文献引用格式： 邵党国 , 黄初升 , 马磊 , 等 . 基于 Bi-LSTM 的医学文本分词模型 [J]. 通信技术 ,2022,55(2):151-159.
doi:10.3969/j.issn.1002-0802.2022.02.003

基于 Bi-LSTM 的医学文本分词模型 *

邵党国，黄初升，马　磊，贺建峰，易三莉
（昆明理工大学 信息工程与自动化学院，云南 昆明 650500）

摘　要： 中文分词（Chinese Word Segmentation，CWS）是自然语言处理（Natural Language Processing，NLP）中一项重要的基础任务。由于不同领域中文词汇的特殊性以及缺乏相关领域的标记数据，面向特定领域的分词任务是近年来中文分词研究面临的挑战之一。因此，提出了一种基于双向长短时记忆网络（Bidirectional Long Short-Term Memory，Bi-LSTM）的分词模型，其中分别使用了大规模的中文通用语料以及少量中文医学语料训练模型来构建这种全新的分词模型。该模型在仅使用少量中文医学领域标记数据的情况下，在该领域的分词任务上获得了较好的分词效果，实验结果中的最优 *F*1 值为 95.54%，相比单独使用中文医学语料训练的分词模型，有比较明显的提升。

关键词： 中文分词；自然语言处理；双向长短期记忆网络；领域分词
中图分类号： TP391.1　　**文献标识码：** A　　**文章编号：** 1002-0802(2022)-02-0151-09

Research on Chinese Word Segmentation in Medical Domain Based on Bi-Lstm

SHAO Dangguo, HUANG Chusheng, MA Lei, HE Jianfeng, YI Sanli
(Faculty and Information Engineering and Automation, Kunming University of Science and Technology, Kunming Yunnan 650500, China)

Abstract: CWS (Chinese word segmentation) is an important basic task in NLP (Natural Language Processing). Due to the particularity of Chinese vocabulary in their respective fields and the lack of labeled data in related fields, Chinese word segmentation tasks for specific fields is a difficult point of research in recent years. Therefore, this paper proposes a Bi-LSTM-based word segmentation model, which combines a large-scale Chinese general corpus training model and a small number of Chinese medical corpus training models to build a new Chinese word segmentation model. On the premise of using only a small scale of labeled data in the Chinese medical domain, a better experimental result is obtained. The optimal F-score in the experimental results is 95.54%, , which is a fairly good improvement compared to the word segmentation model trained with Chinese medical corpus alone.

Keywords: Chinese word segmentation; NLP (Natural Language Processing); Bi-LSTM (Bidirectional Long Short-Term Memory); domain word segmentation

0　引　言

近年来随着信息技术在中国的蓬勃发展，针对医疗机构的信息系统已经得到广泛推广，医疗机构在引入信息系统后，积累了大量的中文医学文本数据[1]。目前医疗机构的信息系统中的医学文本主要以医学影像报告、病理报告、电子病历等形式存在，大多是以自然语言的形式进行记录和保存的，而将这些文本中的相关信息提取出来并加以利用，对医

* **收稿日期：** 2021-10-14；**修回日期：** 2022-01-15　Received date:2021-10-14;Revised date:2022-01-15
基金项目： 国家自然科学基金项目（61702069，61741112）
Foundation Item: Project of National Natural Science Foundation of China (61702069, 61741112)

疗行业的发展有着巨大的意义[2]。

汉字是构成中文最基本的单位，在中文中汉字与汉字之间紧密排列，而相比于单个的汉字，词才是表征中文语义最精确的单位，但是中文中并没有明确的标记将一段话中的前后两词进行分隔，而中文分词的目的就是将中文文本基于某些标准划分为能够表征精确语义的词构成的序列[3]。

在中文的自然语言处理任务中，中文分词大都作为第一步的工作，分词的效果会直接影响后续的实体识别、语义分析、机器翻译、信息检索等自然语言处理相关工作的结果[4]。然而无论是传统的机器学习的分词方法还是基于深度学习的分词方法，其效果依赖于大规模的训练数据[5]，目前中文分词领域开放的数据集大多为通用的分词数据集，其在各领域的分词任务上效果并不好。由于医学领域具有极强的专业性，获取大量的标记的中文医学领域的训练数据比较困难，所以传统架构的中文分词方法暂时还无法在医学领域分词上实现期望的结果。

针对以上问题，笔者提出了一种基于双向长短期记忆网络（Bi-directional Long-Short Term Memory，Bi-LSTM）的分词模型。通过引入开放的分词语料生成的预训练模型，与中文医学领域的分词预训练模型进行融合，构建出这种全新的分词模型。该模型在仅使用少量的中文医学领域标记数据的同时，获得了较好的分词效果。

1 相关工作

目前无论是传统的机器学习模型还是基于神经网络的深度学习模型，大部分中文分词的思想都是将中文分词问题视为对单个汉字字符的序列标注问题[6]。此外，在深度学习进入大众视野之前，就已有学者对中文分词任务进行了大量的研究。

1.1 传统的中文分词机器学习模型

传统的基于机器学习的中文分词模型分为基于字标注的学习和基于词特征的学习两大类[7]。基于字标注学习的方法始于 Xue 等人[8]的研究，该工作通过使用位置标签（tag）来代表一个字在词中的位置，表达该字所携带的切分信息，从而将分词任务转化为字所在词中位置的序列标注学习任务。Low 等人[9]继续完善了这一思想，将严格的串标注学习应用于分词。2004 年，Peng 等人[10]将条件随机场（Conditional Random Fields，CRF）引入中文分词学习，自此，CRF 的多个变种在深度学习时代之前成为了标准分词模型的主要构成部分。基于词特征的分词方法主要通过半条件随机场（semi-Markov Conditional Random Fields，semi-CRF）来实现，与 CRF 不同，semi-CRF 通过将其特征函数定义在序列数据的某个段落上来捕捉词特征信息。Andrew[11]第一次将半条件随机场semi-CRF引入分词任务中，但是其性能却不甚理想。虽然后续仍有学者对其进行改善，但是考虑到 semi-CRF 的训练相比标准的 CRF 更耗时，并且也没有明显的优势，因此基于词特征的分词在后续的分词研究中并未成为主流。

1.2 基于深度学习的神经网络分词模型

随着深度学习在自然语言处理中不断取得突破，越来越多的研究人员也开始探索将其用于中文分词任务中。Zheng 等人[12]在 2013 年首次使用深度学习技术解决中文分词任务，验证了深度学习模型应用到中文分词任务上的可行性。Chen 等人[13]基于长短期记忆网络（Long Short-Term Memory，LSTM）模型特有的记忆性，将其引入分词模型中来解决中文分词问题，充分发挥了该模型在长序列任务处理上的优势，获得了较好的分词效果。2016 年，Yao 等人[14]提出了使用 Bi-LSTM 模型完成中文分词任务，获得了非常不错的效果。

1.3 分词任务的标注方法

分词任务的标注需要体现汉字在一个词中具体的位置信息，针对中文分词任务，目前主流的标注法主要有 2 词位标注法、4 词位标注法和 6 词位标注法。这 3 种方法的定义如表 1 所示。

表 1 中文分词标注方法

标注方法	标　记	定　义
2 词位	S,N	开始，后续
4 词位	S,M,E,S	开始，中间，结束，单字
6 词位	B,B_2,B_3,M,E,S	开始，第一中间，第二中间，其他中间，结束，单字

1.4 跨领域分词

随着计算机技术以及深度学习模型的发展，中文分词任务在准确率上有了明显的提升，但是中文分词任务面临着新的问题：未登录词的影响仍然存在，专业名词的识别失误在跨领域分词中尤为明显[15]。2018 年，Ma 等人[16]探讨了在 Bi-LSTM 模型的基础上预训练字向量、Dropout 和调整超参数以达到最先进结果，实验证明最新的分词模型的错误大多由注释不一致问题或者词汇不足造成，并且

也几乎无法通过调整模型结构来降低犯错的概率，因此使用外部词典或知识库非常重要。Zhang 等人[17]提出在神经网络的基础上结合字典进行分词，该方法可以解决对应领域词汇稀少引起的问题，并且当模型需要应用于不同领域时仅需添加额外的领域词典，其他参数可保持不变。

2　本文方法

本文提出的分词模型结构如图 1 所示，由通用分词模型训练、模型测试和领域分词模型训练 3 个部分构成。通用分词模型训练部分和领域分词模型训练部分，统称为模型训练部分，分别使用两种数据集在相同的模型结构下进行训练，得到对应的通用分词模型和领域分词模型。模型测试部分，对领域测试集分别使用两种模型对测试集进行标签预测，然后对两种模型得到的标签概率按权重比例 λ 进行权重组合，得到组合后新的标签概率，最后经过标签预测层，得到最后分词的结果。

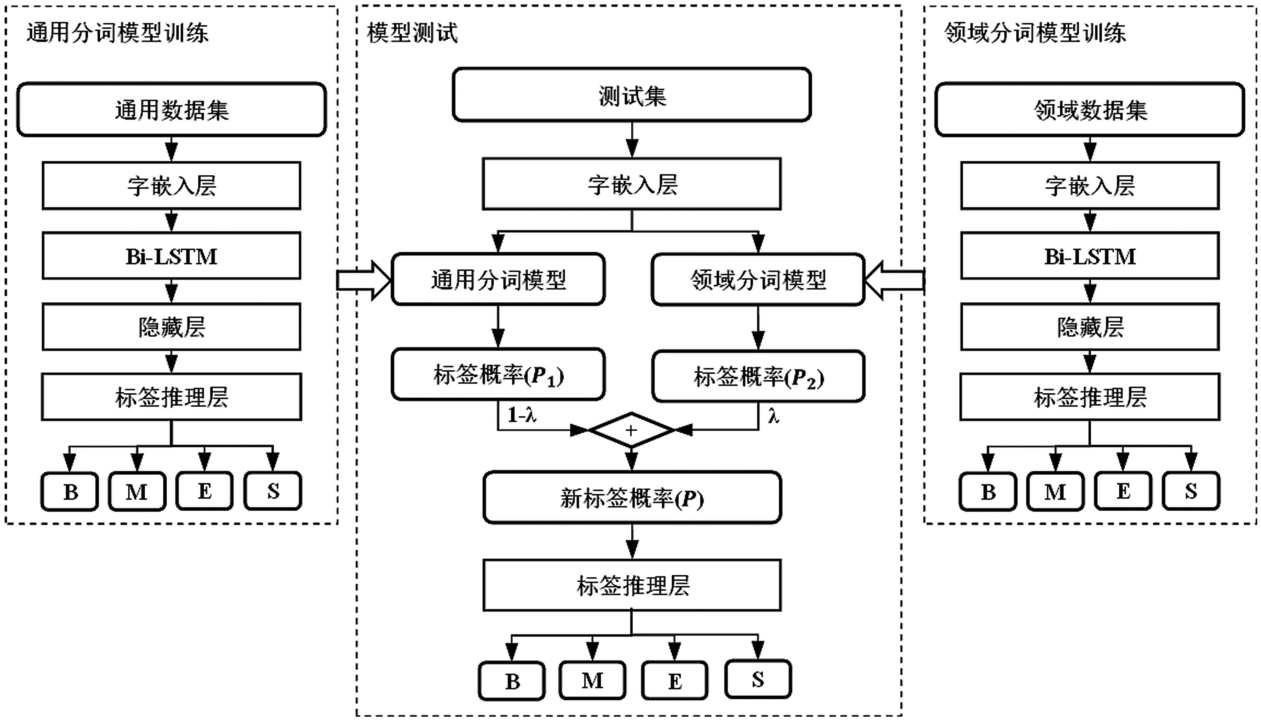

图 1　本文模型框架

2.1　模型训练部分

本方法中需要同时准备一个通用分词模型以及一个领域分词模型，两者使用相同结构的 Bi-LSTM 分词模型进行训练，区别主要在于通用分词模型由通用数据集训练而来，领域分词模型则由领域数据集训练而来。

2.1.1　基于 Bi-LSTM 的中文分词模型架构

基于 Bi-LSTM 的中文分词模型主要由字嵌入层、神经网络层和标签推理层 3 部分组成。具体架构如图 2 所示。

2.1.2　字嵌入

使用深度学习模型进行自然语言处理相关任务，首先需要将输入的文本转化为向量，因为只有经过文本向量化，输入的文本才能变成计算机能够识别和理解的信息。文本向量化的方法主要有独热表示(one-hot representation)和分布式表示(distribute representation) 两种。前者是用一个长向量来表示一个字或者词，向量的长度词典的大小相等，每个词对应的向量的分量只有一项为 1，其他均为 0，1 的位置对应该字在词典中的位置。这种表示方法生成的词向量维度太高且数据过于稀疏，并且不能很好地反映词与词之间的相似性。

分布式表示[18]是将每一个词映射成相同长度的短向量，在所有词对应向量形成的向量空间中，每一个向量即为该空间中的一个点，两个词在这个空间中的“距离”就体现出了它们之间的语义相似性。该表示方法通常又称字嵌入（Embedding）。目前已有充足的研究证明，在自然语言处理领域，对神经网络训练之前将文本进行字嵌入操作，可以明显地提升训练的效果。

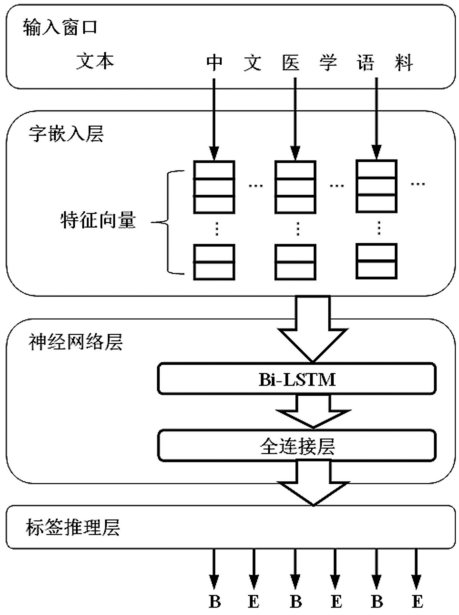

图 2 基于 Bi-LSTM 的中文分词架构

2.1.3 Bi-LSTM

循环神经网络（Recurrent Neural Network，RNN）是一种特殊的神经网络，其结构特点就是其神经元会在时序上传递信息，因此 RNN 在序列数据的处理上有比较大的优势。由 Hochreiter 等人 [19] 首先提出的 LSTM 是一种改进的 RNN 模型，通过其特有的结构，解决了传统 RNN 在较长序列下容易出现的梯度消失和梯度爆炸等问题。LSTM 与传统 RNN 的区别主要在于神经元内部结构不同，其神经元结构如图 3 所示。

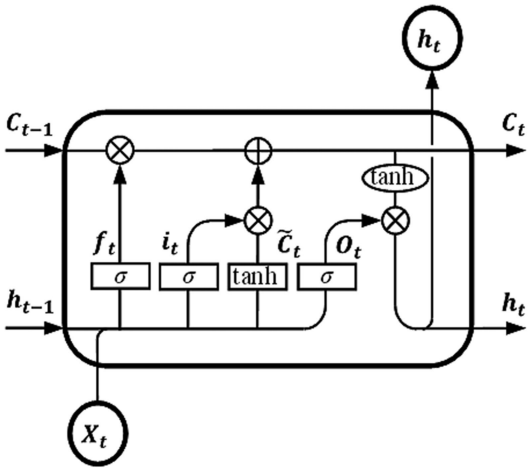

图 3 LSTM 神经元结构

（1）细胞状态

图 3 中穿过神经元顶部的水平线即代表细胞状态，细胞状态穿过了 LSTM 链上的所有神经元，并且只有一些小的线性操作作用其上，因此整个链的信息在传递过程中可以保持相对稳定。

（2）遗忘门

遗忘门决定细胞状态在该神经元中丢弃的信息。遗忘门的输入为上一时刻的隐层状态 $\boldsymbol{h}_{t-1}$ 和当前时刻的输入 $\boldsymbol{x}_t$，遗忘门的输出用于决定上一时刻的隐层状态中需要被遗忘的信息。遗忘门门控状态 $\boldsymbol{f}_t$ 的表达式为：

$$\boldsymbol{f}_t=\sigma(\boldsymbol{W}_f\cdot[\boldsymbol{h}_{t-1},\boldsymbol{x}_t]+\boldsymbol{b}_f) \quad (1)$$

式中：σ 为 Sigmoid 激活函数；$\boldsymbol{W}_f$ 为 $\boldsymbol{f}_t$ 的权重矩阵；$\boldsymbol{b}_f$ 为 $\boldsymbol{f}_t$ 的偏置。

（3）输入门

在本时刻需要在细胞状态中存入的信息是由输入门决定的，其包括输入门的门控状态 $\boldsymbol{i}_t$ 和候选值向量 $\tilde{\boldsymbol{C}}_t$，具体的表达式为：

$$\boldsymbol{i}_t=\sigma(\boldsymbol{W}_i\cdot[\boldsymbol{h}_{t-1},\boldsymbol{x}_t]+\boldsymbol{b}_i) \quad (2)$$

$$\tilde{\boldsymbol{C}}_t=\tanh(\boldsymbol{W}_C\cdot[\boldsymbol{h}_{t-1},\boldsymbol{x}_t]+\boldsymbol{b}_C) \quad (3)$$

式中：$\boldsymbol{W}_i$ 为 $\boldsymbol{i}_t$ 的权重矩阵；$\boldsymbol{b}_i$ 为 $\boldsymbol{i}_t$ 的偏置；$\boldsymbol{W}_C$ 为 $\tilde{\boldsymbol{C}}_t$ 的权重矩阵；$\boldsymbol{b}_C$ 为 $\tilde{\boldsymbol{C}}_t$ 的偏置。

在信息经过输入门后，细胞状态在本时刻的神经元中也随之更新完毕：

$$\boldsymbol{C}_t=\boldsymbol{f}_t\cdot\boldsymbol{C}_{t-1}+\boldsymbol{i}_t\cdot\tilde{\boldsymbol{C}}_t \quad (4)$$

式中：$\boldsymbol{C}_t$ 为本时刻的细胞状态；$\boldsymbol{C}_{t-1}$ 为上一时刻的细胞状态。

（4）输出门

输出门决定细胞的输出。输出门的输入也是上一时刻的隐层状态 $\boldsymbol{h}_{t-1}$ 和当前时刻的输入 $\boldsymbol{x}_t$，在通过 Sigmoid 层后得到输出门的门控状态 $\boldsymbol{O}_t$，$\boldsymbol{O}_t$ 与经过 tanh 函数处理更新后的细胞状态 $\boldsymbol{C}_t$ 相乘，结果即为本时刻神经元的输出，同时该输出又会作为隐层状态输入下一时刻的神经元中，具体的表达式为：

$$\boldsymbol{O}_t=\sigma(\boldsymbol{W}_O\cdot[\boldsymbol{h}_{t-1},\boldsymbol{x}_t]+\boldsymbol{b}_O) \quad (5)$$

$$\boldsymbol{h}_t=\boldsymbol{O}_t\cdot\tanh(\boldsymbol{C}_t) \quad (6)$$

式中：$\boldsymbol{W}_O$ 为 $\boldsymbol{O}_t$ 的权重矩阵；$\boldsymbol{b}_O$ 为 $\boldsymbol{O}_t$ 的偏置；$\boldsymbol{h}_t$ 为隐层状态。

Bi-LSTM 神经网络 [20] 由两个结构相同、传播方向相反的 LSTM 网络组成，这两个网络分别称为前向层和后向层。与 LSTM 相比，Bi-LSTM 在继承其诸多优点的同时，还能同时在正方向和反方向上捕捉序列的信息，其具体结构如图 4 所示。图 4 中 $\boldsymbol{i}_b$、$\boldsymbol{o}_b$、$\boldsymbol{i}_f$、$\boldsymbol{o}_f$ 分别代表后向层和前向层的输入和输出，$\boldsymbol{w}_b$ 和 $\boldsymbol{w}_f$ 分别代表后向层和前向层神经元之间传递的信息，即隐藏状态和细胞状态。

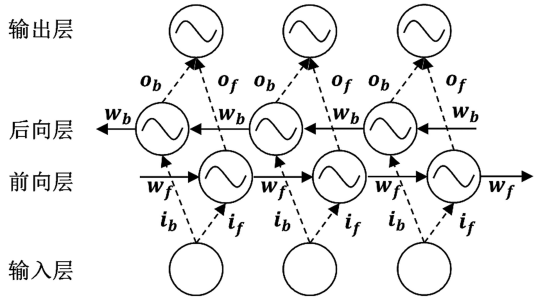

图 4 Bi-LSTM 结构

Bi-LSTM 的最终输出由前向层和后向层的对应节点共同决定，其公式为：

$$\boldsymbol{h}_{ft}=\boldsymbol{O}_{ft}\cdot\tanh(\boldsymbol{f}_{ft}\cdot\boldsymbol{C}_{ft-1}+\boldsymbol{i}_{ft}\cdot\tilde{\boldsymbol{C}}_{ft}) \quad (7)$$

$$\boldsymbol{h}_{bt}=\boldsymbol{O}_{bt}\cdot\tanh(\boldsymbol{f}_{bt}\cdot\boldsymbol{C}_{bt-1}+\boldsymbol{i}_{bt}\cdot\tilde{\boldsymbol{C}}_{bt}) \quad (8)$$

$$\boldsymbol{y}_t=[\boldsymbol{h}_{ft},\boldsymbol{h}_{bt}] \quad (9)$$

式中：$\boldsymbol{h}_{ft}$ 和 $\boldsymbol{h}_{bt}$ 分别为前向层和后向层的输出；$\boldsymbol{y}_t$ 为 Bi-LSTM 的最终输出；$\boldsymbol{f}_{ft}$、$\boldsymbol{i}_{ft}$、$\boldsymbol{O}_{ft}$，$\boldsymbol{f}_{bt}$、$\boldsymbol{i}_{bt}$、$\boldsymbol{O}_{bt}$ 分别为前向层和后向层遗忘层、输入层、输出层的门控状态；$\tilde{\boldsymbol{C}}_{ft}$、$\tilde{\boldsymbol{C}}_{bt}$ 分别为前向层和后向层的候选值向量；$\boldsymbol{C}_{ft-1}$、$\boldsymbol{C}_{bt-1}$ 分别为前向层和后向层上一时刻的细胞状态。

2.2　模型测试部分

2.2.1　标签概率权重组合

将测试集进行字嵌入操作后分别输入至通用分词模型和领域分词模型中，输出的结果即为两种模型对分词结果对应的标签的概率预测。对两种模型得到的标签概率进行权重组合，得到组合后新的标签概率。

$$P_i=\lambda\cdot P_{1i}+(1-\lambda)\cdot P_{2i} \quad (10)$$

式中：P_{1i} 为使用领域数据训练模型预测该汉字标签为 i 的概率；P_{2i} 为通用数据模型预测该汉字标签为 i 的标签概率；λ 为领域数据模型预测的权重比。

2.2.2　Viterbi 算法

维特比算法（Viterbi Algorithm）是动态规划算法中的一种，用于快速求解最佳路径的问题。笔者在标签预测层引入维特比算法对汉字对应的标签进行预测[21]。对于权重组合后的标签概率，笔者要寻找一条隐含状态序列，用它去生成指定的观测序列，使得这个观测序列的概率最大。计算最大路径的概率公式为：

$$\delta_t(i)=\max P(i_t=i,i_{t-1},\cdots,i_1,o_t,\cdots,o_1),\quad i=1,2,3,\cdots,N \quad (11)$$

$\delta_t(i)$ 表示在时刻 t，结束于隐藏状态 i，同时满足观测序列 $(o_1,o_2,o_3,\cdots,o_t)$ 的最大路径概率。其递推公式为：

$$\delta_t(i)=\max(\delta_{t-1}(j)a_{ij}b_{io_t}),1\leqslant j\leqslant N \quad (12)$$

式中：a_{ij} 为状态 i 到 j 的转移概率；b_{io_t} 为状态 o_t 的观察概率。

笔者按照维特比算法中的思路从 $\delta_1(i)$ 开始起步，一步一步推导至 $\delta_t(i)$，求得最大的概率，同时在递推的过程中，在每一个时间点 t 都记录好上一个时间点 $t-1$ 的隐含状态，即可获得最优的隐含状态序列。

笔者在模型测试的标签预测层利用维特比算法，对各汉字的标签进行计算得到各汉字所属标签的最终概率，将最大概率的标签作为预测标签，从而完成中文分词。

3　实验及分析

3.1　实验数据

本次实验中通用数据集来自 SIGHAN 第二届国际中文分词大赛（SIGHAN The Second International Chinese Word Segmentation Bakeoff）的微软研究院（Microsoft Research，MSR）语料。中文医学领域数据（Chinese Medical Corpus，CMC）语料来自“‘万创杯’中医药天池大数据竞赛——中药说明书实体识别挑战”所提供的训练集，其中包括 1 000 条药物说明书，以及对应的实体标注。将数据集中的实体内容提取作为领域词典，使用 jieba 分词导入自定义词典的功能导入领域词典后对其进行分词。最后通过人工对分词结果进行检查，得到一个标准的中文医学语料。随机抽取该语料中的 70% 作为领域分词模型的训练集，剩下的作为测试集。数据详情如表 2 所示。

表 2　实验数据统计

数据集	训练集	测试集
MSR	86 924	–
CMC	700	300

所有的数据集的标注方法均为 4 词位标注法。表 3 为两种数据集中的标注示例。

3.2　实验设置

本文实验软件环境为 python 3.6，模型基于 pytorch1.8.0 深度学习框架和 numpy、pandas、sklearn 等第三方库。使用 torch.nn.Embedding 模块对数据集进行字嵌入操作。其他参数和设置如表 4 所示。

表 3 4 词位标注样例

数据集	样 例
MSR	人/b 们/e 常/s 说/s 生/b 活/e 是/s 一/s 部/s 教/b 科/m 书/e ，/s 而/s 血/s 与/s 火/s 的/s 战/b 争/e 更/s 是/s 不/b 可/m 多/m 得/e 的/s 教/b 科/m 书/e ，/s 她/s 确/b 实/e 是/s 名/b 副/m 其/m 实/e 的/s ‘/s 我/s 的/s 大/b 学/e ’/s 。/s
CMC	如/s 与/s 其/b 他/e 药/b 物/e 同/b 时/e 使/b 用/e 可/b 能/e 会/s 发/b 生/e 药/b 物/m 相/m 互/m 作/m 用/e ，/s 详/b 情/e 请/s 咨/b 询/e 医/b 师/e 或/s 药/b 师/e 。/s

表 4 该处填入表的名称

参 数	数 值
最大长度（max length）	70
Bi-LSTM 层数	1
批量大小（batch size）	64
隐藏层维度（hidden dimension）	64
学习率（learning rate）	1e-5
字嵌入维度（embedding dimension）	128
优化器（optimizer）	Adam
训练轮数（epoch）	30

3.3 评价指标

本实验使用 $F1$ 值作为主要评价指标，同时使用精确率（Precision rate，P）、召回率（Recall rate，R）作为参考，其计算方式为：

$$P=\frac{A}{B} \quad (13)$$

$$R=\frac{A}{C} \quad (14)$$

$$F1=\frac{2PR}{P+R} \quad (15)$$

式中：A 表示正确分词的单词的数量；B 表示分词总数；C 表示标准分词集中的词数。

3.4 实验设计

3.4.1 模型预训练

本实验涉及两个预训练模型，分别使用两种数据集在相同结构的 Bi_LSTM 模型下训练获得。

（1）通用分词模型：使用 MSR 数据集作为训练集获得的 MSR_Bi_LSTM 模型。

（2）领域分词模型：使用 CMC 数据集的训练集获得的 CMC_Bi_LSTM 模型。

3.4.2 权重比例的选择

权重比例 λ 用于调节 MSR_Bi_LSTM 和 CMC_Bi_LSTM 的预测结果之间的权重。图 5 展示了选择不同 λ 得到的对应分词实验的 $F1$ 值。虽然在不同的迭代次数下得到最好实验结果的 λ 值有所不同，但是在迭代 30 轮的训练模型下，当 λ=0.8 时取得了最好的效果。表 5 为迭代 30 轮的分词模型在不同 λ 下的预测结果，此词的准确率、召回率和 $F1$ 值，分别为 95.49%，95.44% 和 95.46%。因此将 λ 固定为 0.8 来进行后续的实验。

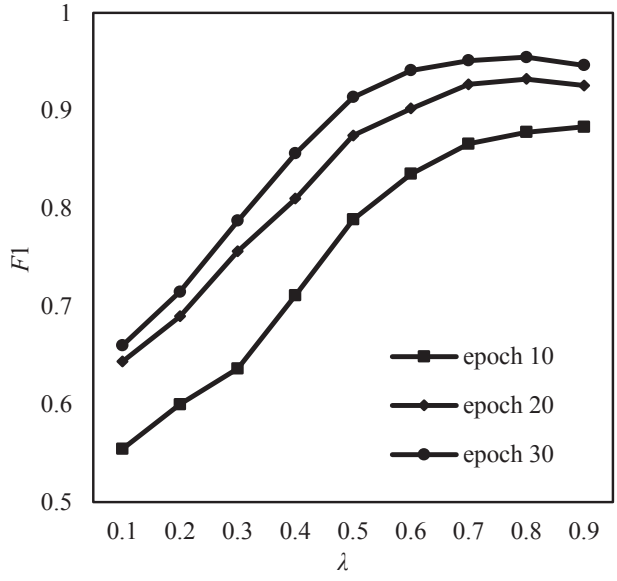

图 5 不同权重比在不同迭代数下的 $F1$ 值

表 5 不同权重比对应的实验结果

λ	P/%	R/%	$F1$/%
0.1	72.94	67.45	66.00
0.2	77.34	72.38	71.49
0.3	82.78	78.99	78.75
0.4	87.89	85.41	85.62
0.5	92.19	91.09	91.38
0.6	94.43	93.91	94.10
0.7	95.26	94.99	95.11
0.8	**95.49**	**95.44**	**95.46**
0.9	94.71	94.53	94.62

3.4.3 标签转移概率矩阵的确定

本实验中标签推理层使用维特比算法进行标签预测，需要对标签转移概率进行确定。对于本文所选的标注方法，其标签转移的可能性仅存在 SS、SB，BM、BE，MM、ME，ES、EB 几种情况。在上文实验中均采用等概率形式的标签转移概率进行实

验，即所有转移概率设为 0.5。然而等概率形式的标签转移概率并不一定能使实验效果最好。

在本实验中，通过采用固定其他 3 个标签的转移概率，变换当前标签的转移概率进行实验来确定当前标签的最佳标签转移概率。表 6 和表 7 为最后得到的最优标签转移概率矩阵以及对应的实验结果。可以看出，使用最优转移概率所取得的实验结果要优于使用等概率的实验结果。

表 6　最优标签转移概率矩阵

	S	B	M	E
S	0.3	0.7	–	–
B	–	–	0.7	0.3
M	–	–	0.6	0.4
E	0.4	0.6	–	–

表 7　最优标签转移概率实验结果

矩阵	*P*	*R*	*F*1
最优转移概率	95.54	95.55	95.54
等概率	95.49	95.44	95.46

3.5　对比实验

为验证本文所提模型的有效性，本文分别选择 MSR 数据集、CMC 数据集以及将两者合并的 MSR&CMC 数据集使用 BiLSTM-CRF 模型进行实验作为对比。在确保模型参数一致的情况下，记录的实验结果如表 8 所示。

根据表 8 的实验结果可知，分词模型 MSR_BiLSTM-CRF 在跨领域的分词任务上效果并不好，并且对比 CMC_BiLSTM-CRF 与 MSR&CMC_BiLSTM-CRF 的实验结果可知，在医学领域的分词任务中，通过并入其他领域的训练数据对训练集进行扩充，可略微提升最后的分词效果。本文所提出的模型相比以上的分词模型在准确率、召回率和 *F*1 值上均获得了最优的结果。

表 8　实验结果对比　　%

实验方法	*P*	*R*	*F*1
MSR_BiLSTM-CRF	68.97	65.56	64.45
CMC_BiLSTM-CRF	92.69	92.01	92.26
MSR&CMC_BiLSTM-CRF	92.96	92.39	92.61
本文模型	**95.54**	**95.55**	**95.54**

表 9 列举了测试集中 3 个例句在对比实验中的结果。从分词结果上可看出，传统的跨领域的分词模型 MSR_BiLSTM-CRF 在中文医学领域上表现很差，对于未登录词没办法生成正确的分词结果，例如例 1 中将“失血性休克”切分为“失 / 血性 / 休克”。

表 9　实验结果示例

实验方法	示例	分词结果
MSR_BiLSTM-CRF	例 1	能 / 显著 / 性 / 地 / 升高 / 白 / 细胞 / 的 / 数目 /、/ 升高 / 血压 /，/ 防止 / 失 / 血性 / 休克 /。
	例 2	本品 / 对 / 角 / 叉 / 菜 / 胶 / 引起 / 的 / 大 / 鼠 / 无 / 菌性 / 前列腺炎 / 和 / 大肠杆菌 / 引起 / 的 / 大 / 鼠 / 细菌性 / 前列腺炎 / 均 / 有 / 减轻 / 肿胀 /，/ 抑制 / 炎症 / 指标 / 作用 /。
	例 3	可 / 抑 / 制氢 / 化 / 可 / 的 / 松 / 所致 / 小鼠 / 胸腺 / 萎缩 / 及 / 脾脏 / 系数 / 下降 /；/ 抑制 / 大鼠 / 热刺激 / 和 / 小鼠 / 腹腔 / 注射 / 醋酸 / 等 / 致 / 痛 / 反应 /。
CMC_BiLSTM-CRF	例 1	能 / 显著性 / 地 / 升高 / 白细胞 / 的 / 数目 /、/ 升 / 高血压 /，/ 防止 / 失血性休克 /。
	例 2	本品 / 对 / 角叉菜胶 / 引起 / 的 / 大鼠 / 无菌性前列腺炎 / 和 / 大肠杆菌 / 引起 / 的 / 大鼠 / 细菌性前列腺炎 / 均 / 有 / 减轻肿胀 /，/ 抑制 / 炎症指标 / 作用。
	例 3	可 / 抑制 / 氢化可的松 / 所致 / 小鼠 / 胸腺萎缩 / 及 / 脾脏系数下降 /；/ 抑制 / 大鼠 / 热刺激 / 和 / 小鼠 / 腹腔注射醋酸 / 等 / 致痛反应 /。
MSR&CMC_BiLSTM-CRF	例 1	能 / 显著 / 性地升 / 高白细胞 / 的 / 数目 /、/ 升高 / 血压 /，/ 防止 / 失血性休克 /。
	例 2	本品 / 对 / 角叉菜胶 / 引起 / 的 / 大鼠 / 无菌性前列腺炎 / 和 / 大肠杆菌 / 引起 / 的 / 大鼠 / 细菌性前列腺炎 / 均 / 有 / 减轻肿胀 /，/ 抑制炎症 / 指标 / 作用 /。
	例 3	可 / 抑 / 制氢 / 化可的松 / 所致 / 小鼠 / 胸腺萎缩 / 及 / 脾脏系数下降 /；/ 抑制 / 大鼠 / 热 / 刺激 / 和 / 小鼠 / 腹腔 / 注射 / 醋酸 / 等 / 致痛 / 反应 /。
本文模型	例 1	能 / 显著性 / 地 / 升高 / 白细胞 / 的 / 数目 /、/ 升高 / 血压 /，/ 防止 / 失血性休克 /。
	例 2	本品 / 对 / 角叉菜胶 / 引起 / 的 / 大鼠 / 无菌性前列腺炎 / 和 / 大肠杆菌 / 引起 / 的 / 大鼠 / 细菌性前列腺炎 / 均 / 有 / 减轻肿胀 /，/ 抑制 / 炎症指标 / 作用 /。
	例 3	可 / 抑制 / 氢化可的松 / 所致 / 小鼠 / 胸腺萎缩 / 及 / 脾脏系数下降 /；/ 抑制 / 大鼠 / 热刺激 / 和 / 小鼠 / 腹腔注射醋酸 / 等 / 致痛反应 /。

使用中文医学领域训练集训练所得的分词模型CMC_BiLSTM-CRF，能够更加正确地切分出该领域的专业词汇，但是由于训练集样本较小，在测试时还是会出现如例 2 中将“升高血压”切分为“升 / 高血压”的情况。将两种数据集合并后训练得出的MSR&CMC_BiLSTM-CRF 模型也获得了不错的切分结果，但是由于两个数据集数据量的悬殊，在出现歧义的句子中的分词结果会偏向于数据量较大的数据集中的样本，如例 3 中的“可抑制氢化可的松……”被切分为了“可抑 / 制氢 / 化可的松 /……”，本文模型由于能够通过权值 λ 来控制两个数据集所训练的模型对预测结果的影响，所以能有效地避免该问题。

综上所述，在小样本的中文医学领域的分词任务上，利用本文提出的分词模型，通过引入跨领域的通用分词模型，能够有效地弥补因为训练样本不足而导致分词结果较差的问题。在调整标签转移概率以及权值等参数后获得的最优结果，相比传统的BiLSTM-CRF 分词模型，在正确率、召回率和 $F1$ 值上均有显著提升。

4 结 论

针对医学领域的中文分词问题，本文提出了一种分别将通用数据集以及医学领域数据集进行模型训练，再将训练所得模型结合形成的全新分词模型。实验结果证明，该模型能充分发挥两种模型在各自领域的优势，提升分词的效果。在仅有少量医学领域标注数据的前提下，该模型相比传统的分词方法在分词效果上有比较显著的提升。

由于条件所限，本文并未将工作重心放在模型选择上，在实验的参数设置、最优转移概率的选择上也比较粗糙。在下一步的研究中，笔者计划引入预训练的词向量，以门控循环单元（Gated Recurrent Unit，GRU）、有序神经元长短时记忆神经网络（Ordered Neurons Long Short-Term Memory，ONLSTM）等 LSTM 变种单元，对模型进行优化，同时针对实验中的参数的优化做进一步的探索。

参考文献：

[1] 丁泽源，杨志豪，罗凌，等 . 基于深度学习的中文生物医学实体关系抽取系统 [J]. 中文信息学报，2021,35(5):70-76.

[2] 陈德华，冯洁莹，乐嘉锦，等 . 中文病理文本的结构化处理方法研究 [J]. 计算机科学，2016,43(10):272-276.

[3] 黄昌宁 . 中文信息处理中的分词问题 [J]. 语言文字应用，1997(1):74-80.

[4] WANG K,ZONG C,SU K Y. A character-based joint model for Chinese word segmentation[C]//23rd International Conference on Computational Linguistics,2010:1173-1181.

[6] XUE N,CONVERSE S P.Combining classifiers for Chinese word segmentation[C]//The First SIGHAN Workshop on Chinese Language Processing,2002:1-7.

[5] 张梅山，邓知龙，车万翔，等 . 统计与词典相结合的领域自适应中文分词 [J]. 中文信息学报，2012,26(2):8-12.

[7] 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.

[8] XUE N W,SHEN L B.Chinese word segmentation as LMR tagging[C]//SIGHAN Workshop on Chinese Language Processing,2003:176-179.

[9] LOW J K,NG H T,GUO W.A maximum entropy approach to Chinese word segmentation[C]//the Fourth SIGHAN Workshop on Chinese Language Processing,2005:161-164.

[10] PENG F C,FENG F F,MCCALLUM A.Chinese segmentation and new word detection using conditional random fields[C]//International Conference on Computational Linguistics - COLING‘04,2004:562-568.

[11] ANDREW G.A hybrid Markov/Semi-Markov conditional random field for sequence segmentation[C]//Conference on Empirical Methods in Natural Language Processing - EMNLP ‘06,2006:465-472.

[12] ZHENG X Q,CHEN H Y,XU T Y.Deep learning for Chinese word segmentation and pos tagging[C]//Conference on Empirical Methods in Natural Language Processing,2013:647-657.

[13] CHEN X C,QIU X P,ZHU C X,et al.Long short-term memory neural networks for Chinese word segmentation[C]//Conference on Empirical Methods in Natural Language Processing,2015:1197-1206.

[14] YAO Y S,HUANG Z.Bi-Directional LSTM recurrent neural network for Chinese word segmentation[C]//Neural Information Processing,2016:345-353.

[15] 王佳楠，梁永全 . 中文分词研究综述 [J]. 软件导刊，2021,20(4):247-252.

[16] MA J,GANCHEV K,WEISS D.State-of-the-Art Chinese word segmentation with Bi-LSTMS[C]//

Conference on Empirical Methods in Natural Language Processing,2018:4902-4908.

[17] ZHANG Q,LIU X,FU J.Neural networks incorporating dictionaries for Chinese word segmentation[C]//AAAI Conference on Artificial Intelligence,2018.

[18] HINTON G E.Learning distributed representations of concepts[C]//the Eighth Annual Conference of the Cognitive Science Society,1986:12.

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

[20] SCHUSTER M,PALIWAL K K.Bidirectional recurrent neural networks[J].IEEE Transactions on Signal Processing,1997,45(11):2673-2681.

[21] 蒋卫丽，陈振华，邵党国，等．基于领域词典的动态规划分词算法 [J]. 南京理工大学学报，2019,43(1):63-71.

作者简介：

邵党国（1979—），男，博士，副教授，主要研究方向为图像处理、文本处理、机器学习和数据挖掘等；

黄初升（1992—），男，硕士研究生，主要研究方向为自然语言处理；

马　磊（1978—），通讯作者，男，硕士，高级工程师，主要研究方向为信息技术、生物医学工程；

贺建峰（1965—），男，博士，教授，主要研究方向为数字图像处理；

易三莉（1977—），女，博士，讲师，主要研究方向为数字图像处理。