Libin. Chinese word segmentation as LMR tagging [C] //Proc of the 2nd SIGHAN Workshop on Chinese Language Processing. Stroudsburg: Association for Computational Linguistics，2003: 176-179.

[7] Peng Fuchun，Feng Fangfang，McCallum A. Chinese segmentation and new word detection using conditional random fields [C] //Proc of the 20th International Conference on Computational Linguistics. Stroudsburg: Association for Computational Linguistics，2004.

[8] 于江德，睢丹，樊孝忠. 基于字的词位标注汉语分词 [J]. 山东大学学报: 工学版，2010，40(5): 117-122.

[9] 于江德，王希杰，樊孝忠. 词位标注汉语分词中特征模板定量研究 [J]. 计算机工程与设计，2012，33(3): 1239-1244.

[10] Zhao Hai，Huang Changning，Li Mu. An improved Chinese word segmentation system with conditional random field [C] //Proc of the 5th SIGHAN Workshop on Chinese Language Processing. 2006: 162-165.

[11] 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.

[12] 罗彦彦，黄德根. 基于 CRFs 边缘概率的中文分词 [J]. 中文信息学报，2009，23(5): 3-8.

[13] 赵海，揭春雨. 基于有效子串标注的中文分词 [J]. 中文信息学报，2007，21(5): 8-13.

[14] 黄德根，焦世斗，周惠巍. 基于子词的双层 CRFs 中文分词 [J]. 计算机研究与发展，2015，47(5): 962-968.

[15] 徐浩煜，任智慧，施俊，等. 基于链式条件随机场的中文分词改进方法 [J]. 计算机应用与软件，2016，33(12): 210-213.

[16] Zheng Xiaoqing，Chen Hanyang，Xu Tianyu. Deep learning for Chinese word segmentation and POS tagging [C] //Proc of Conference on Empirical Methods in Natural Language Processing. 2013: 647-657.

(下转第 1341 页)

可以有效地促进各粒子逼近真实 Pareto 前沿面，提高所求解的质量。

表 3　Pareto 最优解集

序号	目标					
	最大完工时间 $C_{\max}$	总加工成本 M	设备总负荷 W_T	最大设备负荷 W_M	加工质量 Q	欧氏贴近度 τ
1	132	856.8	408	71	1.895 3	0.800 12
2	130	854.7	409	74	1.959 3	0.825 51
3	114	847.6	426	71	1.904 8	0.814 51
4	111	851.6	413	73	1.942 6	0.842 16
5	115	836.6	424	77	1.904 9	0.812 82
6	113	847.6	420	76	1.900 2	0.831 54
7	116	841.5	421	76	1.905 9	0.821 31
8	117	839.9	420	73	1.868 4	0.812 4
9	118	851.7	429	64	1.886 2	0.805 03
10	124	836.5	432	71	1.853 2	0.810 23
11	128	848.7	432	70	1.854 1	0.785 91
12	126	840.4	423	68	1.845 8	0.798 19
13	135	859.6	423	68	1.901 2	0.795 43
14	121	855.4	421	80	2.142 2	0.838 41
15	109	848.6	437	81	1.981 3	0.793 1

仿真实验表明，该方法能够较好地解决高维多目标柔性作业车间调度问题。本文所提算法是基于标准粒子群算法的，缺乏对种群多样性的有效控制，所以接下来将借助小生境技术对所提算法作进一步的改进。

参考文献：

[1] Zhang Qingfu, Li Hui. MOEA/D: a multiobjective evolutionary algorithm based on decomposition [J]. IEEE Trans on Evolutionary Computation, 2007, 11(6): 712-731.

[2] 白俊杰，王宁生，唐敦兵. 一种求解多目标柔性作业车间调度的改进粒子群算法 [J]. 南京航空航天大学学报，2010，42(4)：447-453.

[3] 郭思涵，龚小胜. 正交设计的 E 占优策略求解高维多目标优化问题研究 [J]. 计算机科学，2012，39(2)：276-279.

[4] 章恩泽，陈庆伟. 改进的 r 支配高维多目标粒子群优化算法 [J]. 控制理论与应用，2015，32(5)：623-630.

[5] Said L B, Bechikh S, Ghédira K. The r-dominance: a new dominance relation for interactive evolutionary multicriteria decision making [J]. IEEE Trans on Evolutionary Computation, 2010, 14(5): 801-818.

[6] Kang Zhuo, Kang Lishan, Zou Xiufen, *et al*. A new evolutionary decision theory for many-objective optimization problems [C]//Advances in Computation and Intelligence. Berlin: Springer, 2007: 1-11.

[7] 肖婧，王科俊，毕晓君. 基于改进 K 支配排序的高维多目标进化算法 [J]. 控制与决策，2014，29(12)：2165-2170.

[8] Sinha A, Saxena D K, Deb K, *et al*. Using objective reduction and interactive procedure to handle many-objective optimization problems [J]. Applied Soft Computing, 2013, 13(1): 415-427.

[9] Saxena D. Searching for Pareto-optimal solutions through dimensionality reduction for certain large-dimensional multi-objective optimization problems [C]//Proc of World Congress on Computational Intelligence. 2006: 3352-3360.

[10] Jaimes A L, Coello C A C, Chakraborty D. Objective reduction using a feature selection technique [C]//Proc of the 10th Annual Conference on Genetic and Evolutionary Computation. 2008: 673-680.

[11] Saxena D K, Duro J A, Tiwari A, *et al*. Objective reduction in many-objective optimization: linear and nonlinear algorithms [J]. IEEE Trans on Evolutionary Computation, 2013, 17(1): 77-99.

[12] 车晓毅，罗佑新，汪超. 高维多目标优化设计的灰色微粒群算法 [J]. 机械传动，2009，33(1)：55-58.

[13] 巩敦卫，刘益萍，孙晓燕，等. 基于目标分解的高维多目标并行进化优化方法 [J]. 自动化学报，2015，41(8)：1438-1451.

[14] 张超勇，饶运清，李培根，等. 柔性作业车间调度问题的两级遗传算法 [J]. 机械工程学报，2007，43(4)：119-124.

[15] Zhang Guohui, Shao Xinyu, Li Peigen, *et al*. An effective hybrid particle swarm optimization algorithm for multi-objective flexible Job-Shop scheduling problem [J]. Computers & Industrial Engineering, 2009, 56(4): 1309-1318.

[16] 侯晓莉，刘永，江来臻，等. 多目标 FJSP 的一维编码粒子群优化求解方法 [J]. 计算机工程与应用，2015，51(13)：47-51.

(上接第 1324 页)

[17] Collobert R, Weston J, Bottou L, *et al*. Natural language processing (almost) from scratch [J]. Journal of Machine Learning Research, 2011, 12(1): 2493-2537.

[18] Bengio Y, Schwenk H, Senécal J S, *et al*. Neural probabilistic language models [M]//Innovations in Machine Learning. Berlin: Springer, 2006: 137-186.

[19] Pei Wenzhe, Ge Tao, Chang Baobao. Max-margin tensor neural network for Chinese word segmentation [C]//Proc of Annual Meeting of the Association for Computational Linguistics. 2014: 293-303.

[20] Chen Xinchi, Qiu Xipeng, Zhu Chenxi, *et al*. Long short-term memory neural networks for chinese word segmentation [C]//Proc of Conference on Empirical Methods in Natural Language Processing. 2015.

[21] Chen Xinchi, Qiu Xipeng, Zhu Chenxi, *et al*. Gated recursive neural network for Chinese word segmentation [C]//Proc of Annual Meeting of the Association for Computational Linguistics. 2015.

[22] Tseng H, Chang P, Andrew G, *et al*. A conditional random field word segmenter for sighan bakeoff [C]//Proc of the 4th SIGHAN Workshop on Chinese Language Processing. 2005: 168-171.

[23] Liu Pengfei, Qiu Xipeng, Chen Xinchi, *et al*. Multi-timescale long short-term memory neural network for modelling sentences and documents [C]//Proc of Conference on Empirical Methods in Natural Language Processing. 2005: 2326-2335.

[24] Wang Xin, Liu Yuanchao, Sun Chengjie, *et al*. Predicting polarities of tweets by composing word embeddings with long short-term memory [C]//Proc of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. 2015: 1343-1353.

[25] Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks [C]//Advances in Neural Information Processing Systems. 2014: 3104-3112.

[26] Wang Di, Nyberg E. A long short-term memory model for answer sentence selection in question answering [C]//Proc of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. 2015: 707-712.

[27] Ghosh S, Vinyals O, Strope B, *et al*. Contextual LSTM (CLSTM) models for large scale NLP tasks [J]. ArXiv Preprint ArXiv: 1602.06291, 2016.

[28] Vinyals O, Toshev A, Bengio S, *et al*. Show and tell: A neural image caption generator [C]//Proc of IEEE Conference on Computer Vision and Pattern Recognition. 2015: 3156-3164.

[29] Mikolov T, Chen Kai, Corrado G, *et al*. Efficient estimation of word representations in vector space [J]. Computer Science, 2013.

[30] Chen Tianqi, Li Mu, Li Yutian, *et al*. MXNet: a flexible and efficient machine learning library for heterogeneous distributed systems [J]. Statistics, 2015.

[31] SIGHAN. http://www.sighan.org/bakeoff2005/data/results.php.htm [EB/OL].