基于神经网络的半监督CRF中文分词

罗智勇
北京语言大学
luo_zy@blcu.edu.cn

张明明
北京语言大学
MingmingZhang_blcu@163.com

韩玉蛟
北京语言大学
chloe_hanyu@163.com

赵志琳
北京语言大学
zhi_lin_zhao@163.com

摘要

分词是中文信息处理的基础任务之一。目前全监督中文分词技术已相对成熟并在通用领域取得较好效果，但全监督方法存在依赖大规模标注语料且领域迁移能力差的问题，特别是跨领域未登录词识别性能不佳。为缓解上述问题，本文提出了一种充分利用相对易得的目标领域无标注文本、实现跨领域迁移的半监督中文分词框架；并设计实现了基于词记忆网络和序列条件熵的半监督CRF中文分词模型。实验结果表明,该模型在多个领域数据集上F-值和ROOV值分别取得最高2.35%和12.12%的提升，并在多个数据集上成为当前好结果。

关键词： 半监督 ； 序列条件熵 ； 跨领域 ； 中文分词

Semi-supervised CRF Chinese Word Segmentation based on Neural Network

Zhiyong Luo
北京语言大学
luo_zy@blcu.edu.cn

Mingming Zhang
北京语言大学
MingmingZhang_blcu@163.com

Yujiao Han
北京语言大学
chloe_hanyu@163.com

Zhilin Zhao
北京语言大学
zhi_lin_zhao@163.com

Abstract

Chinese word segmentation (CWS) is a fundamental task of natural language processing. Currently, CWS model using fully supervised learning technology has achieved good results in the common domain. However, it has the problem of relying on the large-scale annotated corpus and poor domain migration capability, especially the cross-domain OOV word recognition is not effective. In order to alleviate these problems, this paper proposes a semi-supervised CWS framework that uses relatively easy-to-obtain unlabeled texts in the target domain to achieve cross-domain transfer. We design a semi-supervised model based on word memory network and sequence conditional entropy. Our model based on this framework achieves significant improvements in F-scores and ROOV on several datasets,some of them are state-of-the-art.The maximum F-value and ROOV improvements are 2.35% and 12.12%.

Keywords: Semi-supervised learning , Conditional entropy for sequence , Chinese word segmentation

第二十一届中国计算语言学大会论文集，第644页-第655页，南昌，中国，2022年10月14日至16日。

1 引言

中文文本可视为由汉字字符（包含标点）组成的连续字符串，且词间无明确标记。因此，在中文信息处理任务中，分词通常是词法分析层面的一项重要任务，准确的分词结果有助于提升基于词的深层次语言信息处理任务的性能。

Xue(2003)将中文分词转化为给汉字标注词位的序列标注任务，自此基于字序列标注方法被广泛应用于中文分词。近年来，随着深度学习技术的发展，基于神经网络的全监督序列标注模型逐渐应用于中文分词任务中 (Chen et al., 2015; Cai and Zhao, 2016; Cai et al., 2017)，不仅提升了歧义切分的性能，还在通用领域（新闻领域）中文分词上取得较好的结果。例如，He et al.(2019)基于多准则的通用领域（PKU数据集）中文分词F-值达96%以上。但是，全监督的分词方法依赖大规模人工标注语料，并且模型领域迁移能力差，训练数据和测试数据之间差异对模型效果影响很大。目前，中文分词的标注语料主要来自于新闻领域，特定专业领域（如医学、知识百科、小说等）标注语料稀少。因此，尽管全监督分词模型在通用领域分词任务上的准确率较高(He et al., 2019)，但是受限于其领域迁移能力，跨领域分词准确率和未登录词的识别性能都有待提升；此外，由于专业领域缺乏大规模标注语料，因而也无法通过全监督方式训练有效的领域分词器。但是，较于昂贵的专业领域标注语料，专业领域无标注文本却相对易得。因此，如何充分利用用现有大规模通用领域标注语料和专业领域无标注语料，构建具备较好的分词准确率和未登录词识别性能的半监督分词模型,是近年来中文分词任务关注的问题之一。

邓丽萍and罗智勇(2017)首次提出以条件熵作为正则化项，利用CRF++工具和人工特征模板构建半监督CRF并应用到跨领域分词任务上，使得百科领域分词F-值与未登录词召回率有效提升。但该模型基于统计机器学习的方法，存在需手工定制特征模板的弊端且不适用于神经网络框架；Fu et al.(2020)研究表明用大规模预训练语言模型BERT、ELMo等作为字符特征编码器有助于更好地提升未登录词识别的性能；Tian et al .(2020)构建了词记忆网络，将n-gram信息编码到上下文表示中来，提高了模型对于词语边界的预测能力，但是该方法未能有效利用目标领域无标注文本中的字符共现特征，因而领域迁移能力有限。

在上述研究的基础上，本文提出了一种基于神经网络的半监督跨领域序列标注框架，并实现了两种基于该框架的半监督中文分词模型：以BERT作为特征提取器的半监督CRF模型（记为BERT-semiCRF），以及增加词记忆网络的半监督CRF模型(记为BERT-WM-semiCRF)。上述模型将已标注通用领域分词语料和无标注专业领域文本作为输入，不仅可以通过编码无标注文本上下文信息，提高模型领域迁移能力和未登录词识别性能，还可以通过减小序列条件熵增强模型预测置信度，从而实现跨领域分词准确率与未登录词识别性能的有效提升。实验表明，在特定专业领域（如专利PT、小说ZX/FR、医学DM等数据集上F-值和未登录词召回率达当前最好，相较于基线模型，F-值提升最高达2.35%，未登录词召回率提升最高达12.12%。

本文第2节介绍中文分词的相关研究；第3节介绍基于神经网络的半监督CRF模型；论文的第4节介绍模型的实验验证、消融研究和实验分析；第5节为总结和研究展望。

2 相关研究

2.1 半监督中文分词

中文分词方法主要分为全监督、无监督和半监督三类。全监督分词方法要求训练数据集全部为有标注数据，因此存在依赖大规模标注语料且跨领域迁移能力差的缺陷；无监督分词方法的训练集则全部为无标注数据，主要用于新词发现任务，其分词准确性较低；而半监督分词方法指训练数据集中既包含有标注数据，也包含无标注数据的分词方法。目前，半监督中文分词方法主要分为两类：一类半监督方法通过设计损失函数 (Liu et al., 2014; Zhao et al., 2018)，使其能够利用标注和部分标注的中文分词数据训练，但是无法避免来自网络的自然标注数据带来的歧义、标注准则不一致的问题；另一类半监督方法通过自采样得到伪标注从而扩充训练样本(Liu and Zhang, 2012)，这种方法存在错误累加和未登录词别受限的问题；Wang and Xu (2017)使用训练好的教师模型自采样无标注数据构建词表，然后使用预训练方法获得该词

基金项目：国家自然科学基金(62076037)、北京语言大学研究生创新基金(中央高校基本科研业务费专项资金)项目成果(22YCX173)

表的向量表示，并将该信息以词嵌入方式增加到学生模型的全监督训练中，这是一种通过增加无标注文本特征实现半监督的方法，但是该方法缺乏对未登录词的关注；针对跨领域分词任务，Ding et al.(Ding et al., 2020)等人提出基于远距离标注和对抗训练的分词模型。该方法在F-值上有一定的提升，但需要为每个领域从初始状态训练模型，耗时较长，且分词效果很受无监督词挖掘器影响。

本文区别于上述研究工作，从增加无标注文本特征表示、设计可同时关注标注数据和完全无标注数据的损失函数这两个角度，构建基于神经网络的半监督CRF分词模型。

2.2 基于神经网络序列标注的中文分词

基于神经网络的序列标注模型通常包含字/词表示、特征提取、推理三大模块(He et al., 2020)。采用序列标注方法的中文分词，是以字为单位，因此第一模块为字表示。字表示模块将字映射到对应的表示向量上，该向量蕴含字意信息,早期使用上下文无关的静态向量，例如word2vec(Mikolov et al., 2013)等，现常用包含语境信息的动态向量，例如ELMo(Peters et al., 2018)、BERT(Devlin et al., 2018)等。特征提取模块又称上下文编码模块，用于捕获序列的上下文信息和边界信息，在基于神经网络的序列标注模型中通常将经过特征提取模块的向量表示作为发射状态矩阵，BILSTM常作为特征提取模块(Huang et al., 2015)。BERT既可作为独立的字表示模块，也可以视为字表示与特征提取为一体的表示层。推理模块的功能是给出输入序列的标注结果，常用可关注标签间转移关系的条件随机场(conditional random filed,CRF)。

本文提出的半监督框架建立在字表示、特征提取和推理模块结构上，BERT和加入词记忆网络的BERT作为字表示和特征提取层，基于神经网络的semiCRF作为推理模块。

3 基于神经网络的半监督CRF分词

本节将详细介绍半监督模型框架、模型具体结构和训练策略。首先形式化定义输入、输出并描述整个模型的半监督框架；接着介绍基于神经网络的半监督CRF模型结构和具体实现细节；最后介绍模型的训练策略。

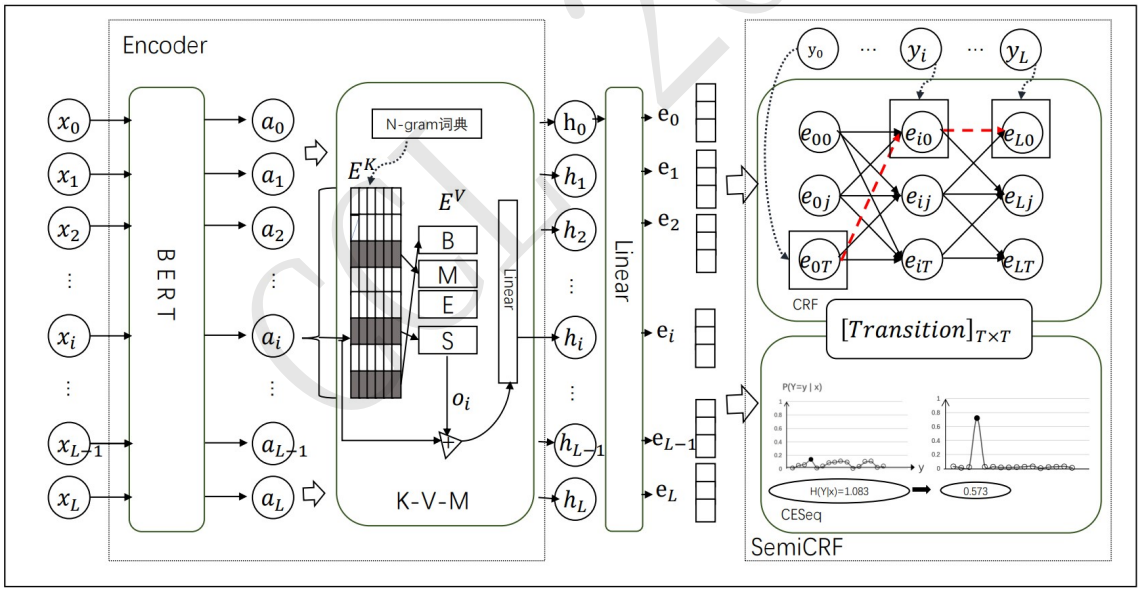

Figure 1: 模型结构

3.1 半监督框架

基于字标注的中文分词任务，通常形式化为：给定长度为n的句子x，其标注分词结果序列为y ,其中：$x = \langle x_1, x_2, \ldots, x_i, \ldots, x_n \rangle$, x_i表示句中第i个字，$y = \langle y_1, y_2, \ldots, y_i, \ldots, y_n \rangle$,且$y_i \in \{B, M, E, S\}$，分别表示词语的开始、中间、结束位置，以及单独成词。例如句子“一/面/鲜艳/的/五星红旗”的切分形式，可以表式为：x = <一，面，鲜，艳，的，五，星，红，旗>，y = $< S, S, B, E, S, B, M, M, E >$。在模型的训练阶段，全监督模型的输入为$(x^i, y^i)$

, $x^i \in X = \left\{x^1, x^2, .., x^N\right\}, y^i \in Y = \left\{y^1, y^2, .., y^N\right\}$ 其中X, Y 分别表示标注数据集的字符序列集合和标签序列集合，N 表示标注数据集样本规模。与全监督分词不同的是，半监督分词模型的输入是标注样本(x^i, y^i) 或无标注样本(x^j)，$x^j \in U = \left\{x^1, x^2, .., x^M\right\}$，其中$U$表示无标注数据的样本集。半监督模型的优化目标是最小化模型在训练数据上的损失：

$$\theta^* = \arg\min_{\theta} \left(M_\theta\left(X, Y\right) + \beta * M_\theta\left(U\right)\right) \tag{1}$$

M表示半监督模型，θ表示模型参数，$M_\theta\left(\bullet\right)$ 表示模型在数据集上的损失值，β为控制模型对无标注数据关注程度的超参数，当$\beta = 0$时，将变为全监督模型。

图1给出了本文提出的半监督模型整体结构，主要包括输入层、编码层$Encoder_{\theta_1}$、发射状态编码层和半监督CRF层$semiCRF_{\theta_2}$。

输入层 对于任何一个作为输入样本的句子$x = \langle x_1, x_2, \ldots, x_n\rangle$，$x \in X$ 或$x \in U$。在样本输入模型前，预处理工作将“$[CLS]$”和“$[SEP]$”加入到样本首尾,并将句子填充为所设最大长度L。此时样本表示为$x = \langle x_0, x_1, \ldots, x_L\rangle$。

编码层($Encoder_{\theta_1}$) 编码层主要获得具有上下文和边界信息的特征表示$h = \langle h_0, h_1, \ldots, h_L\rangle$, $h \in R^{L \times d_h}$, d_h为隐藏层维度,θ_1为模型中参数。主要核心模块包括预训练语言模型和词记忆网络（请见3.3节）。

发射状态编码层 该层将每一个字符的上下文编码特征向量，通过全连接层投射到字标注类别的未归一化概率，即发射状态$e \in R^{L \times T}$,如公式(3)所示，$w^T \in R^{d_h \times T}$，$\mathrm{b} \in R^{1 \times T}$，$T$为发射状态维度，即可预测标签数。

半监督CRF层($semiCRF_{\theta_2}$) 该层主要包含一个CRF模块和一个序列条件熵计算模块CEseq，以及两个模块之间共享的标签转移矩阵B，其中θ_2为标签转移矩阵参数。若原输入x来自标注数据集X，模型将通过CRF模块计算对应标签序列y的负对数似然作为模型预测损失值；若x来自无标注数据集U，模型则由序列条件熵模块CESeq计算模型预测的损失值。具体计算方法详见3.2节。

整个网络前向传递过程如公式(2)-(4)所示：

$$h = Encoder_{\theta_1}\left(x\right) \tag{2}$$

$$e = h * w^T + \mathrm{b} \tag{3}$$

$$loss = \begin{cases} semiCRF_{\theta_2}\left(e\right) & x \in U \\ semiCRF_{\theta_2}\left(e, y\right) & x \in X \end{cases} \tag{4}$$

在预测阶段，对于待预测样本,获得发射状态步骤与训练阶段相同，即式(2)、(3)；在获得发射状态后由semiCRF中CRF模块使用维特比算法解码，获得最大预测概率的标签序列，可形式化为式(5)：

$$\hat{y} = \underset{y \in \mathcal{Y}}{\operatorname{argmax}}\, p_\theta(y|x) \tag{5}$$

其中,$\mathcal{Y}$表示输入序列x所有可能标注序列集合。

本文实现的BERT-semiCRF和BERT-WM-semiCRF的Encoder分别对应BERT、加上词记忆网络的BERT(Tian et al., 2020)，即：$h = BERT_{\theta_1}\left(x\right)$ 和$h = BERT_WM_{\theta_1}\left(x\right)$，其中词记忆网络结构在3.3节说明。此处值得注意的是，本文实现的BERT-WM-semiCRF模型在构建n-gram词典时，数据源范围较原方法增加了无标注数据U，用无监督的邻接多样性(Feng et al., 2004a)方法构建,以便获得目标领域文本的字符组合特征表示信息。

3.2 基于神经网络的半监督条件随机场

本文实现的基于神经网络的半监督条件随机场（记为semiCRF）由推理模块CRF和序列条件熵(Conditional Entropy for Sequence,CESeq)两部分组成，它们共享标签间转移矩阵B。该模型具备提取标签之间的转移特征、计算损失值和解码功能。如式(6)所示，有标注样本损失为条件概率的负对数似然，无标注样本的损失为条件熵。

$$loss = \begin{cases} H_\theta\left(\mathcal{Y}|x\right) & x \in U \\ -logp_\theta\left(y|x\right) & x \in X, y \in Y \end{cases} \tag{6}$$

CRF (Huang et al., 2015; 李航and others, 2012)模块关注于有标注样本，它描述给定输入序列x产生标注序列y的条件概率为：

$$\begin{aligned} \mathrm{p}\left(y|x\right) &= \frac{\Psi(y|x)}{Z(x)} \\ Z\left(x\right) &= \sum_{y\in\mathcal{Y}} \Psi\left(y|x\right) \end{aligned} \tag{7}$$

其中，$\mathcal{Y}$表示x所有可能标注序列集合，当序列长度为L、标签数为T时，$\mathcal{Y}$集合大小为T^L。$\Psi(y|x)$ 是x标注为y的打分函数：

$$\Psi\left(y|x\right) = \prod_{i=1}^{L} \varphi\left(x, i, y_{i-1}, y_i\right) \tag{8}$$

$$\varphi\left(x, i, y_{i-1}, y_i\right) = \exp\left(e_{i,y_i} + B_{y_{i-1},y_i}\right) \tag{9}$$

其中，发射状态e由式(3)获得，$e \in R^{L\times T}$ ，e_{i,y_i}表示x的i号位置标注为标签y_i的发射状态分值；B_{y_{i-1},y_i}表示由标签y_{i-1} 转移到y_i的转移分值。为避免式(7)中$Z\left(x\right)$ 指数级时间复杂度，通常使用动态规划算法，对$Z\left(x\right)$通过以下递推公式求解：

$$Z\left(x\right) = \sum_{i=1}^{T} \exp\left(\alpha_L\left[i\right]\right) \tag{10}$$

$$\alpha_t = \log \sum_{i=1}^{T} \exp\left(\mathrm{M_t}\left[\mathrm{i}\right]\left[\mathrm{j}\right]\right) \tag{11}$$

$$\mathrm{M_t} = \begin{bmatrix} \alpha_{t-1} & \dots & \alpha_{t-1} \end{bmatrix}_{T\times T} + B + \begin{bmatrix} e_t \\ \dots \\ e_t \end{bmatrix}_{T\times T}, t = 1, 2, \dots, L \tag{12}$$

上式中，e_0为样本经过编码层和线性层后的第一个时序，即式(3)中e的第一个时序，$e_0^T \in R^{T\times 1}$ ，T 表示标签数，L 表示序列长度。

序列条件熵计算模块CESeq，关注于无标注数据的序列标注条件熵，表示在给定输入x的情况下,所有可能标注序列$\mathcal{Y}$的概率分布的不确定性，记为$H(\mathcal{Y}|x)$。当$\mathcal{Y}$的不确定性越小时，对于x的各可标注结果y'间的概率区分度越大，满足低密度划分原则。具体计算方式如公式(13)所示：

$$\mathrm{H}\left(\mathcal{Y}|\mathrm{x}\right) = -\sum_{y'\in\mathcal{Y}} p(y'|x)\log(y'|x) \tag{13}$$

上式中，$p(y'|x)$由式(6)定义。为避免遍历$\mathcal{Y}$所带来的指数级时间代价，邓丽萍and罗智勇(2017)采用子序列条件熵对(13)式进行化简，达到了和CRF相同时间复杂度级别$O(LT^2)$：

$$H\left(\mathcal{Y}|x\right) = -\sum_{y_n} p(y_n|x)\left[\log p(y_n|x) + H^{\alpha}(\mathcal{Y}_{1,2,\dots n-1}|y_n, x)\right] \tag{14}$$

$$H^{\alpha}(\mathcal{Y}_{1,2,\dots,t}|y_{t+1}, x) = \sum_{y_t} p(y_t|y_{t+1}, x)[\log p(y_t|y_{t+1}, x) + H^{\alpha}(\mathcal{Y}_{1,2,\dots t-1}|y_t, x)] \tag{15}$$

式中，n表示序列长度，y_i表示序列i号位置被标注的标签类别，$\mathcal{Y}$表示x可标注序列集合，$t \in \{0, \dots, n\}$，$H^{\alpha}\left(|y_0, x\right) = 0$ 。其中$p\left(y_t|y_{t+1}, x\right)$可有由(16)、(17)式获得：

$$\mathrm{p}\left(y_t y_{t+1}|x\right) = \frac{\sum_{y'\in\{\mathcal{Y}\cap y_t y_{t+1}\}} \Psi(y'|x)}{Z(x)} \tag{16}$$

$$\mathrm{p}\left(y_{t+1}|x\right) = \frac{\sum_{y'\in\{\mathcal{Y}\cap y_{t+1}\}} \Psi(y'|x)}{Z(x)} \tag{17}$$

$\mathrm{p}\left(y_{t+1}|x\right)$表示序列$x$的$t+1$号位置被标注为标签$y_{t+1}$ 的条件概率，$\Psi(\cdot|x)$由式(8)定义。集合$\{\mathcal{Y}\cap y_{t+1}\}$表示$x$的$t+1$号位置被标注为$y_{t+1}$的所有可能标注序列集合，集合$\{\mathcal{Y}\cap y_t y_{t-1}\}$同

理，表示t、$t+1$ 号位置被标注为$y_t y_{t+1}$ 的所有可能标注序列集合。本文对式(16-17)进一步推导，直接给出计算对数标签转移概率$p\left(y_t|y_{t+1},x\right)$的前向递推公式：

$$\log p(y_t = i|y_{t+1} = j, x) = \mathrm{M_{t+1}\,[i]\,[j]} - \alpha_{t+1}\,[j] \tag{18}$$

式(18)采用动态规划算法，其中$\mathrm{M_{t+1}}$与α_{t+1}由式(11)、(12)定义。

整体来看，建立在神经网络自动编码结构上的半监督条件随机场具备以下优势：能够关注无标注数据的文本特征，可以提高模型的领域迁移能力；CRF模块具备监督模型习得正确识别词边界的能力，CESeq模块能够扩大可能标注的结果序列之间的概率区分度。上述两部分组合，使模型具备对序列正确标注且正误区别明显的能力。从本文4.3.3节中的消融研究结果来看，加入semiCRF后的半监督中文分词的准确率、未登录词识别性能均有提升，可印证上述观点。

3.3 词记忆网络

词记忆网络(Wordhood Memory Network)(Tian et al., 2020)是对基于预训练语言模型的上下文特征编码的增强。通过加入键值网络结构，为每个输入的上下文编码表示a_i，增加了n-gram信息。在给定的上下文情况下，这些n-gram信息将显式地增加或降低当前字符成为某个词位的可能性或概率。具体计算方法如式(19)、(20)所示，其中,c_i表示输入字符i在当前语境下可查找到的m个n-gram,$E_{c_i}^K$为c_i的向量表示，例如，“分”在语境“原子结合成分子”中的可查找到的ngram有：“成分”、“分子”。n-gram词典由输入文本采用邻接多样性(Feng et al., 2004b)的无监督方法构建；$E_{c_i}^V$为当前字符i在各n-gram中对应的可能词位(BMES)的向量表示。

$$o_i = softmax\left(a_i, E_{c_i}^K\right) \cdot E_{c_i}^V \tag{19}$$

$$h_i = Linear\left(o_i + a_i\right) \tag{20}$$

3.4 模型训练

针对本文提出的半监督框架，我们提出了两种训练策略：联合训练和分步训练策略。为避免模型在无标注数据集上过度关注区分度而忽略正确性，设计如式(21)损失函数，与训练目标式(1)相对应，包含标注数据损失和无标注数据损失，意在使模型在关注区分度时，不偏离标注准则。

$$\mathrm{loss} = -\frac{1}{N}\sum_{i=1}^{N} logp_{\theta_1}\left(y^i|x^i\right) + \frac{\beta}{M}\sum_{j=1}^{M} H_{\theta_2}\left(\mathcal{Y}|x^j\right) \tag{21}$$

回顾模型的整体结构(图1)，来自源领域的标注数据或目标域的无标注数据x都会经过共享的上下文特征编码结构和发射状态编码层得到发射状态编码e；随后，在semiCRF层中,模型根据源领域提供的标注y，由CRF模块计算得到其负对数似然损失；属于目标领域的样例则进入CESeq模块，计算得到序列条件熵损失。模型的损失是上述两部分损失之和。值得说明的是，对于通用领域中文分词任务，本模型源领域与目标领域均为新闻领域，对于跨领域分词，不同目标域仅使用该领域对应的无标注数据。

本文实现的分步训练策略指先在源领域有标注训练集上对模型全监督训练，此时式(21)中β设置为0，在模型习得一定分词能力后再设定非零β进行半监督训练。联合训练策略则不再包含全监督训练步骤，直接使用非零β进行半监督训练，其中，每个batch中标注样本和无标注样本比例为1:1。

4 实验

本节介绍实验使用数据集和实验设置，然后展示我们实现的模型在各领域的分词效果、与其它工作的结果比较、消融研究，并进行分析与总结。

4.1 实验数据集

本文实验在四个领域的六个数据集上进行，包含：新闻领域的标注语料PKU（来自SIGHAN CWS BACKOFF 2005），是1998年人民日报1月份新闻文本；无标注语料pku-07，是1998年人民日报7月份新闻文本；专利领域（PT）、医学领域（DM）、小说领域

（ZX、DL和FR）5份无标注数据集和标注测试集均来源于(Ding et al., 2020)文献，其中小说领域三个语料分别是网络小说《诛仙》(ZX)、《斗罗大陆》(DL)、《凡人修仙传》(FR)。表1中列出了各语料详细信息，其中PKU的训练集和开发集为SIGHAN CWS BACKOFF 2005的训练集根据9:1划分所得，测试集与原测试集保持一致。为方便比较，其他领域数据集划分均与参考文献中保持一致。因语料存在以段落为间隔的情况，导致一个样本长度过长，因此本实验中对数据进行了以句号、问号、引号、省略号为分割符的分句处理。

数据集		句数	领域	标注
PKU	训练集	39.7K	新闻	是
	开发集	4.4K		
	测试集	4.2K		
pku-07	训练集	57.2K	新闻	否
PT	训练集	17.7K	专利	否
	测试集	1.0K		是
DM	训练集	32.0K	医学	否
	测试集	1.0K		是
ZX	训练集	59.0K	小说	否
	测试集	1.0K		是
DL	训练集	40.0K	小说	否
	测试集	1.0K		是
FR	训练集	148.0K	小说	否
	测试集	1.0K		是

Table 1: 实验数据

参数	
序列最大长度L	300
学习率	0.00001
隐藏层维度d_h	768
标签数T	7
超参数β	{0.1，0.05，0.01}
dropout	0.5
随机种子	42
n-gram词典阈值	3

Table 2: 参数设置

4.2 实验设置

实验中BERT使用"bert-base-chinese"作为初始模型参数，BERT-WM-SemiCRF 中词记忆网络的设置参考WMSeg(Tian et al., 2020)。其他参数设置如表2所示。中文分词的测评指标通常包含准确率（P）、召回率（R）、平衡P和R的F-值、未登录词召回率（ROOV）,本实验中采用F-值和ROOV作为主要评价指标，具体计算方式和前人工作保持一致，这里不再赘述。

4.3 实验结果与消融研究

为验证本文方法的有效性，本文以通用领域语料PKU在全监督方式下训练模型BERT-CRF和WMSEeg，获得通用领域分词器，在PKU和其他领域测试集上进行测评，以上述测评结果作为本文比较的基线。此外，本文方法也与近年其他方法在同数据集上的分词结果进行了比较。本文实现的模型在通用领域PKU数据集上最高F-值达96.76%，ROOV达87.48%，成为当前最好结果；在跨领域分词任务上，模型结果较两个基线模型在F-值和ROOV上均有提升，其中，专利领域F-值最大提升达2.78%、未登录词召回率提升达7.93%。

4.3.1 通用领域

本文实现的半监督模型在通用领域的实验结果如表3所示。我们首先将本文实现模型与全监督方法进行对比，表格1-3行为近年来基于神将网络的全监督中文分词在PKU数据集上的分词结果。接着我们列举了半监督分词模型的实验结果，为表格4-6行，WCC-CWS(Hao et al., 2017)是一种增加上下文特征表示的半监督方法，WE_CONV_SEG(Wang and Xu, 2017)则通过特征蒸馏的自采样实现，BILSTM_LM_PL(Zhao et al., 2018)构建了由交叉熵和语言模型组合的损失函数实现半监督；除WMSeg的实验结果为本文复现该论文结果，其他结果摘自原文献。本文实现的模型结果为表格最后三行，BERT-CRF是本文的基线模型，其它两个半监督模型由有标注的PKU训练集和无标注的人民日报中文文本训练。观察实验结果，可以发现：（1）本文方法实现的两个半监督模型，在未登录词召回率上超出所列全监督模型和半监督模型；其中，BERT-WM-semiCRF模型的实验结果达到当前最好，为96.76%和87.48%；（2）本文实现的两个模型较基线模型F-值分别提升0.12%、0.19%，未登录词召回率提升1.12%、1.34%，说明本文的模型不仅能提升分词准确率，在识别未登录词识方面提升更为显著；（3）与BILSTM_LM_PL使用部分标注语料实现半监督分词的结果比较，本文实现的半监督模型

MODEL	F	ROOV
BILSTM_CWS(2018)	96.10%	78.80%
MC_CWS(2019)	96.41%	78.91%
WMSeg(BERT)(2020)	96.73%	86.90%
WCC_CWS(2017)	96.00%	-
WE_CONV_SEG(2017)	96.50%	-
BILSTM_LM_PL(2018)	95.50%	-
our model		
BERT-CRF(BASE)	96.56%	86.13%
BERT-semiCRF(step)	96.69%	87.25%
BERT-WM-semiCRF	**96.76%**	**87.48%**

Table 3: 通用领域PKU数据集分词结果

高约1个百分点，一方面由于本文的特征提取层较BILSTM能更好的提取无标注文本上下文特征，另一方面，相较于构建语言模型和交叉熵损失的半监督方法，本文的semiCRF方法既能关注无标注文本，又避免了部分标注文本带来的噪音。

4.3.2 跨领域分词

	PT		ZX		DM		FR		DL	
	F	ROOV	F	ROOV	F	ROOV	F	ROOV	F	ROOV
Partial-CRF(2014)	85.00%	-	83.90%	-	82.80%	-	90.20%	-	92.50%	-
WCC_CWS(2017)	-	-	89.10%	70.40%	-	-	-	-	-	-
WEB_CWS(2019)	85.10%	-	89.60%	-	82.20%	-	89.60%	-	93.50%	-
DAAT(2020)	89.60%	-	90.90%	-	85.00%	-	93.10%	-	**94.10%**	-
WSEeg(2020)	91.02%	78.64%	90.48%	69.42%	88.55%	75.48%	91.21%	73.74%	92.22%	65.03%
ours:										
BERT-CRF	90.58%	76.36%	90.20%	69.00%	88.39%	74.67%	90.82%	72.78%	92.21%	65.82%
BERT-semiCRF	93.30%	81.79%	91.02%	71.01%	89.72%	78.90%	**93.15%**	83.89%	92.58%	**66.48%**
BERT-WM-semiCRF	**93.37%**	**84.29%**	**91.11%**	**72.98%**	**90.43%**	**80.73%**	93.01%	**85.86%**	92.64%	66.22%

Table 4: 跨领域分词实验结果

表4给出了本文在专利(PT)、医学(DM)、小说（ZX、FR、DL）三个领域5个数据集上的跨领域分词实验结果。表中最后三行列出了我们实现的全监督基线模型（BERT-CRF）、两个半监督模型在各领域数据集上的实验结果，其中半监督模型均由有标注的新闻领域PKU训练集、无标注的目标领域中文文本联合训练得到。表格1-4行，列出了近年使用半监督方法进行跨领域分词在相同的数据集上的实验结果。其中，DAAT模型在各跨领域分词上的F-值为最高(Ding et al., 2020)，本文与该模型训练数据集的设置一致。表格第5行WSEeg(Tian et al., 2020)是用PKU训练的全监督模型的跨领域分词结果。观察实验结果，可以发现：（1）本文实现的半监督模型除DL数据集F值外，其它评估值均达到最好结果，其中BERT-semiCRF在FR小说数据集的F值上最好，达93.15%，其余则是加上词记忆网络的BERT-WM-semiCRF有更高的结果；（2）DAAT为近年来半监督方法在以上数据集最好的结果，本文在各数据集上的分词准确率提升分别为：从89.60%到93.37%，从90.90%到91.11%，从85.00%到90.43%，从93.10%到93.15%，最高提升达5.43%，对应医学数据集；（3）本文实现的半监督模型较全监督的基线模型，在分词准确率和未登录词召回率上都有提升，这一点表5展示的更清晰，专利领域F-值提升最大，为2.72%，FR小说领域未登录词识别提升最大，达12.12%；

4.3.3 消融研究

为更清晰展示本文方法的效果，表5列出了序列条件熵和词记忆网络模块在新闻、医学、专利、小说各领域数据集上的消融实验结果比较。BERT-CRF 、WMSeg(Tian et al., 2020)为两个对比的基线模型，均使用用PKU训练集全监督训；BERT-semiCRF相较于基线模型1，使用了semiCRF和与测试集领域对应的无标注数据U，而semiCRF相较于CRF增加了计算序

列条件熵的辅助推理模块CESeq,因此信息列BERT-semiCRF与对比基线模型BASE1的区别记为：+U+CESeq。BERT-WM-semiCRF在BERT-semiCRF基础上增加了词记忆网络，因此较基线模型BASE1的区别为增加了辅助训练的无标注数据、词记忆网络、辅助推理模块，因此记为：+U+K-V-M+CESeq。分析表格数据，可以发现：

（1）本文提出的半监督框架下的两个分词模型，在同领域（PKU）和跨领域（DM、PT、FR）上的分词F值和未登录词召回率较全监督基线模型均有提升，说明本文的半监督方法在提高模型分词性能和未登录词识别能力上有效；

（2）表中第3、4行，增加无标注数据和模块CESeq模型在各领域结果均有提升，其中专利领域F值提升达2.72%，小说FR未登录词召回率提升达12.12%，说明增加模块CESeq的semiCRF的半监督方法有效；对比3、4行的提升幅度，发现基线模型未登录词识别相对较低的数据集，未登录词召回率提升更明显，说明模型有较强泛化能力和迁移能力；

（3）对比表中4、5行的提升幅度，第5行提升值大于第4行，说明在半监督框架下增加词记忆网络也有助于分词性能提升，但是它的提升效果不如增加模块CESeq；

（4）与通用领域的提升幅度相比较，跨领域分词提升幅度更明显，一方面由于跨领域分词任务难大，基线模型待提升空间大；另一方面由于本文的最小化无标注数据序列条件熵的辅助模块（CESeq）使得模型可以学习目标领域的信息，平衡模型"看到"的目标域和源领域的数据分布，从而达到提升效果。

			PKU		DM		PT		FR	
模型名称	比较	信息	F	ROOV	F	ROOV	F	ROOV	F	ROOV
BERT-CRF	-	BASE1	96.56%	86.13%	88.39%	74.67%	90.58%	76.36%	90.82%	72.78%
WMSeg	-	BASE2	96.73%	86.90%	88.55%	75.48%	91.02%	78.64%	91.21%	73.74%
BERT-semiCRF	BASE1	+U+CESeq	+0.12%	+1.12%	+1.34%	+4.23%	+2.72%	+5.44%	+2.33%	+11.12%
BERT-WM-semiCRF	BASE2	+U+CESeq	+0.03%	+0.58%	+1.87%	+5.25%	+2.35%	+5.65%	+1.80%	+12.12%
	BASE1	+U+K-V-M+CESeq	+0.19%	+1.34%	+2.04%	+6.06%	+2.78%	+7.94%	+2.19%	+13.09%

Table 5: 消融实验

4.4 实验分析

4.4.1 模型预测结果的置信度

数据集	BERT_CRF	BERT_semiCRF	BERT_WM_semiCRF
PKU	0.542	0.577	0.569
DM	0.649	0.694	0.719

Table 6: 模型平均置信度

为进一步说明semiCRF的效果并证明该方法的有效性，我们分别使用基线模型（BERT-CRF）、两个半监督模型为PKU、DM测试集样本的正确标注结果打分，表6给出评分均值，分值取值是[0,1]的概率表示。模型为正确标注序列打出的分值越高，说明模型对正确标注结果的置信度越大。从表中可以看出，增加辅助推理模块的半监督模型较全监督模型的平均分值均有提升，说明增加序列条件熵作为semiCRF的辅助推理模块，模型预测置信度会更高，从而提高了分词效果。

4.4.2 训练策略与超参数的选择

本文根据半监督结构提出了两种训练策略：分步训练策略和联合训练策略。表8给出半监督模型分步、联合两种策略下的实验比较，同领域（PKU）分词，采用联合训练策略有较好的结果，而在跨领域分词任务中，使用分步训练会取得相对较好的实验结果。由于联合训练需要从初始状态训练模型，而分步训练可在半监督训练时，加载同一个已训练好的全监督模型再训练，会节省训练耗时。

	PKU		FR	
joint	96.76%	87.48%	92.58%	83.39%
stepwise	96.69%	87.25%	93.15%	83.89%

Table 7: 不同训练策略下的实验结果

本文在3.2节半监督框架中指出，式(21)中的超参数β是模型对无标注文本的关注程度的调节，表给出了模型在不同β值下在专利领域PT上的评测结果，β的选择对结果存在一定幅度的影响，根据实验经验，选择较大β值会使得模型过分关注无标注文本，忽略分词标准的学习，导致结果降低，因此，在选择β值时应从较小值选取。

MODEL	β	F	ROOV
BERT-semiCRF	0.1	91.77%	83.12%
	0.05	93.01%	83.30%
	0.01	93.30%	81.80%
BERT-WM-semiCRF	0.1	93.00%	82.39%
	0.05	92.93%	83.04%
	0.01	93.37%	84.29%

Table 8: 不同β的影响

4.4.3 未登录词分析

本文对BERT-CRF基线模型和BERT-semiCRF在FR数据集上的未登录词识别情况进行了详细分析。图2为半监督方法较基线模型新增识别的未登录词的分布情况，有新增识别的未登录词种类数87种(图中第一列)，共计328频次（图中第二列）。其中，新发现未登录词57种，占有新增识别的未登录词总种数的66%，主要为“御剑”、“剑诀”、“五行环”一类的低频未登录词，占总增加频数的22%。基线模型已发现未登录词，但是在某些句子中未能正确识别，而半监督模型正确识别，这类有30种，占种数的34%，但是占总频数的78%，为“玄骨”、“冰焰”、“修士”等高频未登录词。上述现象说明，本文实现的半监督模型不仅具备简单高频未登录词在复杂语境下识别能力，还有更强的未登录词发现能力，能够发现低频未登录词。

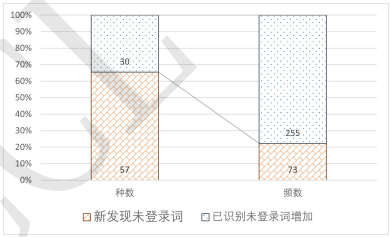

Figure 2: 新增识别的未登录种类和频数分布情况

4.5 总结与展望

词是中文信息处理的基本任务之一，分词结果的准确性将影响基于词的深层次语言信息处理任务的性能。本文提出了一种充分利用相对易得的目标领域无标注文本、实现跨领域迁移的半监督中文分词框架；通过引入词记忆网络和序列条件熵方法，不但提高了跨领域分词任务的准确性，特别是提升了跨领域未登录词识别的召回率。实验结果表明，本文提出的方法增强了分词模型的泛化能力和跨领域迁移能力。

本文还有一些尚未开展的工作和不足之处，包括如何克服不同标注语料在标注准则上的不一致问题，下一步研究可以增加模型多准则学习能力。此外，本文实现的半监督框架是基于神经网络的半监督序列标注框架，在中文分词上较好的提升效果，理论上，该半监督框架可应用于任何序列标注任务，因此，可尝试将该框架应用于命名实体识别、方面级情感分析等序列标注任务。

参考文献

Deng Cai and Hai Zhao. 2016. Neural word segmentation learning for chinese. *arXiv preprint arXiv:1606.04300*.

Deng Cai, Hai Zhao, Zhisong Zhang, Yuan Xin, Yongjian Wu, and Feiyue Huang. 2017. Fast and accurate neural word segmentation for chinese. *arXiv preprint arXiv:1704.07047*.

Xinchi Chen, Xipeng Qiu, Chenxi Zhu, Pengfei Liu, and Xuan-Jing Huang. 2015. Long short-term memory neural networks for chinese word segmentation. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 1197–1206.

Jacob Devlin, Ming Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

N. Ding, D. Long, G. Xu, M. Zhu, and H. T. Zheng. 2020. Coupling distant annotation and adversarial training for cross-domain chinese word segmentation. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

H. Feng, C. Kang, X. Deng, and W. Zheng. 2004a. Accessor variety criteria for chinese word extraction. *Computational Linguistics*, 30(1):75–93.

Haodi Feng, Kang Chen, Xiaotie Deng, and Weimin Zheng. 2004b. Accessor variety criteria for chinese word extraction. *Computational linguistics*, 30(1):75–93.

Jinlan Fu, Pengfei Liu, Qi Zhang, and Xuanjing Huang. 2020. Rethinkcws: Is chinese word segmentation a solved task? *arXiv preprint arXiv:2011.06858*.

Z. Hao, Z. Yu, Z. Yue, S. Huang, and J. Chen. 2017. Word-context character embeddings for chinese word segmentation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*.

Han He, Lei Wu, Hua Yan, Zhimin Gao, Yi Feng, and George Townsend. 2019. Effective neural solution for multi-criteria word segmentation. In *Smart Intelligent Computing and Applications*, pages 133–142. Springer.

Z. He, Z. Wang, W. Wei, S. Feng, X. Mao, and S. Jiang. 2020. A survey on recent advances in sequence labeling from deep learning models.

Z. Huang, X. Wei, and Y. Kai. 2015. Bidirectional lstm-crf models for sequence tagging. *Computer Science*.

Yang Liu and Yue Zhang. 2012. Unsupervised domain adaptation for joint segmentation and pos-tagging. In *Proceedings of COLING 2012: Posters*, pages 745–754.

Yijia Liu, Yue Zhang, Wanxiang Che, Ting Liu, and Fan Wu. 2014. Domain adaptation for crf-based chinese word segmentation using free annotations. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 864–874.

Ji Ma, Kuzman Ganchev, and David Weiss. 2018. State-of-the-art Chinese word segmentation with Bi-LSTMs. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4902–4908, Brussels, Belgium, October-November. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *Computer Science*.

Matthew Peters, M. Neumann, M. Iyyer, M. Gardner, and L. Zettlemoyer. 2018. Deep contextualized word representations.

Yuanhe Tian, Yan Song, Fei Xia, Tong Zhang, and Yonggang Wang. 2020. Improving chinese word segmentation with wordhood memory networks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8274–8285.

Chunqi Wang and Bo Xu. 2017. Convolutional neural network with word embeddings for Chinese word segmentation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 163–172, Taipei, Taiwan, November. Asian Federation of Natural Language Processing.

Nianwen Xue. 2003. Chinese word segmentation as character tagging. In *International Journal of Computational Linguistics & Chinese Language Processing, Volume 8, Number 1, February 2003: Special Issue on Word Formation and Chinese Language Processing*, pages 29–48.

Yuxiao Ye, Yue Zhang, Weikang Li, Likun Qiu, and Jian Sun. 2019. Improving cross-domain chinese word segmentation with word embeddings. *CoRR*, abs/1903.01698.

Lujun Zhao, Qi Zhang, Peng Wang, and Xiaoyu Liu. 2018. Neural networks incorporating unlabeled and partially-labeled data for cross-domain chinese word segmentation. In *IJCAI*, pages 4602–4608.

李航et al. 2012. 统计学习方法. Qing hua da xue chu ban she.

邓丽萍and 罗智勇. 2017. 基于半监督crf 的跨领域中文分词. 中文信息学报, 31(4):9–19.