设计研究与应用

融合 attention 机制的 BI-LSTM-CRF 中文分词模型

黄丹丹，郭玉翠

（北京邮电大学理学院 北京 100876）

摘 要： 中文的词语不同于英文单词，没有空格作为自然分界符，因此，为了使机器能够识别中文的词语需要进行分词操作。深度学习在中文分词任务上的研究与应用已经有了一些突破性成果,本文在已有工作的基础上，提出融合 Bi-LSTM-CRF 模型与 attention 机制的方法,并且引入去噪机制对字向量表示进行过滤,此外为改进单向 LSTM 对后文依赖性不足的缺点引入了贡献率λ对 BI-LSTM 的输出权重矩阵进行调节，以提升分词效果。使用改进后的模型对一些公开数据集进行了实验。实验结果表明，改进的 attention-BI-LSTM-CRF 模型以及训练方法可以有效地解决中文自然语言处理中的分词、词性标注等问题，并较以前的模型有更优秀的性能。

关键词： 中文分词；BI-LSTM；CRF；attention 机制；贡献因子；去噪机制；Dropout

中图分类号： TP391 **文献标识码：** A **DOI：** 10.3969/j.issn.1003-6970.2018.10.050

本文著录格式： 黄丹丹，郭玉翠. 融合 attention 机制的 BI-LSTM-CRF 中文分词模型[J]. 软件，2018，39（10）：260–266

BI-LSTM-CRF Chinese Word Segmentation Model with Attention Mechanism

HUANG Dan-dan, GUO Yu-cui

(*School of Science, Beijing University of Posts and Telecommunications, Beijing* 100876, *China*)

【Abstract】: In English words, spaces are used as natural delimiters between words, and there are no such clear delimiters between Chinese words. Therefore, deep learning models and methods that obtain good results in English natural language processing cannot be directly applied. Deep learning has achieved breakthrough results in the field of natural language processing in English. Based on the existing work, this paper proposes a method to integrate the Bi-LSTM-CRF model and the attention mechanism, and introduces a denoising mechanism to filter the word vector representation.In addition, the contribution rate λ of the unidirectional LSTM is reduced. The output weight matrix of the BI-LSTM is adjusted to improve the word segmentation effect. We conducted experiments using the public data set in the above model. Experimental results show that the improved attention-BI-LSTM-CRF model and training method can effectively solve the problem of word segmentation and part of speech tagging in Chinese natural language processing, and can obtain good performance.

【Key words】: Chinese segmentation; BI-LSTM; CRF; Attention mechanism; Contribution factor; Denoising mechanism; Dropout

0 引言

随着人工智能在越来越多领域的突破，基于深度学习的自然语言处理这一重要领域已经引起了众多研究者的关注。分词、词性标注作为中文自然语言处理中最重要的基础工作之一，已经取得了一些成果。本文在已有结果的基础上深入研究深度学习在中文分词中的应用。分词是指将未加工的自然语言文本分割成单词的顺序。在英语中，单词之间以空格作为的自然分隔符，但在中文中汉字之间没有明显区分。因此需要将中文文本序列进行分割，使之转变成单词序列，以便后续的中文信息处理。

从机器学习角度来看，分词任务可转化成序列标注任务（或者分类任务）。序列标注任务指将观察序列中的每个元素在固定标签集合中为之赋予一个指定标签的过程（分类的过程）。目前，常用的解决

作者简介：黄丹丹(1991–)，女，研究生，主要研究方向：自然语言处理；郭玉翠(1962–)，女，教授，主要研究方向：数学与信息安全。

序列标记任务的模型有隐马尔可夫模型[1]、条件随机场模型[2,3]和最大熵模型[4]。然而，这些传统的模型需要使用大量的语言学知识来手工构造特征，因此不具有广泛的适用性。深度学习有效利用无监督数据，避免繁琐的人工特征提取，从而具有良好的泛化能力。它通过对数据的多层次建模从而得到数据特征的层次结构以及数据的分布式表示。

深度学习用来解决自然语言处理领域的一些难题。语言的高维特性导致了传统的自然语言处理系统需要复杂的语言知识以便手动构造分类器所能使用的特征。深度学习的方法有以下优点：（1）通过构建模型，可以自动学习自然语言处理领域中解决问题所需要的特征。Collobert 等[5]就是利用该特性，抛弃传统的手工提取特征方式，解决了英文序列标注问题。（2）在自然语言处理领域，获得标记数据相对于获得大量的无标记数据成本较大，深入学习可以使用大量的无标记数据来获取特征。（3）自然语言处理领域中的许多问题是密切相关的，如分词、词性标注和命名实体识别等。传统的方法往往单独解决这些问题，而忽略了它们之间的关系。使用深度学习，您可以在特征提取级别构建统一模型以同时处理这些问题，并使用多任务学习方法在模型中建模其相关性以获得更好的性能。Zheng 等[6]利用 SENNA 系统将神经网络运用到中文分词任务上，并提出一个感知器算法加速整个训练过程。Chen 等[7,8]在 GRNN 模型基础上提出了 LSTM（long short-term memory）模型进行中文分词任务，取得了很好的效果。之后，Yao 等人[9]在 LSTM 模型的基础上提出了 BI-LSTM 模型，更进一步提高了中文分词的准确度。

本文在适合于中文自然语言处理的双向长短期记忆条件随机场模型（BI-LISTM-CRF）基础上，进行了以下改进：（1）提出一种去噪机制，对字向量表示进行调整，使得固定窗口内的字嵌入以一定概率出现，不再依赖于左右联合字嵌入的共同作用；（2）引入了贡献因子对前传 LSTM 层和后传 LSTM 层的权重矩阵进行调节以改进单向 LSTM 对后文依赖性不足的缺点；（3）在 BI-LSTM-CRF 中文分词模型中融合 attention 机制，通过注意机制计算 Bi-LSTM 模型的输入和输出之间的相关性的重要性，并根据重要性程度获得文本的整体特征。利用改进的 attention-BI-LSTM-CRF 模型，在 MSRA corpus、PKU corpus 和人民日报 2014 公开数据集上进行了实验。实验结果表明，使用本文改进的模型以及训练方法可以有效地进行中文自然语言处理中的分词问题，并提高了精度。

1 模型建立

本文采用图 1 所示的 attention-BI-LSTM-CRF

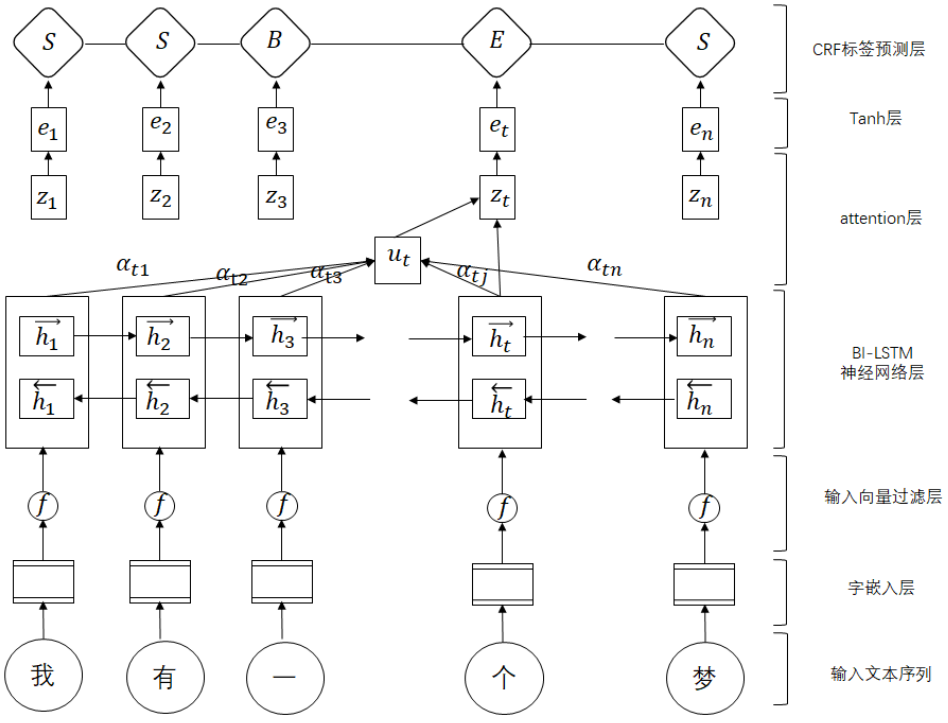

图 1 Attention-BILSTM-CRF 中文分词模型

Fig.1 Attention-BILSTM-CRF Chinese word segmentation model

中文分词模型来进行中文分词处理。自底向上：（1）将待分词的文本序列进行文本向量化，将文本中的每一个字映射成一个固定长度的短向量，以作为当前字的特征向量表示；（2）基于去噪机制对输入的信息进行过滤调整；（3）利用 BI-LSTM 获取每个词长距离的上下文特征；（4）引入 attention 模型对 BI-LSTM 层的输入与输出之间的相关性进行重要度计算，根据重要度获取文本整体特征；（5）最后 CRF 层考虑单词标签之间的制约关系，加入标签转移概率矩阵，给出全局最优标注序列。

1.1 LSTM 和 BI-LSTM

LSTM 和 BI-LSTM 都是递归神经网络[10][11]（RNN）的变体。递归神经网络是一种对连续数据进行操作的神经网络。它们将序列 $(x_1,x_2,\cdots,x_n)$ 作为输入并返回另一个序列 $(h_1,h_2,\cdots h_n)$，该返回序列包含输入中每个步骤的序列信息。具体地运作过程如图 2 所述：t 时刻输入当前信息 x_t 并由神经网络模块 A 接受，之后由 A 得到 t 时刻的输出 h_t，且将当前时刻的部分信息传递到下一刻 t+1。由图 1 我们清晰地看到，每一时刻的神经网络模块都是 A，不同的只是当前时刻的输入以及上一时刻的输出。

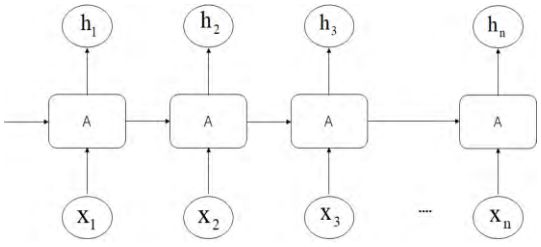

图 2 RNN 神经网络结构
Fig.2 RNN neural network structure

RNN 理论上可以学习长期的依赖关系，但在实际情况中并不是如此，它们更倾向于最近的输入序列。这是由于传统的 RNN 在进行几次链式法则求导后梯度会指数级缩小，导致传播几层后出现梯度消失，无法处理“长期依赖”问题。因此，出现了一种 RNN 的变体即 LSTM。LSTM 的设计旨在通过整合一个存储单元来解决这个问题，并被证明可以捕获远距离依赖。他们使用几个门来控制输入给存储单元的比例，以及从以前的状态中忘记的比例[12]。

LSTM 的结构与 RNN 一致，唯一的不同在于其中间的神经网络模块 A。该模块结构如图 3 所示。

如图 3，其中：输入门 i_t：控制当前时刻的输入和前一时刻输出进入新的 cell 的信息量；忘记门 f_t：决定需要舍弃的信息部分；细胞状态更新 c_t：计算下一个时间戳的状态；输出门 o_t：计算 cell 的输出；最终 LSTM 的输出 h_t：使用一个对当前状态的 softmax 变换进行重变换。

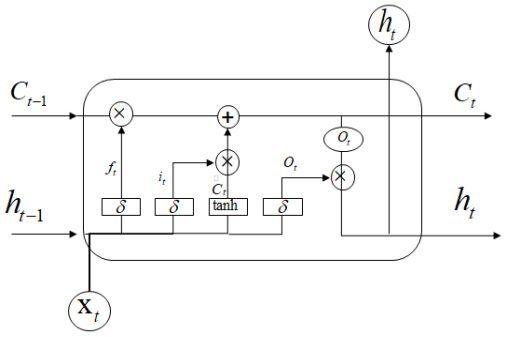

图 3 LSTM 神经网络模块结构
Fig.3 LSTM neural network module structure

根据上图，$x=(x_1,x_2,\cdots,x_t)$ 为输入数据，$h=(h_1,h_2,\cdots,h_t)$ 为 LSTM 单元的输出，相关状态的更新以及计算公式：

$$\begin{aligned}\tilde{c}_t &= g(w_{cx}x_t + w_{ch}h_{t-1} + b_c)\\ i_t &= \sigma(w_{ix}x_t + w_{ih}h_{t-1} + b_i)\\ f_t &= \sigma(w_{fx}x_t + w_{fh}h_{t-1} + b_f)\\ c_t &= f_t \odot c_{t-1} + i_t \odot \tilde{c}_t\\ o_t &= \sigma(w_{ox}x_t + w_{oh}h_{t-1} + b_o)\\ h_t &= o_t \odot g(c_t)\end{aligned} \quad (1)$$

其中 w 代表各个权重矩阵，如 w_{ix} 是输入门到输出的权重矩阵，b 代表偏置向量，如 b_i 是输入门的偏置向量，σ 是 sigmoid 函数，i,f,o,c 分别代表输入门，忘记 s 门，输出门以及 cell 状态更新向量，$\odot$ 代表点乘，g,h 分别为 cell 的输入输出激活函数，通常为 tanh 。

对于包含 n 个词的给定句子 $x=(x_1,x_2,\cdots,x_t)$，每个词表示为 d 维向量，LSTM 计算每个词 t 处句子的上下文的表示。但作为分词模型 LSTM 也有它的局限性，它不能考虑到未来的上下文信息。比如：‘我们’这个词，若无法考虑到“我”后面的上下文信息，那么就会把“我”单独分词。对于这种情况，可以通过使用第二个 LSTM 来反向读取相同的序列。结合向前和向后的 LSTM 对被称为双向 LSTM，简称 BI-LSTM。其结构如图 4 所示。使用 BI-LSTM 模型的单词表示是通过连接其左右上下文表示 $h_t=[\overrightarrow{h_t};\overleftarrow{h_t}]$。这些表示有效地包含了上下文中单词的表示，从而既能保存前面的上下文又能同时考虑到未来的上下

文信息。BI-LSTM 神经网络结构可以同时捕捉到两个方向的长距离信息，因此具有更好的表现。

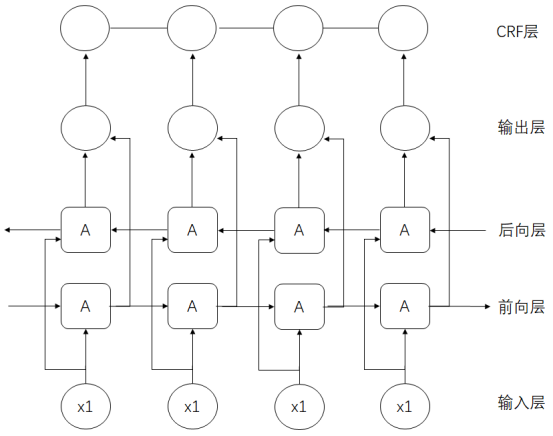

图 4　BILSTM+CRF 模型
Fig.4　BILSTM+CRF model

1.2　标签得分计算

中文分词问题可以转换为字符序列的字符标签分类问题。1.1 节中 BI-LSTM 神经网络的中文分词模型的输出即为字符序列中每一个字符的标签得分。文中采用 BMES 标注方法对分词语料库文本进行标注，即每个字符用{B,M,E,S}来分别表示字符在词中的开始位置、中间位置、结束位置以及单个字为一个独立词。比如文本序列‘我们都是共产主义接班人’用{B,M,E,S}来分割后为‘我/B 们/E 都/B 是/E 共/B 产/M 主/M 义/E 接/B 班/M 人/E’。

对于分词任务来说，自前往后获得的信息量与自后往前获得的信息量是不对等的，前者要大于后者。也就是说 t 时刻，由顺序传播 LSTM 层获得的 $\overrightarrow{h_t}$ 与由逆向传播 LSTM 层获得的 $\overleftarrow{h_t}$ 贡献不同，并且我们猜想，当 $t \geqslant \frac{n}{2}$ 时，$\overrightarrow{h_t}$ 的贡献相比较 $\overleftarrow{h_t}$ 会更多。

由此猜想，我们引入一个贡献因子变量 λ ($0 \leqslant \lambda \leqslant 1$)，且规定：当 $t \geqslant \frac{n}{2}$ 时(n 为序列长度)，$\lambda \geqslant 0.5$。在引入 λ 的条件下，我们的 BI-LSTM 神经网络的输出表示为：

$$h_t = \lambda \overrightarrow{h_t} + (1-\lambda)\overleftarrow{h_t} \tag{2}$$

1.3　CRF 标注模型

一个简单但效果显著的有效标注模型叫条件随机场（CRF）[13]。它根据给定的观察序列来推测出对应的状态序列，属于一种条件概率模型。CRF 由 Lafferty 等人于 2001 年提出，它解决了隐马尔可夫模型的输出独立性假设问题，也解决了最大熵模型在每一个节点归一化导致只能找到局部最优解和标记偏见问题，因此是比较好的命名实体识别模型。CRF 的序列标注思想和 BI-LSTM 模型利用前后上下文特征的思想上有相向之处，在文献[13]和文献[9]中分别证明了该类模型性能相较于只考虑单方面影响的模型性能有更好的表现。

CRF 的工作原理如下：

对于输入的可观测序列 $X=(x_1,x_2,\cdots,x_t)$，状态序列 $y=(y_1,y_2,\cdots,y_n)$ 为最终的预测结果序列。首先我们定义评分：

$$s(X,y)=\sum_{i=0}^{n} A_{y_i,y_{i+1}} + \sum_{i=1}^{n} P_{i,y_i} \tag{3}$$

其中 A 是过渡分数的矩阵，使得 $A_{i,j}$ 表示从标签 i 到标签 j 的过渡分数。P 是由 BI-LSTM 网络输出的分数矩阵。P 的大小为 n×k，其中 k 是不同标注的数量，$p_{i,j}$ 对应于句子中第 i 个单词的第 j 个标签的得分。y_0 和 y_n 是句子的开始和结束标签，我们将其添加到可能的标签集合中。因此 A 是大小为 k + 2 的方阵。

在所有可能的标签序列上产生序列 y 的概率为：

$$P(y \mid X)=\frac{e^{s(X,y)}}{\sum_{\tilde{y}\in Y_X} e^{s(X,\tilde{y})}} \tag{4}$$

训练期间，目标函数是最大化正确标签序列的对数概率：

$$\log(p(y \mid X)) = s(X,y) - \log\left(\sum_{\tilde{y}\in Y_X} e^{s(X,\tilde{y})}\right) = s(X,y) - \log_{\tilde{y}\in Y_X} add\, s(X,\tilde{y}) \tag{5}$$

其中 Y 代表句子 X 的所有可能的标签序列。从上述公式可以看出，CRF 是学习一个从观察序列到标记序列的概率函数映射关系。我们鼓励我们的网络生成一个有效的输出标签序列。在预测过程（解码）中，模型使用动态规划的 Viterbi 算法来获得最大分数的输出序列：

$$y^* = \arg\max_{\tilde{y}\in Y_X} s(X,\tilde{y}) \tag{6}$$

1.4　引入 attention 机制

深度学习中的 attention 机制模拟人脑的注意力特点。Attention 机制可以理解为总是将注意力放在更重要的信息上。Bahdanau 等在论文[14]中第一次提出把 attention 机制应用到了神经网络机器翻译上。

本文中引入 attention 机制以突出特定的字对于

整个文本的重要程度。如图 1 所示，我们在 BI-LSTM 层之后添加 attention 层，用矩阵 T 来计算当前目标字与输入文本中所有字的相似性。注意力权重系数 α_{tj}（矩阵 T 的第 t 行第 j 列）表示第 t 个目标输出与第 j 个输入的相似性，α_{tj} 值越大，表示在生成第 t 个输出的时候受第 j 个输入的影响也就越大。计算方法如下：

$$\alpha_{tj}=\frac{\exp(s_{tj})}{\sum_{k=1}^{n}\exp(s_{tk})} \quad (7)$$

其中：

$$s_{tj}=s(x_t,x_j)=\begin{cases} w_a\,|x_t-x_j|^T|x_t-x_j|\cdots(a) \\ \dfrac{w_a(x_t\cdot x_j)}{|x_t||x_j|}\cdots(b) \end{cases}$$

s_{tj} 被定义为括号中的两种形式，分别表示为欧式距离，余弦定理。当 x_t 和 x_j 越相似的时候余弦距离的值会越大，相反，欧式距离值会越小。s_{tj} 最终结果为 b-a 的值。

接着，用一个全局变量 u_t 表示解码阶段的第 t 时间序列，h_j 为 1.3 节中（7）式表示的 BI-LSTM 层的输出编码的权重之和。计算方法如下：

$$u_t=\sum_{j}^{n}\alpha_{tj}h_j \quad (8)$$

将全局变量 u_t 与 BI-LSTM 层的输出 h_t 合并成一个向量 $[u_t;h_t]$，再将其喂给一个 tanh 函数作为 attention 层的输出。即：

$$z_t=\tanh(w_u[u_t;h_t]) \quad (9)$$

在 attention 层之后用一个 tanh 层用来预测神经网络输出的标签得分：

$$e_t=\tanh(w_e z_t) \quad (10)$$

Attention-BI-LSTM-CRF 模型在 BI-LSTM 网络与 CRF 标签判别层中间添加 attention 层。我们将字嵌入序列作为输入提供给 BI-LSTM，通过 BI-LSTM 层结合了上下文的特征，输出返回每个字的上下文的表示，并结合 attention 机制将更有效的信息输入向 CRF 层，使用 CRF 来考虑相邻标签，从而得出每个字的最终预测结果。

2 输入字嵌入

本节介绍输入字嵌入，用特征向量表示序列文本作为 BI-LSTM 层的输入：（1）将原始序列文本进行向量化，用一个固定长度的向量表示每一个字；（2）由于固定窗口大小带来的上下文不确定性，引入去噪机制对特征向量进行调整；（3）最后使用 dropout 技巧防止训练过程中的过拟合问题。

2.1 文本向量化

为了使机器能够理解自然语言首先需要将自然语言符号数学化，即文本向量化。在深度学习中，将文本向量化的方式使采用分布式表示方法[15]（又称字嵌入）。该方法将词用一种低维实数向量表示，这样的表示既能够使得上下文的词之间的彼此联系，又可以避免向量维度过大带来的不必要的复杂度。

具体地，在中文分词任务中，我们建立一个大小为 d×N 的汉字字典矩阵 D，其中 d 为字向量维度，N 为字典大小。该字典包括我们可以处理的所有汉字以及其他字符（如数字、标点、未登录字等）的替代符号。因此，我们用字典找到对应的字向量来代替每个字。研究表明，将大规模无监督学习得到的字向量作为输入矩阵的初始值要比随机初始化得来的字向量性能上表现更优[16]。本文实验中使用 word2vec 作为第一层，把输入数据预先处理成字嵌入向量。

2.2 输入去噪

本文对当前字设置了特征窗口，即利用固定上下文窗口内的字表示当前字。但是固定窗口内的字不一定每次都出现在一起，有的可能只出现少数次。因此，加入一个去噪层对固定窗口内的信息进行调整，使得固定窗口内的字嵌入以一定概率出现，不再依赖于固定窗口内左右词的字向量的共同作用。

首先，句中每个字的字向量表示作为去噪机制的输入。然后该机制对输入信息进行调整，之后 BI-LSTM 获取每个词长距离的上下文特征并由 attention 机制对 BI-LSTM 层的输入与输出之间的相关性进行重要度计算获取文本整体特征，最后 CRF 层考虑单词标签之间的制约关系，加入标签转移概率矩阵，给出全局最优标注序列。

去噪机制实质上为一个神经网络层，其输出在 0~1 之间。设 x_t 是第 t 个字的原始特征向量表示，经过去噪机制 f_t 的选择，得到输出 k_t：

$$k_t=f_t\odot x_t \quad (11)$$

其中 $\odot$ 表示逐点乘积操作。f_t 定义如下：

$$f_t=f(w_f x_t+b_f) \quad (12)$$

其中，f 函数为

$$f(x)=|\sin x| \quad (13)$$

w_f 表示当前层的权值矩阵，b_f 表示偏置向量。

2.3 Dropout 技巧

为了防止模型训练过程中的过拟合问题，本文采用了 Dropout[17]技术。其主要思想是在模型训练过程中,随机移除一定比例 p（Dropout 比率）的神经元以及其对应的输入输出权重。我们将输入 attention-BI-LSTM-CRF 模型的字嵌入向量使用 Dropout 方法以降低错误率，提升系统性能。

3 实验

为了说明改进的模型的有效性，我们选择常用的 MSRA corpus、PKU corpus 和人民日报 2014 作对比实验。其中 MSRA 和 PKU corpus 是由国际中文分词评测 Bakeoff 提供的封闭语料，包括简体中文和繁体中文。

实验过程中为了公正的评估模型的分词性能，我们采用了分词常用的评价指标:准确率（P），召回率（R），综合指标值（F1）。

3.1 贡献因子与去噪机制测试

为验证本文提出的贡献因子和去噪机制是否会影响到实验效果,我们选取 1 层 BI-LSTM 分词模型，句子长度为 80，在 MSRA 数据集上进行测试，测试结果如表 1 所示。

表 1 贡献因子和去噪机制在 MSRA 测试集上测试结果（F1 值）

Tab.1 Contribution factor and denoising mechanism test results on the MSRA test set (F1 value)

λ 取值	去噪	不去噪
0.5	0.963	0.957
0.6	0.966	0.96
0.7	0.971	0.964
0.8	0.968	0.961
0.9	0.966	0.958
1	0.956	0.952

从表中观察到，当 λ 取相同值时，加入去噪机制时的分词模型表现更优；并且发现当 λ 取值为 0.7 时，实验效果最好，λ 取值 1.0 时相比较 0.5 时表现稍逊。证明了去噪机制和贡献因子的有效性。

为了进一步验证贡献因子的作用，我们作如下测试：在 $t>40$ 时，我们设定 λ 为 0.7；$t<40$ 时我们选取 0.1、0.3、0.5、0.7、0.9 五个不同值做对比实验，结果如表 2。

表 2 t<40 时不同取值的贡献因子测试结果

Tab.2 Contributing factor test results with different values when t<40

λ 取值	F1 值
0.1	0.958
0.3	0.965
0.5	0.970
0.7	0.964
0.9	0.960

由表 2 发现，在 $t>40$ 时设定贡献因子为 0.7 的条件下，当 $t<40$ 时，贡献因子取值 0.5 时优于 0.7 时的表现。这说明在训练模型的不同时刻，顺序传播 LSTM 层获得的 $\vec{h}_t$ 与逆向传播 LSTM 层获得的 $\overleftarrow{h}_t$ 贡献不同。本论文中在训练 attention-BI-LSTM-CRF 模型时为了便捷，统一设定贡献因子为 0.7。

3.2 超参数配置

对于本文改进的 attention-BI-LSTM-CRF 模型，我们使用反向传播算法来训练我们的网络，设定初始学习率为 0.01。本实验采用 word2vec 方法对字向量进行训练预处理。文中使用 PKU 数据集，基于 BMES 词位标注方法，以 BI-LSTM 为模型，我们设定字嵌入向量长度为 100，dropout 大小为 0.3。实验研究过程中，我们发现不断增大的隐藏层单元数当达到一定值以后，对测试结果影响趋于稳定。本文中改进的模型最终选取隐藏层的单元数为 120。

表 3 超参数设置

Tab.3 Hyperparameter setting

超参数	取值
上下文窗口长度	k=5
隐藏层单元数	H=120
初始学习速率	α=0.01
字向量长度	d=100
贡献因子	λ=0.7
Dropout 比例	P=0.3

3.3 实验对比与分析

我们测试 BI-LSTM、BI-LSTM-CRF、和本文改进的 attention-BI-LSTM-CRF 这三个不同的模型分别在 PKU，MSRA 和人民日报 2014 语料库上分词性能的表现。如表 4 所示，本文提出的 attention-BI-LSTM-CRF 模型相比较 BI-LSTM 和 BI-LSTM-CRF 模型性能分别提升为 0.6%、1.0%和 0.6%，分词效果更好。

表 5 为本文训练的 attention-BI-LSTM-CRF 模型

表 4 不同模型在 PKU、MSRA、人民日报 2014 测试集上的实验对比结果

Tab.4 Experimental comparison results of different models on PKU, MSRA, People's Daily 2014 test set

模型	PKU			MSRA			人民日报 2014		
	P	R	F	P	R	F	P	R	F
BI-LSTM	0.953	0.946	0.949	0.962	0.952	0.957	0.963	0.946	0.954
BI-LSTM-CRF	0.962	0.956	0.959	0.971	0.960	0.965	0.974	0.975	0.974
attention-BI-LSTM-CRF	0.966	0.964	0.965	0.975	0.976	0.975	0.982	0.978	0.980

表 5 在 PKU、MSRA 测试集上与前人模型的实验结果对比

Tab.5 Comparison of experimental results with predecessor models on PKU and MSRA test sets

模型	PKU			MSRA		
	P	R	F	P	R	F
Bakeoff-Best	0.946	0.953	0.949	0.966	0.982	0.974
FDU-LSTM	0.958	0.955	0.956	0.967	0.962	0.964
chen-2015	0.965	0.963	0.964	0.974	0.978	0.976
Yao-2016	0.968	0.963	0.965	0.974	0.973	0.973
attention-BI-LSTM-CRF	0.966	0.964	0.965	0.975	0.976	0.975

与前人在分词领域研究结果对比。其中 Bakeoff-best 为 2005 年 Bakeoff 测评最好结果；Chen-2015[7]他们在文本向量化过程中加入了双字符嵌入向量,最佳水平如表 5 所示；Yao-2016 在文献[9]中叠加了 3 层 BI-LSTM 模型。本文中融合了 attention 机制与过滤机制以及引入了贡献因子也取得了不错的分词效果，证明了 attention-BI-LSTM-CRF 分词模型的优越性。

4 结语

文中针对自然语言处理中的中文分词任务，在 BI-LSTM-CRF 模型的基础上提出一种改进的 attention-BI-LSTM-CRF 中文分词模型。该模型融合 attention 机制方法，以计算 BI-LSTM 模型的输入和输出之间相关性的重要性，从而更好的获得文本的整体特征。利用一种去噪机制，使得固定窗口内的字嵌入以一定概率出现，减少了左右联合字嵌入的联合作用。并且引入了贡献因子以改进单向 LSTM 对后文依赖性不足的缺点。实验表明，在中文分词任务中，相比较 BI-LSTM 模型和 BI-LSTM-CRF 模型，本文改进的 attention-BI-LSTM-CRF 模型在选取的测试集上分词表现更加出色。

参考文献

[1] 李月伦，常宝宝. 基于最大间隔马尔可夫网模型的汉语分词方法[J]. 中文信息学报, 2010, 24(1): 8-14.

[2] Peng F, Feng F, Mccallum A. Chinese segmentation and new word detection using conditional random fields[C]. Proceedings of Coling, 2004: 562-568.

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

[4] Nianwen Xue. Chinese word segmentation as char- acter tagging[J]. Computational Linguistics and Chi- nese Language Processing, 2003, 8(1): 29-48.

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

[6] Zheng X, Chen H, Xu T. Deep learning for Chinese word segmentation and POS tagging[C]. Confer- ence on Empirical Methods in Natural Language Processing, 2013: 647-657.

[7] Chen X, Qiu X, Zhu C, et al. Gated recursive neu- ral network for Chinese word segmentation[C]. Proc of Annual Meeting of the Association for Computational Linguistics, 2015: 1744-1753.

[8] Chen X, Qiu X, Zhu C, et al. Long short-term memory neural networks for Chinese word seg- mentation[C]. Conference on Empirical Methods in Natural Language Processing, 2015: 1197-1206.

[9] Yushi Yao, Zheng Huang. Bi-directional LSTM recurrent neural network for Chinese word segmen- tation[C]. International Conference on Neural In-formation Processing, 2016: 345-353.

[10] Y. Bengio; P. Simard; P. Frasconi, Learning long-term dependencies with gradient descent is difficult. IEEE Transactions on Neural Networks, 2002, 5(2): 157-166.

[11] 张玉环，钱江. 基于两种 LSTM 结构的文本情感分析[J]. 软件, 2018, 39(1): 116-120.

[12] S Hochreiter, J Schmidhuber, LSTM can solve hard long time lag problems. International Conference on Neural Information, 1996, 9: 473-479.

[13] Lafferty J D, Mccallum A, Pereira F C N. Conditional random fields: probabilistic models for seg- menting and labeling sequence data[C]. Proc of ICML, 2002, 3(2): 282-289.

[14] Neural Machine Translation by Jointly Learning to Align and Translate. D Bahdanau, K Cho, Y Bengio - arXiv preprint arXiv: 1409. 0473, 2014.

[15] Hinton G E. Learning distributed representations of concepts[C]//Proceedings of the eighth annual conference of the cognitive science society. 1986: 1-12.

[16] Mulder W D, Bethard SMoens M F. A Survey on the application of recurrent neural networks to statistical language modelingJ. Computer Speech &Language, 2014, 30(1): 61-98.

[17] Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: a simple way to prevent neural networks from overfitting[J]. Journal of Machine Learning Research, 2014, 15(1): 1929-1958.