文章编号：1003-0077(2019)09-0017-07

基于联合学习的跨领域法律文书中文分词方法

江明奇，严 倩，李寿山

（苏州大学 计算机科学与技术学院，江苏 苏州 215006）

摘 要：中文分词任务是自然语言处理的一项基本任务。但基于统计的中文分词方法需要大规模的训练样本，且拥有较差的领域适应性。然而，法律文书涉及众多领域，对大量的语料进行标注需要耗费大量的人力、物力。针对该问题，该文提出了一种基于联合学习的跨领域中文分词方法，该方法通过联合学习将大量的源领域样本辅助目标领域的分词，从而提升分词性能。实验结果表明，在目标领域标注样本较少的条件下，该文方法的中文分词性能明显优于传统方法。

关键词：中文分词；法律文书；联合学习

中图分类号：TP391 **文献标识码**：A

Cross-domain Chinese Word Segmentation for Legal Documents with Joint Learning

JIANG Mingqi, YAN Qian, LI Shoushan

(School of Computer Science and Technology, Soochow University, Suzhou, Jiangsu 215006, China)

Abstract: To deal with legal documents involving multi-domain texts, this paper proposes a cross-domain approach on Chinese word segmentation with joint learning. In the method, a large number of source domain samples are used to assist word segmentation in target domain through joint learning, which improves the performance of word segmentation. Experimental results demonstrate that, even with a few annotation samples from target domain, the performance of proposed method is obviously better than that of the traditional method.

Keywords: Chinese word segmentation; legal documents; joint learning

0 引言

中文分词作为中文信息处理的基础任务，其准确性直接影响其它中文信息处理任务的性能[1]。基于机器学习的方法在中文分词领域上有优异的结果。例如，最大熵(maximum entropy)模型[2]、条件随机场(conditional random field，CRF)模型[3-5]以及长短期记忆(long short-term memory，LSTM)神经网络[6]。然而，传统的方法需要大规模的分词语料以训练性能优异的分词器，分词语料的获得需要大量人工参与，所耗费的成本太高。因此，传统的方法在法律文书上不能取得较好的中文分词性能[7]。

由于法律文书中各领域的语料匮乏，学者们使用跨领域的方法进行分词性能的提升。然而，不同领域的样本有一定的差异性，因此在跨领域任务上难以直接使用不同领域的样本提升分词性能。其主要原因在于各领域的词语分布不同，当使用源领域的分词器对目标领域进行分词时，未登录词(out of vocabulary，OOV)的数目快速增加，因此该分词器在目标领域上进行中文分词时无法获得较好的性能。除此之外，法律文书中的专有名词的构词规则和通用领域不同，同一个字符在不同领域中具有不同的标签分布。例如，在法律文书中，“一审”和“二审”为常用词。其中，“审”为词尾，其标签为“E”。但该字在通用领域中情况有所不同，“审”通常以词首的形式出现，如“审稿”，其标签为“B”。

鉴于获得法律文书的各个领域的少量已标注语料的难度较小，各领域拥有相同的标注规则，本文提出了一种基于联合学习的跨领域中文分词方法。该

收稿日期：2018-10-17 **定稿日期**：2019-02-22

基金项目：国家自然科学基金(61672366)

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

方法通过使用法律文书某一领域(源领域)数据辅助另一领域(目标领域)的方式可以提高相应的分词性能。此外,本文并没有采用直接混合源领域的样本和目标领域的样本的方法,而是采用联合学习的方法,在模型中存在主任务(目标领域的分词)和辅助任务(源领域的分词),辅助任务对主任务有一定的辅助作用。具体而言,首先,使用共享 LSTM 层得到两个任务的辅助表示,让主任务和辅助任务共同参与分词结果的评判;其次,将该辅助表示与主任务的表示进行融合;最后,利用 LSTM 对主任务的结果进行预测。本文创建了法律文书的分词语料库,并针对法律文书的特点把本文的方法应用到法律文书的分词当中。

本文的组织结构如下:第 1 节介绍本文相关的一些工作;第 2 节详细介绍基于联合学习的跨领域中文分词方法;第 3 节对中文分词结果进行分析;第 4 节对本文做出总结,并对下一步工作进行展望。

1 相关工作

1.1 中文分词

近些年来,随着神经网络的广泛应用,越来越多的研究人员把神经网络使用到分词系统中。

Zheng 等首次在中文分词任务中使用深度学习模型,同时提出了一种感知器方法用来加速训练过程[8]。Chen 等将改进的长短期记忆神经网络运用到中文分词任务中,该模型是一种可以学习到长期依赖关系的循环神经网络(recurrent neural network,RNN)[9]。实验结果表明,该模型与传统模型在中文分词任务上性能相当。Yao 等提出使用双向 LSTM 层进行中文分词任务,并在实验中对比了含有不同数目的双向 LSTM 层的实验结果[10]。Xu 和 Sun 提出了基于长短期记忆神经网络的中文分词方法,在其中采集局部特征,并通过门控递归神经网络生成具有长距离依赖性的局部特征[11]。金等提出了一种双向 LSTM 模型,并把它用在了分词任务上[12]。Kamper 等通过声音的词的嵌入进行无监督的中文分词任务[13]。

1.2 跨领域的中文分词

邓等将基于通用领域标注文本的有指导训练和基于目标领域无标记文本的无指导训练相结合,即在全监督 CRF 中加入最小熵正则化框架,提出了半监督的 CRF 模型,提高了中文分词上的 F_1 值[14]。佟等提出了一种称为上下文变量(context variables)的数据来衡量某个候选词在篇章内的上下文信息,并使用语义资源,用其同义词的节点代价作为自己的代价,提高了未登录词的召回率[15]。许等针对目标领域分词语料的匮乏问题,提出主动学习(active learning)算法与 N-gram 统计特征相结合的领域自适应方法,用主动学习算法训练的分词系统各项指标上均有提高[16]。

2 基于联合学习的跨领域中文分词方法

2.1 基于 LSTM 模型的中文分词方法

RNN 模型是 Rumelhart 等提出的具有循环结构的网络结构[17]。为了解决 RNN 中的长期依赖问题,Hochreiter 和 Schmidhuber[18] 提出了一种新网络,称为长短期记忆(long short-term memory,LSTM)神经网络,该网络适用于处理和序列中长间隔的事件。LSTM 神经网络在每个时间步包含一个更新历史信息的神经单元。Graves[19] 对 LSTM 模型进行了相应的优化,此后 LSTM 广泛应用于语音等领域。

如图 1 所示,LSTM 单元设置了记忆单元 c_t 用于保存历史信息。LSTM 的 t 时刻的神经单元计算如式(1)~式(6)所示。

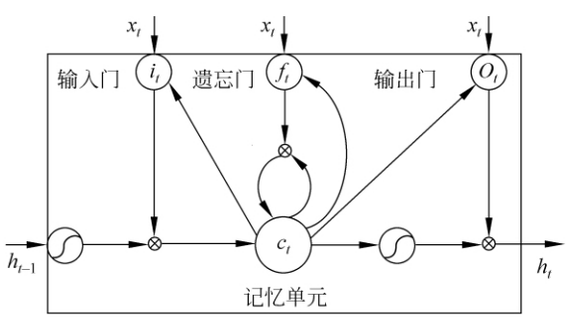

图 1 LSTM 单元

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + V_i C_{t-1}) \quad (1)$$

$$\tilde{c}_t = \tanh(W_c x_t + U_c h_{t-1}) \quad (2)$$

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + V_f c_{t-1}) \quad (3)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \quad (4)$$

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + V_o c_t) \quad (5)$$

$$h_t = o_t \odot \tanh(c_t) \quad (6)$$

其中,σ 为 logistic sigmoid 函数,i_t,f_t,o_t,c_t 分

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

别为输入门、输出门、遗忘门和记忆单元在 t 时刻的值。$\tilde{c}_t$ 为记忆单元的候选记忆状态值，h_t 为 t 时刻 LSTM 单元的输出。图 2 所示的是基于 LSTM 模型的法律文书中文分词方法的基本框架，分为 LSTM 层、全连接层、Dropout 层和 Softmax 层。下面将分别对各层进行介绍。

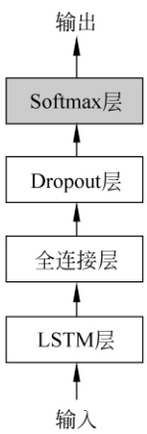

图 2　基于 LSTM 模型的法律文书中文分词方法的基本框架

2.1.1　LSTM 层

我们将所有关于当前字符的特征拼接得到的向量作为 LSTM 的输入。例如，若当前字符为 ch_i，我们将前一个字符 ch_{i-1} 、当前字符 ch_i 以及 ch_{i+1} 的一元和二元字的信息进行拼接，形成输入向量。输入向量通过 LSTM 层得到隐层向量。

2.1.2　全连接层

全连接层用于接收 LSTM 层的输出，我们在全连接层中加入激活函数，如式(7)所示。

$$h^* = \varphi(\theta^{\mathrm{T}} h + b) \tag{7}$$

其中，h 为 LSTM 层的输出，φ 为非线性激活函数，本文使用中“ReLU”作为全连接层的激活函数。

2.1.3　Dropout 层

Dropout 为当前深度学习中主流方法之一，能有效避免深度神经网络中的过拟合问题，提高基于深度学习模型的性能[20]，Dropout 层在训练时屏蔽某些神经元，从而避免过拟合问题，Dropout 操作的过程如式(8)所示。

$$g = h_i^* \cdot D(p) \tag{8}$$

其中，D 为 dropout 操作符，p 为可调超参(神经元以该超参的概率决定是否被屏蔽)。

2.1.4　Softmax 层

Softmax 层用于接收 Dropout 层的输出，并输出最终结果，如式(9)所示。

$$p = \mathrm{softmax}(W^d g + b^d) \tag{9}$$

其中，p 表示预测标签的概率集，W^d 为相应的权重向量，b^d 为偏置。

2.2　基于联合学习的跨领域中文分词方法

法律文书的特点在于其中存在许多专业词汇，各领域的差异性体现在各领域存在不同的专业词汇。然而，各领域含有共通的专业词汇和相似的专业词汇分词规则。对于未登录词，由于目标领域的已标注样本数目不足，不能较好地识别未登录词，因此我们使用将源领域的大量样本加入辅助任务的方法来提升未登录词识别的性能，该方法的提升体现在以下两点。(1)源领域可能含有目标领域中的某些未登录词(如“诉讼”“原告”)，辅助任务能够学习该未登录词中字对于分词任务的表示。在共享中间表示的情况下，这些表示能够帮助任务学习。(2)在辅助任务(源领域)和主任务(目标领域)都是未登录词的情况下，由于目标领域的训练样本规模有限，识别未登录词的性能有限，因此加入大量源领域的样本，可能提升识别未登录词的性能。例如，在各领域的法律文书中，某些字(如“为”)常常出现在词的开头位置(如“为由”)，加入大量样本有利于提升识别这些字开头的未登录词的能力。针对法律文书的上述问题，本文提出了一种基于联合学习的跨领域中文分词方法，有效利用源领域数据辅助目标领域。

图 3 给出了基于联合学习的跨领域中文分词模型的基本框架。其中，目标领域的预测为主任务，源领域的预测为辅助任务，使用辅助 LSTM 层得到两个任务的辅助表示，主任务利用该辅助表示进行预测。模型分为主任务、辅助任务以及联合学习，其介绍如下。

2.2.1　主任务

首先，使用主 LSTM 层(Main LSTM Layer)和辅助 LSTM 层(Auxiliary LSTM Layer)分别生成目标领域的隐层表示 h_{main1} 和 h_{main2}，如式(10)、式(11)所示。

$$h_{\mathrm{main1}} = \mathrm{LSTM}_{\mathrm{main}}(T_{\mathrm{target}}^{\mathrm{input}}) \tag{10}$$

$$h_{\mathrm{main2}} = \mathrm{LSTM}_{\mathrm{aux}}(T_{\mathrm{target}}^{\mathrm{input}}) \tag{11}$$

其次，通过全连接层接收 h_{main1} 和 h_{main2}，进行全链接操作，获得表示 h_{main1}^d 和 h_{main2}^d，如式(12～13)所示。

$$h_{\mathrm{main1}}^d = \mathrm{dense}_{\mathrm{main1}}(h_{\mathrm{main1}}) \tag{12}$$

$$h_{\mathrm{main2}}^d = \mathrm{dense}_{\mathrm{main2}}(h_{\mathrm{main2}}) \tag{13}$$

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

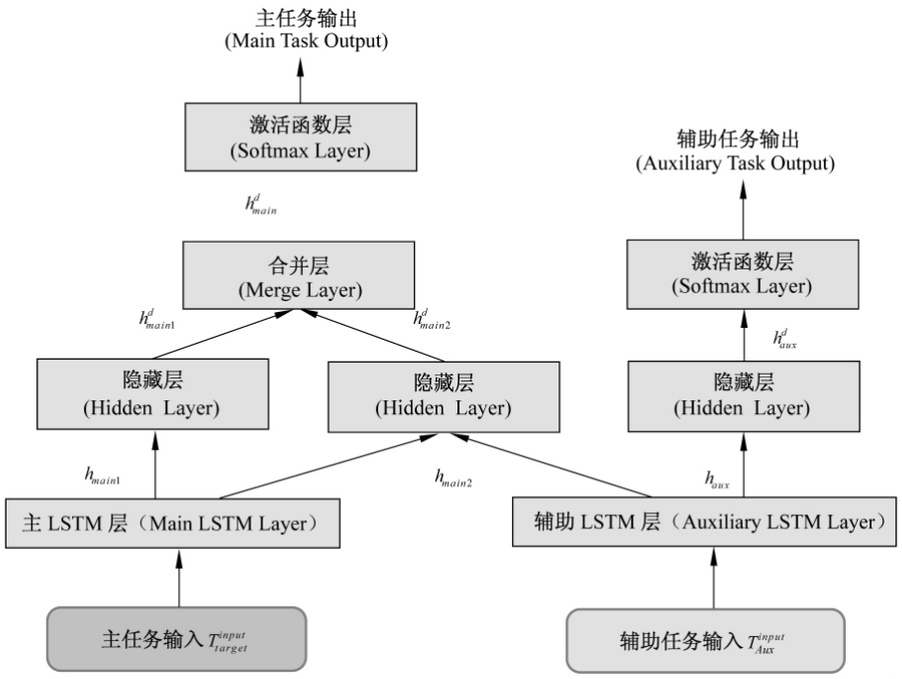

图 3 基于联合学习的跨领域中文分词方法框架

再次，我们将表示 h^d_{main1} 和 h^d_{main2} 进行拼接获得表示，并作为隐藏层（Hidden Layer）的输入，如式(14)所示。

$$h^d_{\text{main}} = \text{dense}_{\text{main}}(h^d_{\text{main1}} \oplus h^d_{\text{main2}}) \tag{14}$$

其中，h^d_{main} 为主任务中全连接层的输出，$\oplus$ 为拼接操作。

2.2.2 辅助任务

首先，源领域的数据通过辅助 LSTM 层获得相应的表示，辅助 LSTM 层连接了目标领域和源领域，使主任务和辅助任务对分词同时参与评判，辅助 LSTM 层对于不同领域的数据有相同的权重，如式(15)所示。

$$h_{\text{aux}} = \text{LSTM}_{\text{aux}}(T^{\text{input}}_{\text{source}}) \tag{15}$$

其中，h_{aux} 是共享的 LSTM 层对源领域进行编码生成的表示。

其次，h_{aux} 通过辅助任务的隐藏层获得新的表示，辅助任务的隐藏层与主任务中的相同，如式(16)所示。

$$h^d_{\text{aux}} = \text{dense}_{\text{aux}}(h_{\text{aux}}) \tag{16}$$

2.2.3 联合学习

使用 Softmax 层（Softmax Layer）对主任务中的表示进行预测，并使用另一个 Softmax 层对辅助任务中的表示进行预测，如式(17)、式(18)所示。

$$\hat{p}_\theta(y^{\text{main}} \mid T^{\text{input}}_{\text{target}}) = \text{softmax}(W^m h^d_{\text{main}} + b^m) \tag{17}$$

$$\hat{p}_\theta(y^{\text{aux}} \mid T^{\text{input}}_{\text{source}}) = \text{softmax}(W^a h^{\text{d}}_{\text{aux}} + b^a) \tag{18}$$

其中，$\hat{p}_\theta(y^{\text{main}} \mid T^{\text{input}}_{\text{target}})$ 为主任务的预测结果，$\hat{p}_\theta(y^{\text{aux}} \mid T^{\text{input}}_{\text{source}})$ 为辅助任务的预测结果。

最后，基于联合学习的中文分词模型的损失函数为主任务的损失函数和辅助任务的损失函数的加权线性损失之和，如式(19)所示。

$$\begin{aligned} J(\theta) = &-\lambda \cdot \sum_{i=1}^{N}\sum_{j=1}^{C} y_j^{\text{main}} \cdot \log \hat{p}_\theta(y_j^{\text{main}} \mid T^{\text{input}}_{\text{target},i}) \\ &-(1-\lambda)\cdot \sum_{i=1}^{N}\sum_{j=1}^{C} y_j^{\text{aux}} \cdot \log \hat{p}_\theta(y_j^{\text{aux}} \mid T^{\text{input}}_{\text{source},i}) \\ &+\frac{l}{2}\|\theta\|_2^2 \end{aligned} \tag{19}$$

其中，λ 为权重，是主任务和辅助任务的参数，y_j^{main} 和 y_j^{aux} 表示被预测为标签 j 的概率，N 为训练样本数目，C 为标签的数目，$T^{\text{input}}_{\text{target},i}$ 和 $T^{\text{input}}_{\text{source},i}$ 分别为目标领域和源领域的第 i 个训练样本，l 为控制模型复杂度的正则化参数。

3 实验设计与分析

本节将给出本文提出的基于联合学习的中文分词模型的结果，同时对各个方法进行详细分析。

3.1 实验设置

(1) 数据设置：本文的实验数据集的来源为中

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

国裁判文书网①，我们对其中的婚姻领域和合同领域进行数据的收集和分词的标注。在实验时，本文在两个领域中分别随机选取 100 篇作为实验样本，据统计，在这 100 个样本中，合同领域共存在 66 755 个词，婚姻领域共存在 46 425 个词。主任务的训练数据分别为目标领域数据的 10%、20%、30% 和 40%样本，验证集为目标领域数据的 10%，测试集为目标领域的 20%，辅助任务的训练集为源领域相同数量的样本，测试集与主任务中的样本一致。

(2) 文本表示：本文使用 Word2Vec 方法对样本中字的一元特征和二元特征进行训练，得到字对应的向量。我们将上下文窗口的长度设置为 2。

(3) 参数设置：本文使用 LSTM 模型进行中文分词实验，模型中的具体超参数如表 1 所示。

表 1　模型中的参数值

参数表述	参数值
词一元特征总数	40 000
LSTM 层输出维度	128
全连接层输出维度	64
Dropout 程度	0.25
迭代次数	30

在实验中，采用 F_1 值作为衡量分词效果的标准。F_1 值具体的计算方法如式(20)所示。

$$F_1\text{-Score} = \frac{P + R \times (1+\beta^2)}{P + R \times \beta^2} \tag{20}$$

其中，P 表示分词准确率，R 表示分词召回率，β 为平衡因子。β 大于 1 时，准确率比召回率更重要；β 小于 1 时，召回率比准确率更重要；β 等于 1 时，二者同等重要。在本文的实验中，β 取 1。

3.2　实验结果

本节中，我们将介绍几种中文分词方法，对所有方法进行结果统计和实验结果进行分析。

(1) 面向源领域的中文分词方法(Source_LSTM)：该方法仅使用 LSTM 模型对源领域数据进行训练。

(2) 面向目标领域的中文分词方法(Target_LSTM)：该方法仅使用 LSTM 模型对目标领域数据进行训练。

(3) 面向混合领域的中文分词方法(Mix_LSTM)：该方法将目标领域数据与源领域数据一起作为训练样本，并使用 LSTM 模型对所有训练样本进行训练。

(4) 基于特征增强的中文分词方法(Feature_Augmentation)：该方法由 Daumé III[21] 通过扩展不同领域中数据的特征提升领域适应性，在跨领域任务中可以有效改善性能。本节将源领域和目标领域混合后的样本进行特征的扩展，并使用 LSTM 模型对其训练。

(5) 基于联合学习的跨领域中文分词(Multi_LSTM)：该方法为本文提出的分词方法。在实验中，我们将 λ 设为 0.7、0.8 和 0.9，表 2 和表 3 展示了 λ 值为 0.7、0.8 和 0.9 在验证集中的实验结果。

从表 2 和表 3 得知，λ 为 0.8 时，在各个规模的样本下均取得了最优结果，源领域的样本产生的噪声最小，因此本文联合学习模型中的 λ 定为 0.8。

表 2　基于不同参数的跨领域联合学习 *F* 值(%)
(源领域：婚姻领域 目标领域：合同领域)

训练数据 / λ	10%	20%	30%	40%
λ=0.7	82.6	87.4	92.1	93.5
λ=0.8	85.5	91.0	93.2	94.1
λ=0.9	82.5	88.4	92.8	93.3

表 3　基于不同参数的跨领域联合学习 *F* 值(%)
(源领域：合同领域 目标领域：婚姻领域)

训练数据 / λ	10%	20%	30%	40%
λ=0.7	75.7	86.1	88.9	89.8
λ=0.8	77.5	86.9	89.3	90.9
λ=0.9	75.8	86.3	88.5	90.3

表 4 和表 5 给出了所有方法的结果，从表中可以得出以下 4 点。

(1) 面向目标领域的法律文书中文分词方法在训练样本规模相同时分词性能明显优于面向源领域的法律文书中文分词方法。原因在于跨领域训练具有领域适应性问题。具体而言，针对源领域训练获得的分词器不能有效适应目标领域的数据，因此获得了较差的性能。

(2) 在样本较少时，Target_LSTM 方法的性能远不及其余的跨领域的方法(Mix_LSTM、Feature_Augmentation 和 Multi_LSTM)。因此，单任务的中文分词方法难以适应目标领域语料匮乏的情况。

① http://wenshu.court.gov.cn/

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

(3) Mix_LSTM 方法直接混合源领域和目标领域的样本,增加了训练样本数目,因此其分词性能优于基线方法。

(4) 本文的 Multi_LSTM 方法分词结果优于所有基线方法。在语料匮乏的情况下提升更为显著。该方法的有效性体现在可以快速地加入法律专业词汇(如"裁定书""判决书""纠纷""诉讼"等),其他方法则不能在缺乏语料的情况下快速识别。因此,本文的方法能够有效地降低人工标注数据的成本,并在此基础上拥有优异的分词性能。

表 4 法律文书上各方法的 F 值(%)
(源领域:婚姻领域 目标领域:合同领域)

方法名 \ 训练数据	10%	20%	30%	40%
Source_LSTM	69.4	75.3	78.7	81.2
Target_LSTM	77.6	87.4	92.3	92.9
Mix_LSTM	82.5	88.5	90.9	93.3
Feature_Augmentation	75.6	87.1	91.0	93.0
Multi_LSTM	85.9	90.9	93.0	94.0

表 5 法律文书上各方法的 F 值(%)
(源领域:合同领域 目标领域:婚姻领域)

方法名 \ 训练数据	10%	20%	30%	40%
Source_LSTM	66.3	75.6	79.3	82.3
Target_LSTM	69.8	84.6	87.6	89.9
Mix_LSTM	75.1	85.2	87.0	90.6
Feature_Augmentation	72.3	84.5	87.9	89.6
Multi_LSTM	77.3	87.1	89.2	91.1

4 结语

针对法律文书,本文提出了一种基于联合学习的跨领域中文分词方法。该方法通过联合学习结合源领域和目标领域的少量已标注样本,使其更加准确地识别目标领域中的词语。具体而言,首先,本文将目标领域的预测作为主任务,将源领域的预测作为辅助任务;其次,使用共享 LSTM 层得到两个任务的辅助表示,让主任务和辅助任务共同参与分词结果的评判;再次,将该辅助表示与主任务的表示进行融合;最后,利用 LSTM 对主任务的结果进行预测。实验结果表明,本文的方法在目标领域拥有较少的训练样本条件下可以获得较好的分词性能。

下一步的工作,我们将使用目标领域中的未标注样本进行半监督的分词任务。此外,我们想收集更多相关领域法律文书,并对本文提出的方法进行有效性检验,希望解决跨领域的中文分词中的领域适应性问题。

参考文献

[1] 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.

[2] 李荣陆,王建会,陈晓云,等. 使用最大熵模型进行中文文本分类[J]. 计算机研究与发展,2005,42(1):94-101.

[3] 迟呈英,于长远,战学刚. 基于条件随机场的中文分词方法[J]. 情报杂志,2008,27(5):79-81.

[4] Lafferty J D, Mccallum A, Pereira F C N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data[J]. Proceedings of Icml, 2001, 3(2):282-289.

[5] Peng F, Mccallum A. Chinese segmentation and new word detection using conditional random fields[J]. Proceedings of Coding, 2004:562-568.

[6] 任智慧,徐浩煜,封松林,等. 基于 LSTM 网络的序列标注中文分词法[J]. 计算机应用研究,2017,34(5):1321-1324.

[7] 严倩. 面向法律文书的中文分词方法研究[D]. 苏州:苏州大学硕士学位论文,2018.

[8] Zheng X, Chen H, Xu T. Deep learning for Chinese word segmentation and POS tagging[C]//Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, USA: Association for Computational Linguistics, 2013: 647-657.

[9] Chen X, Qiu X, Zhu C, et al. Long short-term memory neural networks for Chinese word segmentation [C]//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon, Portugal: Association for Computational, 2015: 1197-1206.

[10] Yao Y, Huang Z. Bi-directional LSTM recurrent neural network for Chinese word segmentation[C]//Proceedings of International Conference on Neural Information Processing. Springer, Cham, 2016: 345-353.

[11] Xu J, Sun X. Dependency-based gated recursive neural network for Chinese word segmentation[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Berlin, Germany: Association for Computational Linguistics, 2016: 567-572.

[12] 金宸,李维华,姬晨,等. 基于双向 LSTM 神经网络

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

模型的中文分词[J]. 中文信息学报，2018，32(2)：29-37.
[13] Kamper H，Jansen A，Goldwater S. Unsupervised word segmentation and lexicon discovery using acoustic word embeddings[J]. IEEE/ACM Transactions on Audio，Speech and Language Processing (TASLP)，2016，24(4)：669-679.
[14] 邓丽萍，罗智勇. 基于半监督 CRF 的跨领域中文分词[J]. 中文信息学报，2017，31(4)：9-19.
[15] 佟德琴. 基于字词联合解码的中文分词研究[D]. 大连：大连理工大学硕士学位论文，2011.
[16] 许华婷，张玉洁，杨晓晖，等. 基于 Active Learning 的中文分词领域自适应[J]. 中文信息学报，2015，29(5)：55-62.
[17] Rumelhart D E，Hinton G E，Williams R J. Learning representations by back-propagating errors[M]. Neurocomputing：foundations of research. MIT Press，1988：533-536.
[18] Hochreiter S，J Schmidhuber. Long short-term memory[J]. Neural Computation，1997，9(8)：1735-1780.
[19] Graves A. Generating sequences with recurrent neural networks[J]. Computer Science，2013，(1308.0850)：1-43.
[20] Hinton G，N Srivastava，A Krizhevsky A，et al. Improving neural networks by preventing co-adaptation of feature detectors[J]. Computer Science，2012，3(4)：212-223.
[21] Daumé III H. Frustratingly easy domain adaptation[C]//Proceedings of the 45th Annual Meeting on Association for Computational Linguistics. Prague，Czech：Association for Computational Linguistics，2007：256-263.

江明奇(1994—)，硕士研究生，主要研究领域为自然语言处理。
E-mail：mqjiang@stu.suda.edu.cn

严倩(1993—)，硕士研究生，主要研究领域为自然语言处理。
E-mail：1361003647@qq.com

李寿山(1980—)，博士，教授，主要研究领域为情感分析、自然语言处理。
E-mail：lishoushan@suda.edu.cn

(上接第 16 页)

[17] Mikolov T，Chen K，Corrado G，et al. Efficient estimation of word representations in vector space[C]//Proceedings of International Conference on Learning Representations，Arizona，USA，2013：1388-1429.
[18] Pei W，Ge T，Chang B. Max-margin tensor neural network for Chinese word segmentation[C]//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing，Lisbon，Portugal，2015：1197-1206.
[19] Hochreiter S，Schmidhuber J. Long short-term memory[J]. Neural Computation，1997，9(8)：1735-1780.
[20] 成于思，施云涛. 面向专业领域的中文分词方法[J]. 计算机工程与应用，2018，54(17)：30-34，109.
[21] 俞士汶，段惠明，朱学锋，等. 北京大学现代汉语语料库基本加工规范[J]. 中文信息学报，2002，16(5)：49-64.
[22] 尹海良. 现代汉语类词缀研究[D]. 济南：山东大学博士学位论文，2007.

成于思(1983—)，博士，讲师，主要研究领域为文本挖掘与工程法律。
E-mail：xchengyusi@163.com

施云涛(1985—)，硕士，高级工程师，主要研究领域为自然语言处理。
E-mail：shiyuntao@js.chinamobile.com

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net