一种采用机器阅读理解模型的中文分词方法

周裕林[1,2]，陈艳平[1,2]，黄瑞章[1,2]，秦永彬[1,2]，林川[1,2]
（1.公共大数据国家重点实验室，550025，贵阳；2.贵州大学计算机科学与技术学院，550025，贵阳）

摘要：针对中文分词序列标注模型很难获取句子的长距离语义依赖，导致输入特征使用不充分、边界样本少导致数据不平衡的问题，提出了一种基于机器阅读理解模型的中文分词方法。将序列标注任务转换成机器阅读理解任务，通过构建问题信息、文本内容和词组答案的三元组，以有效利用句子中的输入特征；将三元组信息通过Transformer的双向编码器（BERT）进行预训练捕获上下文信息，结合二进制分类器预测词组答案；通过改进原有的交叉熵损失函数缓解数据不平衡问题。在Bakeoff2005语料库的4个公共数据集PKU、MSRA、CITYU和AS上的实验结果表明：所提方法的F_1分别为96.64%、97.8%、97.02%和96.02%，与其他主流的神经网络序列标注模型进行对比，分别提高了0.13%、0.37%、0.4%和0.08%。

关键词：中文分词；序列标注；歧义词；机器阅读理解
中图分类号：TP391 **文献标志码**：A
DOI：10.7652/xjtuxb202208010 **文章编号**：0253-987X(2022)08-0095-09

OSID码

Machine Reading Comprehension Model for Chinese Word Segmentation

ZHOU Yulin[1,2], CHEN Yanping[1,2], HUANG Ruizhang[1,2], QIN Yongbin[1,2], LIN Chuan[1,2]
(1. State Key Laboratory of Public Big Data, Guiyang 550025, China;
2. College of Computer Science & Technology, Guizhou University, Guiyang 550025, China)

Abstract: Conventional sequence models for Chinese word segmentation are difficult to encode long distance semantic dependencies of a sentence, cannot make full use of input features, and have few boundary samples for use, which leads to data imbalance. In view of this, this paper proposes a machine reading comprehension (MRC) model for Chinese word segmentation. First, the sequence labelling task for Chinese word segmentation is converted into a machine reading comprehension task. This model constructs a triple relationship among question information, text content and answers to enrich the input features. Then, the triple relationship information is pre-trained by bidirectional encoder representation from transformers (BERT) to capture the contextual information, and a binary classifier is used to predict the word answers. Finally, the original cross-entropy loss function is improved to alleviate the data imbalance between examples. The experiment results show that machine reading comprehension model achieves F_1 value of 96.64%, 97.8%, 97.02% and 96.02% with four public datasets used: PKU, MSRA, CITYU and AS. Compared with other neural network sequence labeling models, this model improves the corresponding F_1 value by 0.13%, 0.37%, 0.4% and 0.08%.

Keywords: Chinese word segmentation; sequence annotation; ambiguous word; machine reading comprehension

收稿日期：2021-12-16。 作者简介：周裕林(1997—)，男，硕士生；陈艳平(通信作者)，男，副教授。 基金项目：国家自然科学基金资助项目(62166007)。
网络出版时间：2022-06-13 网络出版地址：http://kns.cnki.net/kcms/detail/61.1069.T.20220611.1612.002.html

中文分词是中文信息处理中的首要任务。与英语使用分隔符来分割单词不同,汉语是一种由本族语素(汉字)发展而成的多合成语言[1]。在汉语中,语素也可以独立成词。语素和复合词的模糊导致了汉语中对词的概念比较弱。另外,与英语使用分隔符来分割单词不同,汉语句子采用连续书写。词与词之间没有分隔符。因此,在中文文本里,经常存在分词歧义。单个句子会产生多种可能的切分路径。例如,“世界冠军”“抽象概念”“银行流水”等,这些词既可单独成词,又可以切分成粒度更小的多个词语,例如“世界/冠军”“抽象/概念”“银行/流水”等。

中文分词作为中文信息处理的第一步,直接用于支撑多种下游任务,如文本分类、机器翻译等。分词结果的不同将会对下游任务产生不同的影响。错误的分词结果会产生错误扩散,直接影响下游任务的性能。所以,中文分词是中文信息处理中一项重要任务。

现有的神经网络模型很难捕获句子中的长距离语义依赖,使得模型对文本语义特征理解不够充分,从而对文本中的歧义词边界识别性能较差。然而,在序列标注任务中,歧义词的边界样本又相对较少,存在样本不平衡问题。例如,“世界冠军”“抽象概念”等歧义词属于难分类样本,文本中存在着的歧义词与大量的非歧义词样本造成了难易样本不平衡。传统的序列标注模型(如 LSTM、CRF、Transformer 的双向编码器(BERT)等)在歧义词上识别性能都较差,不能很好地解决难易样本不平衡问题。尽管在解决中文分词歧义性上提出了各种解决方案,但仍然存在不足。目前,主流的中文分词模型主要采用序列标注模型。序列标注模型只依赖每个字周围的局部特征对字的分类标签进行预测。该模型存在输入特征使用不充分、难易样本中难分类样本得不到重点关注的问题,使得模型识别歧义词性能较差。

针对中文分词模型输入特征使用不充分、难易样本不平衡的问题,本文提出了基于机器阅读理解的中文分词模型。本文设计模型的动机是构建问题信息作为先验知识以丰富模型输入特征。针对中文词组的歧义性带来的难易样本不平衡问题,本文改进了损失函数。在 Bakeoff2005 语料库的 4 个公共数据集上进行验证,实验结果表明了本文方法的有效性。本文的主要贡献如下。

(1)采用基于机器阅读理解模型的方法,通过构建问题信息作为先验知识以丰富模型输入特征,增强模型对文本语义特征的理解,实现歧义词的更好识别。

(2)在充分分析中文词组特点的基础上,改进损失函数以缓解歧义词所带来的难易样本不平衡问题。

(3)本文首次将机器阅读理解模型应用于中文分词任务中,为中文分词提供了一种新思路。

1 相关工作

目前,主流的中文分词方法是基于神经网络模型。许多方法都将中文分词作为序列标注任务进行处理,然而这些方法都存在输入特征使用不充分和无法缓解难易样本不平衡问题。

随着深度神经网络模型不断发展[2],出现了许多应用于各项自然语言处理任务的神经网络模型[3]。Collobert 等[4]提出将神经网络的方法应用于序列标注任务。此后,许多方法相继应用于中文分词。Chen 等[5]提出用长短记忆(LSTM)神经网络来解决传统神经网络存在的长期依存关系的问题;Yao 等[6]在 Chen 等的基础上,提出双向长短记忆(Bi-LSTM)神经网络来充分利用上下文信息进行分词;Chen 等[7]提出带门结构的递归神经网络(GRNN)来保留上下文;Chen 等[8]使用对抗神经网络来使用多个语料库进行联合训练;Ma 等[9]在 Bi-LSTM 上引入预训练、dropout、超参数调参这 3 项深度学习技术,以简单的模型实现复杂模型的性能;Yang 等[10]利用外部知识提高中文分词的准确率;Gong 等[11]提出一个将每个标准分割成若干标准的 Swich-LSTM 结构;Zhou 等[12]引入多种汉字 Embedding 来增强语义;He 等[13]提出利用多标准进行中文分词;郭振鹏等[14]提出结合词典的 CNN-BiGRU-CRF 网络中文分词模型。

大规模预训练模型 BERT[15]和 ELMo[16]的出现刷新了 NLP 领域各项任务的记录。Diao 等[17]提出基于 BERT 的 N-gram 增强中文文本编码器,以方便识别出可能的词组合;Tian 等[18]提出基于双通道注意力机制的分词及词性标注模型;Tian 等[19]提出基于键值记忆神经网络的中文分词模型;Chen 等[20]提出在基于全局字符联机制的神经网络模型 GCA-FL,通过联邦学习的方式增强模型在中文分词上的性能。

以上模型尽管在公共数据集上取得了不错的效果,但还存在以下的不足:①传统的序列标注模型对文本语义特征使用不充分;②中文分词文本存在难易样本不平衡问题无法得到有效缓解。近年来,有

把序列标注任务转换成智能问答（QA）任务的趋势。Li 等[21]将实体识别任务转换成机器阅读理解（MRC）任务，每个实体类型 $R(x,y)$ 都能被参数化为带答案 y 的一个问题 $q(x)$；Li 等[22]将关系抽取任务转换为一个多回合的问答任务。此外，构建问题信息作为先验知识，能使输入特征更加丰富。然而，以上模型无法缓解难易样本不平衡问题。Lin 等[23]在目标检测中通过降低易分类样本的损失权重，从而更加关注难分类样本，能够有效缓解难易样本之间的不平衡；Liu 等[24]引入密度函数，在目标检测中既抑制了易分类样本损失权重，又不太过于关注难分类样本。

2　机器阅读理解模型

2.1　BERT 预训练模型

Vaswani 等[25]最早提出 Transformer 的模型架构。它能够更好地学习到句子当中单词与单词之间的联系，并完全依赖于自注意力机制来计算其输入和输出从而结合上下文语境来提高模型的性能。自注意力机制的公式为

$$Z = \text{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{1}$$

式中：$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$ 表示 3 个矩阵向量；d 为 $\boldsymbol{Q}$ 向量的维度；通过 softmax 对得到的分数归一化。由于 BERT 的目标是生成语言模型，只需要用到 Transformer 编码器的机制，所以对 Transformer 的解码器部分不再作过多叙述。

BERT 预训练模型中的 Embedding 层是由 3 种 Embedding 求和得到。其中，Token Embeddings 是词向量。Segment Embeddings 是用来区分两种句子，因为预训练不只做语言模型，还要做以两个句子为输入的分类任务。Position Embeddings 是用来表示句子中单词的位置。BERT 预训练模型通过 3 个 Embeddings 相加能更好地提取句子语义特征。

2.2　阅读理解分词标注

本文是在大规模预训练 BERT 模型上构建的机器阅读理解模型。给定一个输入句子 $X=\{x_1,x_2,\cdots,x_n\}$，其中，n 代表句子中第 n 个字，然后在 X 中发现每一个词组。首先，需要将数据集转换成（QUESTION，ANSWER，CONTEXT）的三元组形式，其中，QUESTION 表示问题生成模板，ANSWER 使用 $x_{\text{start,end}}$ 来表示在句子中词组的开始和结束下标，CONTEXT 为整个句子的文本。对于词组，产生一个问题 $q=\{q_1,q_2,\cdots,q_m\}$，其中 m 代表问题中第 m 个字。通过产生一个问题 q_y 就能获得一个三元组 $(q_y,x_{\text{start,end}},X)$，也就是先前定义的三元组（QUESTION，ANSWER，CONTEXT）。由于构建了关于词组先验知识的问题，生成问题的内容对最后的结果有一定影响。Li 等[22]采用基于规则的过程来构建问题。在本文中，采用问句和词定义的方式来构建问题。词定义表示为词概念的描述，它描述得尽可能通用、精准且没有歧义。两种问题的构建方式如表 1 所示。

表 1　问题的构建方式

Table 1　Question construction methods

方式	问题构建内容
问句	在这段话中哪些是词组？
词定义	词和短语（又称词组）的合称。词是最小的能够独立运用的语言单位，而短语是有多个词组成的整体。

问题内容构建的不同，与文本拼接输入模型时会带有不同的先验知识，从而对最后的预测结果产生一定的影响，如图 1 所示。本文给定文本“学生会组织义演活动，他马上从南京市长江大桥过来”。由于词定义构建的问题内容相较于问句式产生的输入特征更加丰富，使得对“学生会”“南京市长江大桥”上分词更加准确。

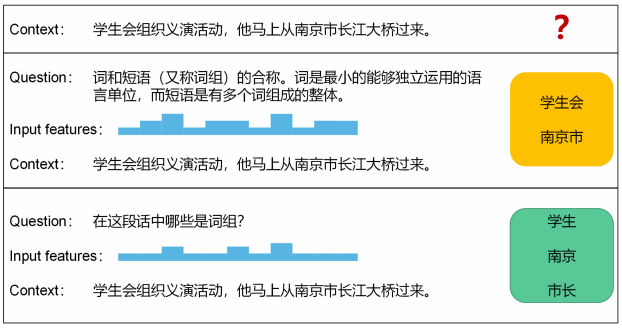


图 1　不同问题内容构建对分词结果的影响

Fig. 1　Influence of different question content constructions on word separation result

2.3　机器阅读理解

机器阅读理解分词模型结构如图 2 所示。在 BERT 预训练模型的基础上，加入已构建问题词组的先验知识，输入 BERT 编码器后得到隐藏层特征，最后通过解析特征输出结果。

输入包含了问题和文本两个部分，通过 BERT 预训练模型获得隐藏层表征矩阵

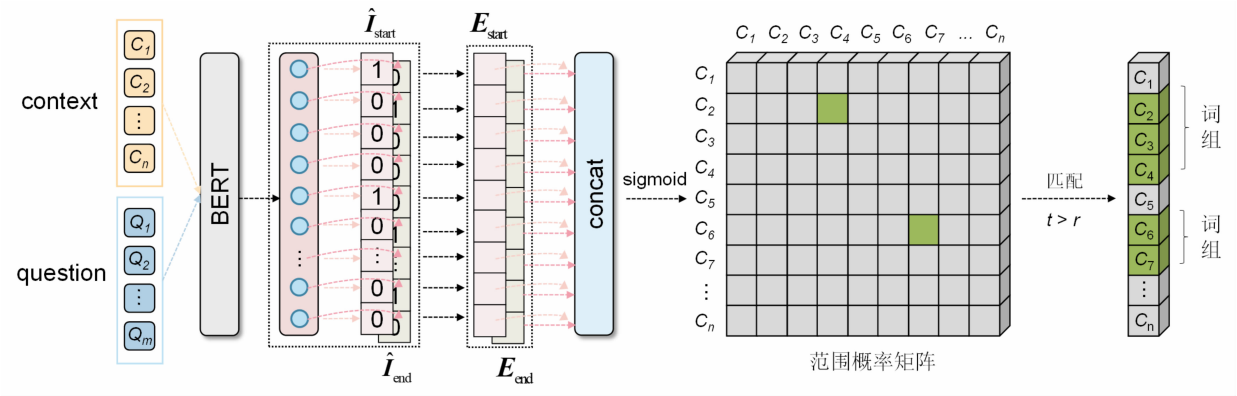


图 2 机器阅读理解分词模型

Fig. 2 Machine reading comprehension word separation models

$$\boldsymbol{E}=f(Q,C) \tag{2}$$

式中：f 为 BERT 编码函数；Q 为问题信息；C 为文本信息；$\boldsymbol{E}$ 为 BERT 编码器输出的表征矩阵。

通过多层感知向量机(MLP)[26]得到预测的词组索引。在 MLP 中，获得句子中每个字是词组的开始和结束索引的概率

$$\begin{cases}\boldsymbol{P}_{\text{start}}=\text{softmax}_{\text{each row}}(\boldsymbol{E}\boldsymbol{T}_{\text{strat}})\in\mathbb{R}^{n\times 2}\\ \boldsymbol{P}_{\text{end}}=\text{softmax}_{\text{each row}}(\boldsymbol{E}\boldsymbol{T}_{\text{end}})\in\mathbb{R}^{n\times 2}\end{cases} \tag{3}$$

式中 T_{start} 和 T_{end} 是学习权重。对 $\boldsymbol{P}_{\text{start}}$ 和 $\boldsymbol{P}_{\text{end}}$ 每一行使用 arg max 函数，得到预测的每个词组的开始和结束索引

$$\begin{cases}\hat{I}_{\text{start}}=\{i\mid\arg\max(P_{\text{start}}^{(i)})=1,i=1,2,\cdots,n\}\\ \hat{I}_{\text{end}}=\{j\mid\arg\max(P_{\text{end}}^{(i)})=1,j=1,2,\cdots,n\}\end{cases} \tag{4}$$

式中上标$^{(i)}$表示矩阵的第 i 行。给定任意开始索引 $i_{\text{start}}\in\hat{I}_{\text{start}}$ 和结束索引 $i_{\text{end}}\in\hat{I}_{\text{end}}$，训练一个二元分类器来预测 $x_{\text{start},\text{end}}$ 匹配的概率

$$P_{i_{\text{start}},j_{\text{end}}}=\text{sigmoid}(m\text{concat}(\boldsymbol{E}_{i_{\text{start}}},\boldsymbol{E}_{j_{\text{end}}})) \tag{5}$$

式中 m 是学习权重。将获得的结果合并得到范围概率分布矩阵

$$\begin{bmatrix}p_{00}&p_{01}&\cdots&p_{0j}\\ p_{10}&p_{11}&\cdots&p_{1j}\\ \vdots&\vdots&&\vdots\\ p_{i0}&p_{i1}&\cdots&p_{ij}\end{bmatrix}$$

式中 p_{ij} 表示句中索引 i 到索引 j 组成词组的概率。最后，通过人工设定阈值，输出匹配的词组结果。

2.4 改进损失函数

尽管机器阅读理解模型通过编码问题信息丰富了输入特征，但在数据集中存在着很多易分类样本和难分类样本。这使得难易样本之间存在不平衡，从而降低了分词的准确度。为了解决上述问题，本文改进了交叉熵损失函数

$$g(p,y)=\begin{cases}-\text{lb}p, & y=1\\ -\text{lb}(1-p), & \text{其他}\end{cases} \tag{6}$$

式中：$y\in\{-1,1\}$ 是一个真实类；$p\in[0,1]$ 是模型对标签为 $y=1$ 的类的估计概率。交叉熵函数在机器阅读理解模型使用为

$$\begin{cases}L_{\text{start}}=g(P_{\text{start}},Y_{\text{start}})\\ L_{\text{end}}=g(P_{\text{end}},Y_{\text{end}})\\ L_{\text{start,end}}=g(P_{\text{start,end}},Y_{\text{start,end}})\end{cases} \tag{7}$$

式中 $Y_{\text{start,end}}$ 表示每个起始索引的真实标签。总的损失函数为上述 3 个损失函数之和。然而，即使是容易识别的样本也会因为交叉熵损失而遭受非显著程度的损失。这些微小的损失值在大量容易识别的样本中汇总起来，可以淹没稀有类。通常，在样本不平衡问题上，普遍存在着的是正负样本不平衡，即正(负)例太多、负(正)例太少。一个解决正负类别不平衡的常用方法是为类别引入一个权重因子 $\alpha\in[0,1]$[27]。最后，权重之和重写为

$$L=\alpha_{\text{t}}L_{\text{start}}+\beta_{\text{t}}L_{\text{end}}+\gamma_{\text{t}}L_{\text{span}} \tag{8}$$

在本文实验中，训练机器阅读理解中文分词模型时会遇到普遍不平衡现象压倒了交叉熵损失。易分类样本占了损失值的大部分，并主导了梯度。尽管 α_{t}、β_{t}、γ_{t} 能够平衡正负样本不平衡，但是无法平衡难易样本。因此，需要降低易分类样本权重并关注难分类样本。本文借鉴目标检测中平衡正负难易样本的方法，对交叉熵函数引入一个平滑因子[23]

$$p_{\text{t}}=\begin{cases}(1-p)^{\theta}, & y=1\\ p^{\theta}, & \text{其他}\end{cases} \tag{9}$$

式中 $\theta\geqslant 0$ 是关注度参数。因此，定义新的损失函数

$$F(p,y)=p_{\text{t}}g(p,y) \tag{10}$$

在式(10)中可以通过参数 θ 来平滑地调整易分类样

本的损失权重。例如，在 $\theta=2$ 和样本概率 $p=0.9$ 时，可计算出与没有平滑因子相比，这个样本对损失的贡献权重降低为原来的 1%。在 $p=0.1$ 时，这个样本显然是难分类样本，计算出的平滑因子要比易分类样本要高，意味着模型在梯度更新过程中应该更加关注这个样本。合并式(7)～(10)，得到最终的损失函数

$$L=\alpha_t F(P_{\text{start}},Y_{\text{start}})+\beta_t F(P_{\text{end}},Y_{\text{end}})+\gamma_t F(P_{\text{start,end}},Y_{\text{start,end}}) \tag{11}$$

通过改进交叉熵函数缓解了难易样本间的不平衡问题。

3　实验与结果分析

3.1　数据集

进行基于机器阅读理解模型的中文分词任务，实验所用数据集来自 Bakeoff2005 语料库中的 4 个公共数据集 PKU、MSRA、CITYU、AS。因机器阅读理解任务不同于序列标注任务，需将原本的训练集和测试集转换成 MRC 所需格式(MRC 所用训练集和测试集均与原数据集相同)，转换后数据集样本数详细信息如表 2 所示，表中显示了未登录词(out of vocabulary，OOV)在测试集中的比例。

表 2　训练集、测试集和 OOV 样本数的统计信息

Table 2　Statistical information on the number of samples in the training sets，test sets and OOV

数据集	原训练集词数/10^6	原测试集词数/10^3	MRC 训练集样本数	MRC 测试集样本数	OOV 在测试集中比例/%
PKU	1.10	104	19 056	1 945	5.8
MSRA	2.37	107	86 909	3 979	2.7
CITYU	1.46	41	53 002	1 492	7.2
AS	5.45	122	698 122	14 361	4.3

3.2　评测指标

实验采用精准率 P、召回率 R、F_1 值为评测指标，其中主要以 F_1 值为主要评测指标。P、R、F_1 值的计算公式分别为

$$P=\frac{W_c}{W_a}\times 100\% \tag{12}$$

$$R=\frac{W_c}{W_t}\times 100\% \tag{13}$$

$$F_1=\frac{2PR}{P+R}\times 100\% \tag{14}$$

式中：W_c 为正确分词样本数；W_a 为全部样本数；W_t 为测试集中正确的样本数。

未登录词是指已知词典中不存在的新词，识别出未登录词也是评价中文分词模型性能优劣的重要指标之一。未登录词召回率计算公式为

$$V_{\text{recall}}=\frac{V(W_s\cap W_p)}{V(W_s)}\times 100\% \tag{15}$$

式中：W_s 为数据集中正确的分词答案；W_p 为模型预测分词的结果；$V(W_s)$ 为 W_s 中的词组未在词典中出现的词数。

3.3　超参数及训练设置

超参数的选择对模型训练结果有很大影响，超参数设计如下：优化算法使用 Adam，初始学习率为 0.000 05，以 0.05 速度进行衰减；每个 batch_size 为 16，Dropout 为 0.2，迭代 20 轮；概率分布矩阵阈值为 0.5。本文选择 BERT 中的 base 版本。

3.4　实验结果及分析

将本文模型与中文分词常用的经典模型 CRF、LSTM、ELMo、BERT 以及近年来其他中文分词模型等进行实验对比，结果如表 3 所示。

表 3　中文分词模型实验结果对比

Table 3　Experimental comparison of models of Chinese word separation

模型	F_1/%			
	PKU	MSRA	CITYU	AS
CRF	92.74	93.30	92.80	93.76
LSTM	93.30	95.34	94.00	94.18
Bi-LSTM	93.30	95.84	94.07	94.20
ELMo	95.25	96.52	95.43	95.03
BERT	96.51	97.43	96.62	95.94
AMCL[8]	94.32	96.04	95.55	94.75
SiBiLSTM[9]	95.40	97.50	95.70	95.50
TBNT[10]	96.30	97.50	96.90	95.70
Swich-LSTM[11]	96.20	97.80	96.20	95.20
ESMCL[13]	96.00	97.20	96.10	95.40
CNN-BiGRU-CRF[14]	**97.40**	97.50		
本文模型	96.64	**97.80**	**97.02**	**96.02**

注：表中加粗数据为最优值。

从表 3 可以看出，本文模型尽管在 PKU 数据集上效果略差，但与近年深度学习的中文分词模型相比还是取得了不错的结果。这主要缘于以下 3 点。一是本文模型在构建时区别于序列标注任务，将模型的构建分为 3 个步骤：首先，将序列标注数据集格式转换成机器阅读理解格式；其次，构建问题信息以丰富输入特征；最后，改进损失函数缓解难易样

本不平衡，从而提高模型的性能。二是问题内容构建上采用词定义的方式，比问句式所获得的输入特征要更加丰富。三是本文模型适用于中文分词中结构明确、词组边界清晰、存在歧义词等特点的特定领域数据。

为验证改进损失函数对于平衡难易样本的有效性，改变改进损失中的参数 θ 进行实验，结果如表 4 所示。

表 4 不同 θ 下的实验结果对比

Table 4 Experimental result comparison of different θ

θ	F_1/%			
	PKU	MSRA	CITYU	AS
0	95.18	96.91	96.32	94.98
1	95.39	97.21	96.20	95.32
2	**96.64**	**97.80**	**97.02**	**96.02**
5	95.29	96.78	96.29	95.10

注：表中加粗数据为最优值。

当 $\theta=0$ 时，交叉熵损失函数与改进的损失函数相等。从表 4 可以看出，相比于不加平滑因子，加入平滑因子后性能都有提升。在 PKU、MSRA、CITYU、AS 数据集上，F_1 分别提升了 1.46%、0.89%、0.7%、1.04%。式(9)以及实验结果表明，当 $\theta=1$ 时，模型对易分类样本的损失抑制和对难分类样本关注度变化较小，使得在缓解难易样本不平衡上效果较弱。$\theta=2$ 时模型性能最好。这是因为 $\theta=2$ 时，本文模型能够较好地抑制易分类样本损失和关注难分类样本。但是，当 $\theta=5$ 时，过度的抑制易分类样本损失和关注难分类样本，使得模型性能反而下降。这是因为过度抑制易分类样本损失反而会造成模型对易分类样本识别错误。若在模型已经收敛的情况下去过度关注那些非常难分类的样本，也会使模型产生误判。上述两种情况会导致模型的准确度降低。

未登录词是影响中文分词准确性的关键问题之一。为验证阅读理解分词模型在 OOV 上的表现，对比了阅读理解模型和表 3 中的经典模型在 OOV 上的性能，结果如表 5 所示。可以看出，本文方法在 OOV 识别效果上均优于经典模型。其原因在于：①大规模预训练模型 BERT 通过海量数据进行预训练，掌握更好的通用语言能力，下游任务只需微调即可获得优异性能；②OOV 中也包含歧义词和难分类样本。本文在预训练模型 BERT 基础上丰富输入特征和改进损失函数，在提高歧义词识别的基础上，也增强了 OOV 的识别。尽管如此，由于新词的不断出现，中文分词中 OOV 的识别仍具挑战性。

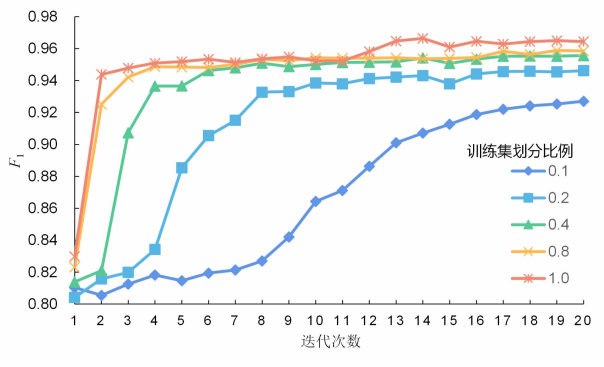
表 5 不同模型的 OOV 召回率实验结果对比

Table 5 Experimental comparison of OOV recall for different models

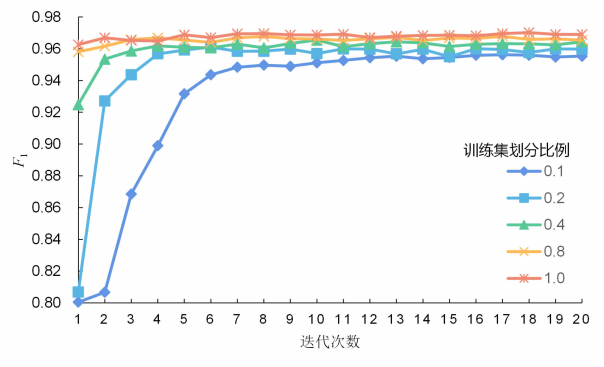
模型	V_{recall}/%			
	PKU	MSRA	CITYU	AS
CRF	62.80	61.30	63.10	58.90
LSTM	66.34	63.60	65.48	69.83
Bi-LSTM	66.09	66.28	65.79	70.70
ELMo	75.26	78.51	76.78	73.41
BERT	86.34	86.50	83.70	78.30
本文模型	**87.12**	**86.78**	**84.24**	**79.10**

注：表中加粗数据为最优值。

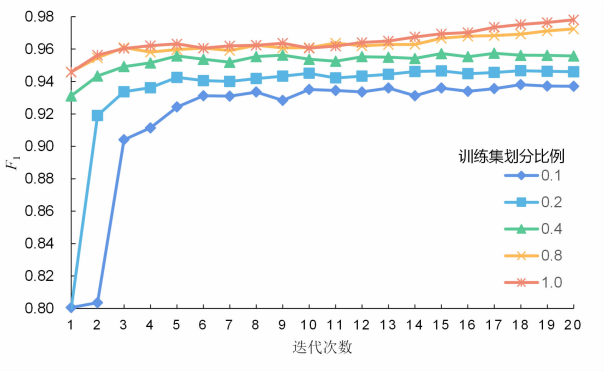
在对实验过程进一步分析后，发现本文方法在样本数较少的情况下也呈现出不错的结果。本文将 4 个公共数据集按 10%、20%、40%、80%、100%的比例划分训练集，测试集保持不变，实验结果如图 3 所示。可以看出：4 个公共数据集按比例划分训练集，送入模型训练 20 个 epoch 后，在测试集上得到的实验结果相差不大；随着训练集规模的增大，实验结果提升较小；本文提出的机器阅读理解模型能够在样本数较少的情况下，达到较好的中文分词结果。



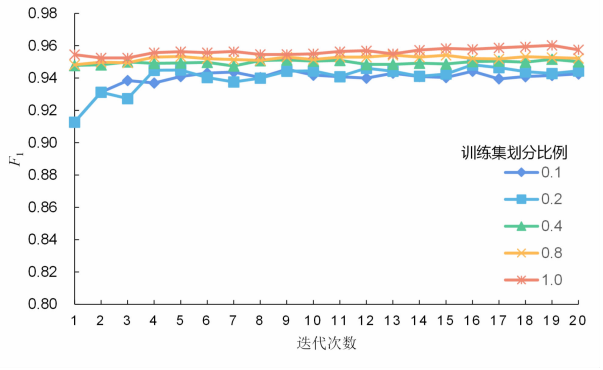
(a)PKU 数据集



(b)CITYU 数据集



(c)MSRA 数据集



(d)AS 数据集

图 3　数据集划分不同比例时的实验结果变化

Fig. 3　Variation in experimental results for different proportions of data set divisions

最后，对本文模型进行消融实验分析。表 6 对原始 MRC 方法[21]和本文方法进行了实验对比，并在问题信息构建上采用了表 1 中的两种方式。

从表 6 可以看出，改进损失函数和问题信息构建的不同会带来明显的提升。在 PKU、MSRA、CITYU、AS 数据集上，比原始 MRC 方法[21]分别提升了 1.52%、1.05%、0.91%、1.29%。由此说明对基础的 MRC 模型改进损失函数后，能更好地缓解难易样本的不平衡。在问题信息构建上，词定义的方式比问句的方式能带来更加丰富的特征，使得模型在 4 个数据集上都获得了一定的提升。

表 6　消融实验结果

Table 6　Results of ablation experiments

模型		问题		F_1/%			
MRC	$F(p,y)$	问句	词定义	PKU	MSRA	CITYU	AS
✓		✓		95.12	96.75	96.11	94.73
✓			✓	95.18	96.91	96.32	94.98
✓	✓	✓		96.42	97.54	96.73	95.97
✓	✓		✓	**96.64**	**97.80**	**97.02**	**96.02**

注：表中加粗数据为最优值。

综上所述可知，本文提出的基于机器阅读理解的中文分词方法可以有效解决中文分词领域的分词问题。

4　结　论

本文提出了一种机器阅读理解模型的中文分词方法，解决了序列标注模型很难获取句子长距离依赖导致输入特征使用不充分、边界样本少导致数据不平衡问题。本文将序列标注任务转换为机器阅读理解任务并改进损失函数，进而有效地增强输入特征使用和缓解数据不平衡。实验结果表明，本文提出的方法相比于序列标注模型的中文分词方法具有明显优势。

本文是机器阅读理解模型在中文分词上的初步探索，该方法还有进一步改进的空间。在下一步工作中，可以使用不同的预训练模型和改进注意力机制来更好地捕获上下文信息。通过探索新的模型架构、设计新的问题构建方式，进一步提升机器阅读理解模型在中文分词上的应用。

参考文献：

[1] GONG Chen, LI Zhenghua, ZHANG Min, et al. Multi-grained Chinese word segmentation [C]// Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA, USA: ACL, 2017: 692-703.

[2] 马海辉，余小玲，吕倩，等. 一维卷积神经网络在往复式压缩机气阀故障诊断中的应用 [J]. 西安交通大学学报，2022，56(4)：101-108.
MA Haihui, YU Xiaoling, LÜ Qian, et al. Application of one-dimensional convolutional neural network in fault diagnosis of reciprocating compressor air valve [J]. Journal of Xi'an Jiaotong University, 2022, 56(4): 101-108.

[3] 方军，管业鹏. 基于双编码器的会话型推荐模型 [J]. 西安交通大学学报，2021，55(8)：166-174.
FANG Jun, GUAN Yepeng. A session recommendation model based on dual encoders [J]. Journal of Xi'an Jiaotong University, 2021, 55(8): 166-174.

[4] COLLOBERT R, WESTON J, BOTTOU L, et al. Natural language processing (almost) from scratch [J]. The Journal of Machine Learning Research, 2011, 12: 2493-2537.

[5] CHEN Xinchi, QIU Xipeng, ZHU Chenxi, et al. Long Short-term memory neural networks for Chinese word segmentation [C]// Proceedings of the 2015 Con-

ference on Empirical Methods in Natural Language Processing. Stroudsburg, PA, USA: ACL, 2015: 1197-1206.

[6] YAO Yushi, HUANG Zheng. Bi-directional LSTM recurrent neural network for Chinese word segmentation [C]// ICONIP 2016: Neural Information Processing. Cham, Germany: Springer International Publishing, 2016: 345-353.

[7] CHEN Xinchi, QIU Xipeng, ZHU Chenxi, et al. Gated recursive neural network for Chinese word segmentation [C]// Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. Stroudsburg, PA, USA: ACL, 2015: 1744-1753.

[8] CHEN Xinchi, SHI Zhan, QIU Xipeng, et al. Adversarial multi-criteria learning for Chinese word segmentation [C]// Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA, USA: ACL, 2017: 1193-1203.

[9] MA Ji, GANCHEV K, WEISS D. State-of-the-art Chinese word segmentation with Bi-LSTMs [C]// Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA, USA: ACL, 2018: 4902-4908.

[10] YANG Jie, ZHANG Yue, DONG Fei. Neural word segmentation with rich pretraining [C]// Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA, USA: ACL, 2017: 839-849.

[11] GONG Jingjing, CHEN Xinchi, GUI Tao, et al. Switch-LSTMs for multi-criteria Chinese word segmentation [C] // Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence. Palo Alto, CA, USA: AAAI Press, 2019: 6457-6464.

[12] ZHOU Jianing, WANG Jianing, LIU Gongshen. Multiple character embeddings for Chinese word segmentation [C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop. Stroudsburg, PA, USA: ACL, 2019: 210-216.

[13] HE Han, WU Lei, YAN Hua, et al. Effective neural solution for multi-criteria word segmentation [M] // Smart Intelligent Computing and Applications. Singapore: Springer Singapore, 2019: 133-142.

[14] 郭振鹏，张起贵. 基于结合词典的 CNN-BiGRU-CRF 网络中文分词研究 [J]. 电子设计工程，2021，29(16)：64-69，74.
GUO Zhenpeng, ZHANG Qigui. Research on Chinese segmentation of CNN-BiGRU-CRF network based on combined dictionary [J]. Electronic Design Engineering, 2021, 29(16): 64-69, 74.

[15] DEVLIN J, CHANG Mingwei, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding [C] // Proceedings of NAACL-HLT 2019. Stroudsburg, PA, USA: ACL, 2019: 4171-4186.

[16] PETERS M E, NEUMANN M, IYYER M, et al. Deep contextualized word representations [C] // Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, PA, USA: ACL, 2018: 2227-2237.

[17] DIAO Shizhe, BAI Jiaxin, SONG Yan, et al. ZEN: pre-training Chinese text encoder enhanced by N-gram representations [C]// Findings of the Association for Computational Linguistics: EMNLP 2020. Stroudsburg, PA, USA: ACL, 2020: 4729-4740.

[18] TIAN Yuanhe, SONG Yan, AO Xiang, et al. Joint Chinese word segmentation and part-of-speech tagging via two-way attentions of auto-analyzed knowledge [C] // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA, USA: ACL, 2020: 8286-8296.

[19] TIAN Yuanhe, SONG Yan, XIA Fei, et al. Improving Chinese word segmentation with wordhood memory networks [C] // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA, USA: ACL, 2020: 8274-8285.

[20] TIAN Yuanhe, CHEN Guimin, QIN Han, et al. Federated Chinese word segmentation with global character associations [C]// Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Stroudsburg, PA, USA: ACL, 2021: 4306-4313.

[21] LI Xiaoya, FENG Jingrong, MENG Yuxian, et al. A unified MRC framework for named entity recognition [C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA, USA: ACL, 2020: 5849-5859.

[22] LI Xiaoya, YIN Fan, SUN Zijun, et al. Entity-relation extraction as multi-turn question answering [C]// Proceedings of the 57th Annual Meeting of the Associ-

ation for Computational Linguistics. Stroudsburg, PA, USA: ACL, 2019: 1340-1350.

[23] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection [C]// 2017 IEEE International Conference on Computer Vision (ICCV). Piscataway, NJ, USA: IEEE, 2017: 2999-3007.

[24] LI Buyu, LIU Yu, WANG Xiaogang. Gradient harmonized single-stage detector [C]// Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence. Palo Alto, CA, USA: AAAI Press, 2019: 8577-8584.

[25] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc., 2017: 6000-6010.

[26] GARDNER M W, DORLING S R. Artificial neural networks (the multilayer perceptron): a review of applications in the atmospheric sciences [J]. Atmospheric Environment, 1998, 32(14/15): 2627-2636.

[27] XIE Saining, TU Zhuowen. Holistically-nested edge detection [C]// 2015 IEEE International Conference on Computer Vision (ICCV). Piscataway, NJ, USA: IEEE, 2015: 1395-1403.

（编辑　陶晴　亢列梅）