计算机科学
COMPUTER SCIENCE

基于信息熵-切分概率模型的新词发现方法

祝钰莹，郭燕，万亿兆，田凯

引用本文

祝钰莹，郭燕，万亿兆，田凯.基于信息熵-切分概率模型的新词发现方法[J]．计算机科学, 2023, 50(7): 221-228.

ZHU Yuying, GUO Yan, WAN Yizhao, TIAN Kai. New Word Detection Based on Branch Entropy-Segmentation Probability Model [J]. Computer Science, 2023, 50(7): 221-228.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.220700074

基于信息熵-切分概率模型的新词发现方法

祝钰莹[1,2]　郭　燕[1,2]　万亿兆[2]　田　凯[2]
1 中国科学技术大学苏州高等研究院　江苏 苏州 215123
2 中国科学技术大学软件学院　江苏 苏州 215123
(hiyazyy@mail.ustc.edu.cn)

摘　要　新词发现是中文自然语言处理的基本任务,对于提升各种下游任务的性能至关重要。文中提出了一种基于信息熵-切分概率模型的新词发现方法,该方法首先基于信息熵从待处理文本中生成候选词集,然后对候选词集进行切分概率计算,从而筛选出真正的新词。针对有无待处理文本相关的标注语料,提出了两种不同的模型。在缺少待处理文本相关标注语料的情况下,使用通用的分词基准数据集训练了多标签 Transformer-CRF 模型;在具有专业标注语料的情况下,则引入了基于键值的记忆神经网络,以充分融合词语成词信息。实验结果表明,多标签 Transformer-CRF 模型在 Top900 词中法律相关词的 MAP 高达 54.00%,较无监督方法生成的候选词集提升了 2.15%;在对法律专业语料提取新词时,键值记忆神经网络的表现进一步超过了多标签 Transformer-CRF 模型,达到了 3.43%的效果提升。

关键词:新词发现;信息熵;互信息;Transformer;条件随机场;键值记忆神经网络
中图法分类号　TP391

New Word Detection Based on Branch Entropy-Segmentation Probability Model

ZHU Yuying[1,2],GUO Yan[1,2],WAN Yizhao[2] and TIAN Kai[2]
1 Suzhou Institute for Advanced Research,University of Science and Technology of China,Suzhou,Jiangsu 215123,China
2 School of Software Engineering,University of Science and Technology of China,Suzhou,Jiangsu 215123,China

Abstract　As a basic task of Chinese natural language processing,new word detection is crucial for improving the performance of various downstream tasks. This paper proposes a new word detection method based on branch entropy and segmentation probability. The method firstly generates a candidate word set from the text based on branch entropy,and then calculates the segmentation probability of each candidate,so as to filter out the noisy word. Two different models are proposed to respectively deal with situations whether or not there are annotated corpus related to the text to be processed. In the absence of related segmented corpus,the multi-criteria Transformer-CRF model is trained using general segmented benchmark data sets. A key-value based memory neural network is introduced to fully extract the wordhood information if there is field-specific segmented corpus. Experimental results show that the multi-criteria Transformer-CRF model has a MAP of 54.00% of legal texts in the top 900 resulted words,which is 2.15% higher than that of the unsupervised method. As with segmented legal corpus,the performance of the key-value memory neural network further exceeds the former model,has an improvement of 3.43%.

Keywords　New word detection,Branch entropy,Mutual information,Transformer,Conditional random fields,Key-value memory neural networks

1　引言

随着现代科技和互联网的快速发展,“专有术语”“流行词汇”等大量涌现。新词发现是自然语言处理中的一项基本任务,它可以和很多其他任务相结合。譬如,将新词加入分词词典,可以提高分词准确率;将新词和纠错任务相结合,可以避免新词导致的误判;同时也是提升很多下游任务,如情感分析[1]、热点事件提取[2]、文本倾向性分析[3]性能的基础。因此,新词发现对于中文文本 NLP 任务具有非常重要的意义。另一方面,新词产生速度快,构成规则多样化,很难使用模板规则或者统计特征进行匹配,因此亟需不依靠先验知识和人工处理,就能完成从一定规模文本中自动抽取出新词的新词发现方法。

传统的无监督新词发现,如信息熵、互信息、邻接变化数等都是基于频率的算法。文本预处理阶段进行 n-gram 分割产生候选词串时,不可避免地包含高频的噪声词串,如“之间的”“哄抬物价”等。而基于频率的算法很难排除此类噪声词串,因此需要进一步过滤无监督方法生成的候选词,以得到更准确的新词结果。

针对以上两点需求,本文提出了基于信息熵-切分概率

到稿日期:2022-07-07　返修日期:2022-10-23
通信作者:郭燕(guoyan@ustc.edu.cn)

模型的新词发现方法。一方面，无监督方法可以从大规模原始语料中自动抽取数量可观的候选词；另一方面，本文使用神经网络模型对候选词上下文进行切分概率预测，将新词发现转化为预测新词边界是否有效，从而充分过滤“哄抬物价”等噪声词串。针对有无专业领域标注语料，我们提出了两种不同的模型。在缺少相关标注语料的情况下，使用通用分词基准数据集训练，在具有专业标注语料的情况下，则引入了可以融合词语成词信息的记忆神经网络。

本文的主要贡献包括3个方面：

(1)提出了通用的基于中文分词基准数据集SIGHAN Bakeoff训练的多标签Transformer-CRF切分概率模型，模型在Top900词中法律相关词的MAP高达54.00%，较候选词集提升了2.15%。

(2)爬取了大规模法律文本，并对其中部分文本进行人工标注，形成了法律领域分词数据集。

(3)基于该专业分词数据集，训练了键值记忆网络切分概率模型，以充分利用领域的成词特征；在法律文本上，该模型的新词发现效果优于多标签Transformer-CRF切分概率模型，MAP值达到了3.43%的提升。

本文第2节总结了新词发现的相关工作；第3节对本文方法的总体流程和模型结构进行具体阐述；第4节分别给出了两个切分概率模型的实验验证及结果评估；最后总结全文并展望未来。

2 相关工作

新词发现方法可以大致分为两类：基于规则的方法和基于统计的方法。

基于规则的方法的关键点在于挖掘新词的构词特征、词性特点或语义信息，从而建立规则库、模式库或专业词库，继而以规则匹配的方式给出潜在的新词。Zhao等[4]基于依存句法分析和TF-IDF，构建了领域句法词典，然后用词向量计算候选新词与词典中已登录词的相似度，从而完成领域新词的判定。基于规则的方法优势在于识别精准，缺点在于规则总结困难且迁移能力差。

基于统计的方法利用统计策略来提取候选新词。第一类基于字标注，通过给句子中每个字打上标签来进行切分，将分词和新词发现统一为序列标注和分类问题，这类方法往往需要进行大规模的分词标注。统计机器学习的方法可以很好地对数据进行标注和分类，常用模型有隐马尔可夫、最大熵、条件随机场和支持向量机。Liu等[5]提出了一种基于古文语料的新词识别方法AP-LSTM-CRF。首先通过改进的类Apriori算法产生候选词，然后训练LSTM-CRF切分概率模型，在宋词和史记数据集上F1值分别达到了89.68%和81.13%。第二类为无监督的基于频繁模式的方法。它无需任何先验知识，也无需经过分词预处理的数据，而是使用预定义的标准来预测候选词串能否称为合理词汇。传统计量标准有信息熵[6]、点间互信息[7]、邻接变化数[8]、多词表达距离[9]等。此外，也有一些无监督模型，如Deng等[10]提出的TopWords。在其基础上，Chen等[11]提出了D-TopWords模型来进一步抽取专业领域中出现较多或意义不同的领域新词；Pan等[12]提出了基于贝叶斯推理的TopWORDS-Seg模型，可以在开放域中同时实现有效的文本分词和新词发现。此类方法的不足之处在于易产生大量噪声词串，需要进一步过滤。因此，本文将在无监督方法的基础上，设计神经网络模型对候选词上下文进行切分概率预测，以完成噪声词串的过滤。

除新词发现外，本文相关工作包括如何更有效地利用通用标准分词语料以及如何将成词特征融入神经网络中。为了充分利用上下文特征，Tian等[13]提出了神经网络框架WM-SEG，使用记忆网络将成词信息融入分词网络中。针对多标准分词语料问题，Qiu等[14]提出多标签统一模型，基于Transformer编码器和多个分词标准标签，进一步提升了每个标准下的分词性能。

3 方法描述

3.1 问题描述

本文要解决的是如何有效地从大规模原始语料中得到候选词集并过滤噪声词串，发现语料相关的有价值的词。本文引入通用词典D_c，基于新词为通用领域不常出现的词汇，给出定义如下。

定义1(语料新词) 对大规模原始语料L，记其预处理后的有效文本集合为T_L，语料新词为T_L中出现的所有语义独立且有意义的词汇集合W，大小为n，具体可表示为$\{w_1, w_2, \cdots, w_i, \cdots, w_n\}$，其中$w_i \notin D_c$。

从另一个角度来说，若第一阶段无监督新词发现的结果为候选词集W_{candi}，第二阶段切分概率模型过滤掉的噪声词串集为W_{junk}，则最终结果W即可表示为$W_{candi} - W_{junk} - D_c$。

3.2 总体框架

基于信息熵-切分概率模型的新词发现方法的总体框架分为3部分：候选词集的获取、多标签Transformer-CRF切分概率模型，以及键值记忆网络切分概率模型。具体描述如图1所示。

(1)候选词集的获取

在文本集合上进行n-gram切割，用基于信息熵和互信息的无监督方法提取大量候选新词，组成候选词集。此阶段构建的候选词集即是下一阶段切分概率模型的基础，对于其中的每个候选新词，找到文本集合中对应出现的一定量的上下文句子，送入训练好的切分概率模型，并结合过滤规则进行判断，以决定是否保留。

(2)多标签Transformer-CRF切分概率模型

当缺乏与待处理文本相关的专业领域标注数据集时，用中文分词数据集Bakeoff2005/2008对Transformer-CRF模型进行训练。由于该基准数据集包括简繁中文，并且对应的标注规则并不兼容，本文使用多标签的方式在输入语料中加入语料相关的标识符，并联合多个不同分词标准的基准数据集进行训练，从而使得共享编码器可以学习到分词标准敏感的语境特征。

(3)键值记忆网络切分概率模型

在具有待处理文本相关的专业领域标注数据集时，为充分利用文本的上下文信息，特别是领域相关的成词信息，本文引入了记忆网络切分概率模型，使用键值信息，将字符的成词

信息加入神经网络中,使得模型可以学习到更有效的领域相关词信息,更好地完成噪声词串的过滤。

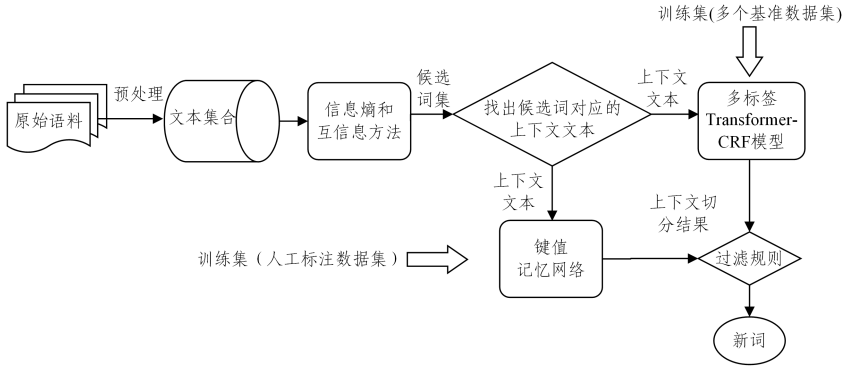

图 1 总体框架图

Fig. 1 Overall framework

3.3 多标签 Transformer-CRF 切分概率模型

图 2 给出了多标签 Transformer-CRF 切分概率模型的具体结构。模型将上下文句子加上分词标准对应的标签作为输入,每个输入句子将被映射成向量。根据文献[14],本文使用了字符级别的向量、Bigram 级别的向量和位置向量 3 种向量。处理后的向量会被送入 Transformer 编码器,提取每个输入字符的上下文特征。编码器的输出进入 CRF 解码器,用于边界预测,最终将输出句子中所有对应位置的切分概率,并按照预设的候选词中每个内部位置的切分概率和两边的外部切分概率,完成无监督方法生成的候选词集的过滤操作。下文将分别介绍该神经网络的向量层、编码器和解码器。

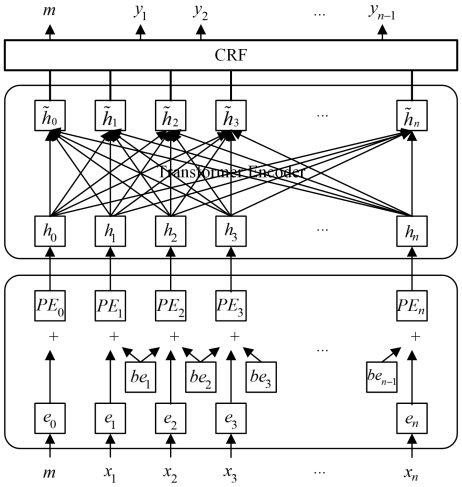

图 2 多标签 Transformer-CRF 切分概率模型结构图

Fig. 2 Structure of multi-label Transformer-CRF segmentation probability model

3.3.1 向量层

在神经网络模型中,首先需要将离散的语言符号映射到分布式向量空间中。本节的向量层分 3 部分:字向量、Bigram 向量和位置向量。

对输入句子 $X=x_1x_2\cdots x_n$,记 m 为其语料库标签,如〈PKU〉,模型尝试在每个句子开头位置添加语料库标签来学习不同标签对应的信息。经过字符级别的映射,输入句子中每个位置都会被映射到一个 d_{model} 维的字向量,具体如式(1)所示:

$$X\rightarrow[\boldsymbol{e}_m,\boldsymbol{e}_{x_1},\boldsymbol{e}_{x_2},\cdots,\boldsymbol{e}_{x_n}] \tag{1}$$

其中$\boldsymbol{e}_m$ 为标签向量,$\boldsymbol{e}_{x_i}$ 为句子位置 i 处的字向量。为了进一步提升切分效果,本文引入 Bigram 向量。输入句子 X 中位置 i 处的字符 x_i 对应的向量由字符特征和 Bigram 特征共同计算得到,具体描述如式(5)所示:

$$\boldsymbol{E}_{x_i}=FC(\boldsymbol{e}_{x_i}\oplus\boldsymbol{e}_{x_{i-1}x_i}\oplus\boldsymbol{e}_{x_ix_{i+1}}) \tag{2}$$

其中,$\boldsymbol{E}_{x_i}$ 维数为 d_{model},由字向量$\boldsymbol{e}_{x_i}$ 与包含字符 x_i 的前后两个 Bigram 对应的向量$\boldsymbol{e}_{x_{i-1}x_i}$ 和$\boldsymbol{e}_{x_ix_{i+1}}$ 进行级联后接入一个全连接层得到。

为了学习文本中的位置信息,此处引入位置向量 $\boldsymbol{PE}$。对句子中位置 i 处的字符,以及位置向量的维数 k,其位置向量被定义为:

$$\boldsymbol{PE}_{i,2k}=\sin(i/10\,000^{2k/d_{\text{model}}}) \tag{3}$$

$$\boldsymbol{PE}_{i,2k+1}=\cos(i/10\,000^{2k/d_{\text{model}}}) \tag{4}$$

综上,对于长度为 n 的输入句子 X,其向量长度为 $n+1$,维数为 d_{model},最终得到的输入矩阵 $\boldsymbol{H}\in\boldsymbol{R}^{(n+1)\times d_{\text{model}}}$,具体描述为:

$$[\boldsymbol{e}_m+\boldsymbol{PE}_0,\boldsymbol{E}_{x_1}+\boldsymbol{PE}_1,\cdots,\boldsymbol{E}_{x_n}+\boldsymbol{PE}_n] \tag{5}$$

3.3.2 编码器和解码器

所有不同分词标准的语料将共享同一个编码器和解码器,由于语料库标签的存在,编码器和解码器仍可有效提取到分词标准相关的信息。

Transformer 可以解决循环神经网络由于序列的固有顺序约束无法并行化而产生的效率问题,以及位置距离、传递缺陷造成的远程依赖[15]。编码器接收向量层的输出阵 $\boldsymbol{H}$,其中每个向量会被投影到对应的 ***query***,***key***,***value*** 这 3 个向量,生成方法为分别乘以 3 个权重矩阵,如下式所示:

$$\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}=\boldsymbol{HW}^Q,\boldsymbol{HW}^K,\boldsymbol{HW}^V \tag{6}$$

$$Atten(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\text{softmax}\left(\frac{\boldsymbol{QK}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{7}$$

其中,$\boldsymbol{Q},\boldsymbol{K}\in\boldsymbol{R}^{(n+1)\times d_k}$,$\boldsymbol{V}\in\boldsymbol{R}^{(n+1)\times d_v}$,权重矩阵$\boldsymbol{W}^{\boldsymbol{Q}},\boldsymbol{W}^{\boldsymbol{K}}\in\boldsymbol{R}^{d_{\text{model}}\times d_k}$,$\boldsymbol{W}^{\boldsymbol{V}}\in\boldsymbol{R}^{d_{\text{model}}\times d_v}$,为可训练的参数。多头注意力机制则通过随机初始化多组权重矩阵将输入向量映射到不同的注意力子空间中,单独计算后进行拼接,具体如下:

$$MH(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=Concat(\boldsymbol{h}_1,\cdots,\boldsymbol{h}_h)\boldsymbol{W}^O \tag{8}$$

$$\boldsymbol{h}_i=Atten(\boldsymbol{HW}_i^Q,\boldsymbol{HW}_i^K,\boldsymbol{HW}_i^V) \tag{9}$$

其中,参数矩阵$\boldsymbol{W}^O\in\boldsymbol{R}^{hd_v\times d_{\text{model}}}$,$h$ 为注意力头数,$\boldsymbol{W}_i^{\boldsymbol{Q}},\boldsymbol{W}_i^{\boldsymbol{K}}\in$

$\boldsymbol{R}^{d_{\text{model}}\times d_k}$，$\boldsymbol{W}_i^V \in R^{d_{\text{model}}\times d_v}$。

编码器中的前馈神经网络模块由 ReLU 激活函数加一个线性变换组成，如式(10)所示：

$$FFN(x)=\max(0,x\boldsymbol{W}_1+b_1)\boldsymbol{W}_2+b_2 \tag{10}$$

编码器每个子模块后都有一个 Add&Norm 模块，Add 表示残差连接，Norm 表示归一化。记 Sublayer 为多头注意力机制模块或前馈神经网络模块，则该模块的输出为：

$$LayerNorm(x+Sublayer(x)) \tag{11}$$

编码器部分会输出每个字的所有对应类别的分数，以 BMES 4 标注为例，输出 S,B,M,E 各自对应的分数，而这部分输出将作为解码器的输入。考虑到模型输出标签之间的依赖关系，譬如 S 标签后只会出现 S 和 B，本文使用 CRF 作为解码器。记 $Y=y_1y_2\cdots y_n$ 为对应的标签序列，为了得到最佳标签序列 Y^*，需要最大化标签序列得分，如下式所示：

$$s(Y)=\sum_{i=2}^{n}\boldsymbol{A}_{y_{i-1}y_i}+\sum_{i=1}^{n}\boldsymbol{P}_i[y_i] \tag{12}$$

$$Y^*=\arg\max_{Y\in Set(Y)} s(Y) \tag{13}$$

其中，l 为标签数，$\boldsymbol{A}\in\boldsymbol{R}^{l\times l}$ 为转移分数矩阵，记录从上一个标签 y_{i-1} 转移到当前标签 y_i 的分数；$\boldsymbol{P}\in\boldsymbol{R}^{n\times l}$ 为编码器输出的标签分数矩阵，$\boldsymbol{P}_i[y_i]$对应句子中第 i 个字标注为 y_i 的分数；$Set(Y)$表示输入句子对应的所有标签序列的集合。

3.4 键值记忆网络切分概率模型

若有与待处理文本相关的标注数据集，则本文采用键值记忆网络切分概率模型，结构如图 3 所示。以“尖沙嘴”为例，模型将该候选词上下文句子送入编码器，之后整合键值记忆网络中来自不同度量方式的词汇信息，由解码器进行标签预测并输出每个对应位置的切分概率，继而根据过滤规则来决定是否保留该词。

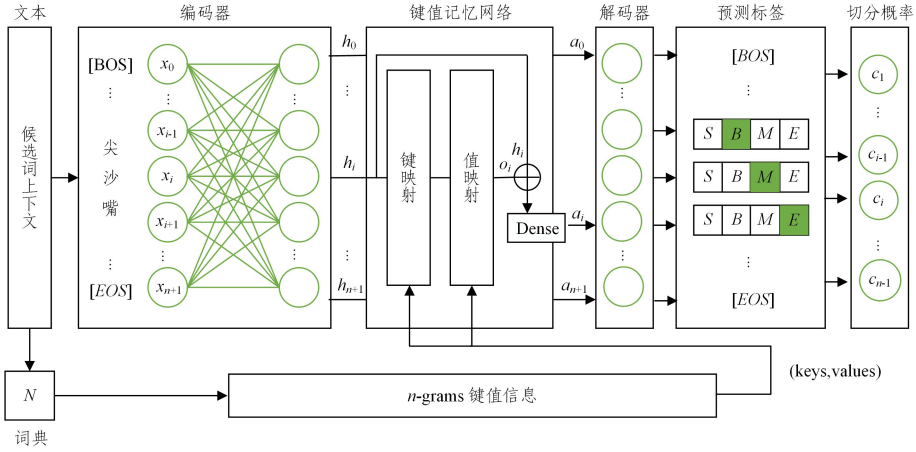

图 3 键值记网络切分概率模型结构图

Fig. 3 Structure of key value memory network segmentation probability model

由于键值记忆网络能够学习到词汇度量方式中蕴含的 n-gram 统计特征，从而更好地挖掘该位置字与其他字的关系，并计算得到使整体语义更完整的切分方式，因此本文在切分概率模型的编码器和解码器间使用类似的键值记忆网络结构。不同于文献[14]中的单标准度量方式，如 AV，PMI，DLG，本文尝试将 AV 与 PMI 结合的双标准度量方式。下文将分别介绍各词汇度量方式和网络。

3.4.1 邻接变化数

邻接变化数是一种用于衡量文本片段在语境中使用自由程度的度量方式。对文本片段 $x_{i,j}$，其邻接变化数 $AV(x_{i,j})$ 的计算方式为：

$$AV(x_{i,j})=\min(L_{AV}(x_{i,j}),R_{AV}(x_{i,j})) \tag{14}$$

其中，$L_{AV}(x_{i,j})$是候选 n-gram $x_{i,j}$ 的左邻接变化数，定义为不同左邻接字的数目加上出现在句首的次数；$R_{AV}(x_{i,j})$为右邻接变化数，即不同右邻接字的数目加上出现在句尾的次数。邻接变化数越大，文本片段成词概率就越高。

3.4.2 信息熵

信息熵可以衡量一个文本片段的左邻接字集合和右邻接字集合的随机程度，即文本片段的上下文丰富性。计算式如下：

$$h(x_{i,j})=-\sum_{x\in V}p(x|x_{i,j})\log p(x|x_{i,j}) \tag{15}$$

其中，$x_{i,j}$是一个候选 n-gram，V 是 $x_{i,j}$ 的邻接字集，x 是其邻接字，$p(x|x_{i,j})$为 x 和 $x_{i,j}$ 的共现概率。由于邻接字有左或右两种情况，因此 h 可分为 h_L 和 h_R，即左熵和右熵。通过对语料中所有 n-gram 计算其 h_L 和 h_R，然后将 $\min(h_L,h_R)$与阈值进行比较，以确定是否将其作为固定搭配。

3.4.3 互信息

文本片段内的凝固程度通过点间互信息进行衡量，对 2-gram xy，其点间互信息被定义为：

$$PMI(x,y)=\log\frac{p(xy)}{p(x)p(y)} \tag{16}$$

其中，$p(xy)$是 x 和 y 的共现概率，$p(x)$和 $p(y)$分别是 x 和 y 的独立概率。对于概率计算，本文遵循文献[7]的设置，通过将其频率 $f(x)$(语料库中的出现次数)除以语料库字符总数来估计任何 n-gram 中 x 的出现概率 $p(x)$。从该定义中可以发现，互信息值越大，文本片段的内部凝固程度越高，成词可能性就越大。

若 $n>2$，则需要将 n-gram 进行 $n-1$ 次切分，将其分成两部分，计算得到所有切分可能的 PMI 的最小值，即该 n-gram 的 PMI。以“ABC”为例：

$$PMI(\text{“}ABC\text{”})=\min\left\{\log\frac{p(\text{“}ABC\text{”})}{p(\text{“}A\text{”})p(\text{“}BC\text{”})},\ \log\frac{p(\text{“}ABC\text{”})}{p(\text{“}AB\text{”})p(\text{“}C\text{”})}\right\} \tag{17}$$

3.4.4 键值记忆网络

对一个具体句子来说，N 其实是一个包含了句子中可能 n-gram 的词表，这个词表通过 AV 与 PMI 结合的双标准度量方式构建。此种方式能够将词汇信息合并到通用序列标注框架中，并通过键值转化对这种成对的知识进行建模。譬如，句子“九铁尖沙嘴支线”构建的 N 如下：

{九，铁，尖，沙，嘴，九铁，沙嘴，支线，九铁尖，尖沙嘴，九铁尖沙嘴}

首先是键映射。记忆网络的键对应 N 中的 n-gram，对句子的每一个字构建词表 N，记第 i 个字 x_i 对应的 n-gram 集 $K_i=\{k_{i,1},k_{i,2},\cdots,k_{i,m_i}\}$。比如对例句“九铁尖沙嘴支线”中的第四个字“沙”构建词表，则其内容可表示为 $K_4=\{$沙，沙嘴，尖沙嘴，九铁尖沙嘴$\}$，包含 4 个 n-gram，m_i 为 4。“尖沙嘴”即可表示为 $k_{4,3}$。将这些 n-gram 的向量表示与编码器传来的 $\boldsymbol{h}_i$ 相乘后做 softmax 得到概率分布。记 $\boldsymbol{e}^k_{i,j}$ 是 K_i 中 $k_{i,j}$ 对应的词向量，概率大小 $p_{i,j}$ 可描述为：

$$p_{i,j}=\frac{\exp(\boldsymbol{h}_i\cdot\boldsymbol{e}^k_{i,j})}{\sum_{j=1}^{m_i}\exp(\boldsymbol{h}_i\cdot\boldsymbol{e}^k_{i,j})} \tag{18}$$

值映射需要将每个 $k_{i,j}$ 对应映射到一个值，由于每个字在不同的 n-gram 中的位置不同，模型使用 BMES 4 标注，如“沙嘴”中的“沙”在词首，则对应 V_B；“尖沙嘴”中的“沙”在词中，则对应 V_M。$K_4=\{$沙，沙嘴，尖沙嘴，九铁尖沙嘴$\}$中“沙”字对应的位置标签依次为 S,B,M,M，则 $V_4=\{V_S,V_B,V_M,V_M\}$。这样，每个 K_i 都有对应值集 $V_i=\{v_{i,1},v_{i,2},\cdots,v_{i,m_i}\}$。将其向量表示 $\boldsymbol{e}^v_{i,j}$ 与键映射的概率 $p_{i,j}$ 相乘后累加得到字 x_i 对应的结果 $\boldsymbol{o}_i$，具体如下式所示：

$$\boldsymbol{o}_i=\sum_{j=1}^{m_i}p_{i,j}\boldsymbol{e}^v_{i,j} \tag{19}$$

之后将 $\boldsymbol{o}_i$ 与编码器的输出 $\boldsymbol{h}_i$ 相加后接入一个全连接层得到键值记忆网络的输出 $\boldsymbol{a}_i$，即有：

$$\boldsymbol{a}_i=\boldsymbol{W}_O\cdot(\boldsymbol{h}_i+\boldsymbol{o}_i) \tag{20}$$

3.4.5 解码器

解码器的设计是在 CRF 后接一个 softmax 层来得到所有可能的标签序列的概率。记 $Y=y_1y_2\cdots y_n$ 为对应的标签序列，为了得到每个位置的最佳标签 y_i^*，需要最大化标签概率，如式(21)所示：

$$s(y_i)=\frac{\exp(\boldsymbol{W}_c\cdot\boldsymbol{a}_i+\boldsymbol{b}_c)}{\sum_{y_{i-1}y_i}\exp(\boldsymbol{W}_c\cdot\boldsymbol{a}_i+\boldsymbol{b}_c)} \tag{21}$$

$$y_i^*=\arg\max_{y_i\in tag}s(y_i) \tag{22}$$

其中，l 为标签数，则 $\boldsymbol{W}_c\in\boldsymbol{R}^{l\times l}$ 和 $\boldsymbol{b}_c\in\boldsymbol{R}^l$ 为标签转移中可训练的参数矩阵。tag 表示 y_i 对应的所有可能的标签，大小为 l。

3.4.6 切分概率

字标注的各个标签之间存在一定的约束，以 BMES 4 标注为例：

(1)S 标签后只会出现 S,B；

(2)B 标签后只会出现 M,E；

(3)M 标签后只会出现 M,E；

(4)E 标签后只会出现 S,B。

记 S_{pair} 为合理的相邻标签组合，则有 $S_{\text{pair}}=\{SS,SB,BM,BE,MM,ME,ES,EB\}$。因此，位置 $i-1$ 到位置 i 处的切分情况可表示如下：

$$cut=\begin{cases}0, & \text{if } y_{i-1}y_i\in S_{\text{pair}} \text{ and } y_{i-1}=B/M\\ 1, & \text{else}\end{cases} \tag{23}$$

其中，0 表示不切分，1 表示切分，y_i 为位置 i 处对应的标签，$i\in[1,n-1]$。模型最终将输出每个位置 $cut=1$ 对应的切分概率 c_i。

4 实验

4.1 实验数据集和预处理

本文的实验部分主要用到了两类中文数据集，分别是分词基准数据集和法律领域数据集。

(1)中文分词基准数据集：用来训练多标签 Transformer-CRF 切分概率的模型。本文使用来自 SIGHAN2005/2008 Bakeoff[16-17]的 8 个中文分词基准数据集，即 AS，CITYU，MSR，PKU，CKIP，CTB，NCC 和 SXU。预处理后的训练集细节如表 1 所列。其中标星号对应的是繁体中文数据集，需简化。Sentence 列为文本行数；Wordlist(>1)列是去除单字后的词表大小，Size 列为数据集大小。

表 1 预处理后的 SIGHAN2005/2008 Bakeoff 训练集细节

Table 1 Details of SIGHAN2005/2008 Bakeoff training set after preprocessing

Corpus(Train)		Sentence	Wordlist(>1)	Size/M
Sighan2005	AS*	708953	137346	25.20
	CITYU*	53019	66539	6.81
	MSR	86918	84423	11.70
	PKU	19054	52342	5.25
Sighan2008	CITYU*	36227	41273	5.08
	CKIP*	94169	45458	3.25
	CTB	23444	40076	3.06
	NCC	32099	53649	4.17
	SXU	17115	29932	2.47

(2)法律领域数据集：用于无监督方法中生成候选词。本文爬取了 3 个主流的中文法律数据库：全国人大数据库、北大法律信息网和北大法宝。具体内容包括法规、白皮书、法律新闻、法律释义与问答、实务探讨等。不同网站的文本格式有所区别，对其进行数据清洗，解决一些无效文本、换行、重复的问题，汇总整理后最终得到 360MB 的大规模法律领域数据集，以标题-文本或问题-答案的格式统一存储在 csv 文件中。

为了训练键值记忆网络切分概率模型，本文挑选了结构规整、内容合理的小部分内容进行人工标注。将 9.9MB 的法律释义书(共 66369 个文本行)按大小分为 11 份，交给 11 个标注者进行标注。对于不确定切分的词，所有标注者一起讨论决定。由于 11 个标注者是分开的，相同的词在不同的文本中可能对应着不同的切分结果，最终汇总文本时需要进行一致性校对工作。

4.2 实验结果

4.2.1 基于信息熵和互信息的无监督新词发现

本节给出在大规模法律领域数据集上进行无监督新词发现的实验结果。首先对文本进行基于 n-gram 的分割，将所有

长度在[2,6]范围内、出现次数不小于10的文本串作为候选。设置阈值过滤掉明显不合理的文本串后,综合信息熵和互信息的值排序输出。本文过滤每个指标下后30%的候选词,对PMI来说是3,对BE_min来说是0.8。然后将PMI和BE_min的结果归一化后相加并按序输出,共产生40 522个候选词。

4.2.2 多标签Transformer-CRF切分概率模型

本节根据3.3节的内容在中文分词基准数据集上训练一个多标签Transformer-CRF切分概率模型。字向量采用character.vec[1)]方法,它考虑了汉字的偏旁部首等构字信息。Bigram向量采用FastText在维基百科语料上训练的300维中文词向量,并用Aaunak等[18]提供的降维方式降到100维。实验中使用的超参数如表2所列。

表2 超参数设置

Table 2 Hyper-parameter settings

Hyper-parameter	Value
Embedding Size d	100
Hidden State Size	256
Batch Size	256
Dropout Ratio	0.2
Epoch	100

本文在SIGHAN2005/2008 Bakeoff的各个数据集上加以分词标准标签联合训练,然后分别对各个数据集采用精准率(P)、召回率(R)、$F1$值及R_{OOV}指标进行测试评估,并对比不同解码器,结果如表3所列。可以看出,CRF解码器在所有数据集上的表现全面优于Softmax解码器。对于各数据集来说,MSR上的P,R,$F1$和R_{oov}指标结果全面优于其他数据集。

表3 Transformer-Softmax/CRF模型各数据集实验结果

Table 3 Experimental results of each data set in Transformer-Softmax/CRF model

Corpus	Transformer-Softmax				Transformer-CRF			
	P	R	$F1$	R_{oov}	P	R	$F1$	R_{oov}
AS	96.48	95.34	95.91	67.42	96.56	95.62	96.08	70.88
CITYU	96.79	96.11	96.45	83.02	96.82	96.21	96.52	83.47
CKIP	96.73	95.57	96.15	79.00	96.67	95.76	96.21	79.75
CTB	96.59	96.84	96.72	83.90	96.57	96.86	96.71	84.42
MSR	97.67	97.84	97.76	74.30	97.96	97.97	97.96	75.54
NCC	95.64	95.78	95.71	75.53	95.69	95.90	95.80	76.21
PKU	95.95	96.75	96.35	74.62	95.75	96.77	96.26	75.68
SXU	97.14	97.18	97.16	81.05	97.32	97.32	97.32	81.66
Avg.	96.62	96.43	96.52	77.35	96.67	96.55	96.61	78.45

对每个候选词,找到原始语料中出现该词的上下文句子。由于数据集较大,本文只记录随机抽取的1 000个,出现次数不足1 000的则记录其所有上下文句子。用训练好的Transformer-CRF切分概率模型预测每个句子任意两个字之间的切分概率。综合考虑表3中各数据集的结果,以及MSR数据集保留命名实体构成大量长词的特点,候选词过滤时标签统一选用〈*MSR*〉。

记候选词与其前、后字的切分概率最小值为外部切分概率,候选词中每个位置的切分概率的最大值为内部切分概率,过滤规则定义为:对候选词w,若存在上下文句子S,w在S中的外部切分概率大于a,且内部切分概率不大于b,则保留该候选词,否则将其作为噪声词串过滤掉。

为了探索合理的a,b,本文对不同阈值进行了对比实验,如图5所示,图5(a)中b恒定为0.5,保留的候选词数随外部切分概率阈值的增大而减少,图5(b)中a恒定为0.5,保留的候选词数随内部切分概率阈值的增大而增加。可以发现,保留词数对内部阈值的变化更为敏感。因此,实验选取阈值$a=0.5$,以及一个更为宽容的阈值$b=0.8$。

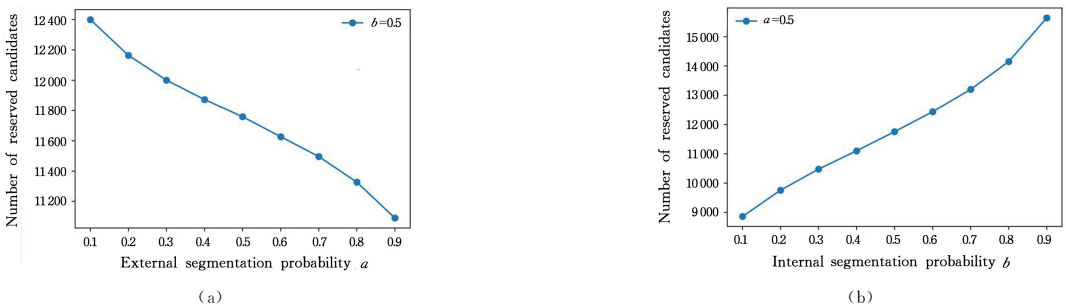

图5 不同外部/内部切分概率阈值下保留的候选词数

Fig. 5 Number of reserved candidates with different external/internal segmentation probability thresholds

遵循上述设置,Transformer-CRF切分概率模型最终保留了14 151个候选词,过滤掉了26 371个。表4列出了部分top输出结果示例,remain列为模型最后保留的词,del列为过滤掉的词。本文对比了相同语料模型D-TopWords[11]的抽取结果,如D-TopWords列所示。由于候选词是基于n-gram方法得到的,因此我们需要对该词是否为合理词(fW列),以及在为合理词的前提下是否为法律相关词(fS列),进行人工判断。从表4中的结果可以看出,模型能对合理词

1) https://github.com/hankcs/multi-criteria-cws/tree/master/data/embedding

进行充分有效的判断，并去除了类似“怂使或促致”“哄抬物价”“之间的”这种不合理词组，并且模型在整体 top K 候选词的提取效果上优于 D-TopWords 模型。

表 4 切分概率模型输出结果示例

Table 4 Example of outputs of segmentation probability model

remain	fW	fS	del	D-TopWords	fW	fS
澳大利亚	1	1	著作权法	民法典	1	1
能够	1	0	怂使或促致	号法律公告	0	0
槟榔	1	1	哄抬物价	根据第	0	0
继续	1	0	檐篷	释义	1	0
选择	1	0	猥亵儿童罪	作者简介	0	0
儿童	1	0	碎片化	ii	0	0
积极	1	0	惰性气体系统	规例	1	1
被害人	1	1	驯养繁殖	iii	0	0
咳嗽	1	0	之间的	本公约	0	0
父母	1	0	借款人	司法解释	1	1
经济	1	0	泡沫灭火	年以下有期徒刑	0	0
瑕疵	1	1	蒸汽容器	视属何情况而定	0	0
劳动者	1	1	诬告陷害	物权法	1	1
宪法	1	1	吹哨人	即属犯罪	0	0
国家	1	1	医疗机构	CLI	1	1
撤销	1	1	摊还债款	个人信息	1	1
匈牙利	1	1	进行了	附表	1	1
协议	1	1	嫖娼的	号第 3 条修订	0	0
墨西哥	1	1	聊天记录	附注	1	1
债权人	1	1	绿色原则	民法总则	1	1

4.2.3 键值记忆网络切分概率模型

本节根据 3.4 节在人工标注的法律释义数据集上训练键值记忆网络切分概率模型。实验按照测试集行数占整个数据集行数的 0.1 来随机分割训练集和测试集。本节采用 BERT 编码器[19]来提取每个输入字符的上下文特征，采用 CRF 解码器来进行切分预测。实验中使用的超参数如表 5 所列。

表 5 超参数设置

Table 5 Hyper-parameter settings

Hyper-parameter	Value
Hidden State Size	768
Hidden State Layers	12
Batch Size	256
Learning Rate	1×10^{-5}
Dropout Ratio	0.1
Epoch	30

BERT-CRF 键值记忆网络切分概率模型在法律释义数据集上表现优越，精准率(P)、召回率(R)、$F1$ 值及 R_{OOV} 指标结果分别达到了 98.98，98.80，98.89 和 68.86。由于训练集和测试集是随机划分的，所以 R_{OOV} 指标的结果随训练集和测试集的划分不同而具有随机性。为了验证 3.4 节提出的 AV 与 PMI 结合的双标准度量方式的有效性，我们采用不同词汇度量方式进行对比实验，结果如表 6 所列。可以看出，双指标度量方式的整体表现普遍优于单指标度量方式，且 AV 与 PMI 结合的度量方式在该数据集上表现最优。

表 6 不同词汇度量方式的实验结果

Table 6 Experimental results on different wordhood measures

Measure	Threshold	P	R	$F1$	R_{oov}
AV	2	98.96	98.75	98.85	68.69
PMI	0	98.92	98.78	98.85	67.68
BE	0	98.93	98.77	98.85	67.21
AV&PMI	2&0	98.98	98.80	98.89	68.86
BE&PMI	0&0	98.94	98.79	98.87	68.32

此处遵循 4.2.2 节设定的过滤规则与探索得到的 a，b 阈值，用键值记忆网络切分概率模型过滤候选词。模型最终保留了 14 006 个结果，过滤掉了 26 516 个结果。为了定量评估，我们对过滤前的候选词集和两个切分概率模型输出结果的 top900 是否为法律相关词进行了人工判断，并使用 MAP 指标[11]量化，结果如表 7 所列。Model1 多标签 Transformer-CRF 模型在 Top900 词中法律相关词的 MAP 高达 54.00%，较候选词集提升了 2.15%；Model2 键值记忆网络的表现进一步超过了前者，达到了 3.43%的效果提升，并在除 100 之外的各个截取点上全面优于 D-TopWords 模型。

表 7 不同模型的 MAP 结果

Table 7 MAP results of different models

（单位：%）

MAP(top K)	100	300	500	700	900
候选词集	37.05	45.49	48.84	50.45	51.85
Model 1	37.27	46.27	50.16	52.08	54.00
Model 2	37.44	**46.67**	**51.00**	**53.18**	**55.28**
D-TopWords	**45.53**	46.50	47.51	47.81	48.04

表 8 列出了本节输出结果中保留及删除的候选词中，与 4.2.2 节模型结果有差异的 top5 输出。从表中可以看出，本节的键值记忆网络切分概率模型能有效保留如“著作权法”“借款人”此类有效词，以及进一步去除如“缔约双方”“珍贵文物”此类噪声词，得到更精确的结果。

表 8 有差异的 top5 候选词

Table 8 Different top5 candidates

remain	del
著作权法	疲劳审讯
哄抬物价	缔约双方
借款人	疫情防控
摊还债款	珍贵文物
行为人	旅游经营者

4.2.4 不同法律领域文本上的新词发现

本节尝试用 4.2.1 节和 4.2.2 节的方法，在有着不同侧重对象和适用范围的法律文本上进行对比实验，如香港法、司法案例、经济法。不同语料下输出的 top10 候选词结果如表 9 所列。

表 9 香港法、司法案例、经济法语料下的 top10 候选词

Table 9 Top10 candidates on Hong Kong law, judicial cases and economic law corpus

Hong Kong Law	Judicial Cases	Economic Law
狂犬病	聂树斌	巴塞尔
帕斯卡	吉利德	居住地
兄弟	蜀都实业公司	宏观调控
诽谤	歌力思	教育
噪声	疲劳审讯	死亡
恐怖分子	饲养动物	剩余
玻璃	慈善信托	剥夺
餐厅	溯及力	储蓄
抚恤金	玩忽职守罪	遗嘱
腐蚀	巡回检察	徇私舞弊

从表 9 可以发现，不同语料的代表词特征明显不同。以司法案例为例，其输出结果包括一些公司（如蜀都实业公司、吉利德、歌力思）、人名（如聂树斌）、一些场景情形（如疲劳审讯、巡回检察），以及一些专有名词（如溯及力、玩忽职守罪）。

实验结果表明，本文方法可以适配任意文本语料，并且在输出质量和数量上都非常优越。此外，不同语料的叙述侧重点不同，对应的新词结果也明显不同，通过取 top 的方式，可以直观地看出该文本的特点，或选取一定量的代表词进行下一阶段的任务。

结束语 本文提出了一种基于信息熵-切分概率模型的新词发现方法。基于信息熵和互信息的无监督新词发现方法可以不依靠任何先验知识和人工处理，从一定规模的文本中自动抽取出数量可观的候选新词。但这种基于文本频率的方法不可避免地会产生噪声词串，而切分概率模型可以完成噪声词串的过滤，得到更准确的新词结果。本文针对有无原始语料相关的标注数据集，通过实验测试了多标签 Transformer-CRF 切分概率模型和键值记忆网络切分概率模型。实验结果表明，本文方法可以适配任意文本语料，并且输出在质量、数量上都很优越的结果，在 top K 候选词的提取效果上优于 D-TopWords 模型。

本文在神经网络训练过程中并未区分合理词与特定领域词，因此接下来会尝试加入特定领域相关参数或者对特定领域词进行标注，来进一步提升模型提取新词的性能。

参 考 文 献

[1] NGUYEN T H, SHIRAI K. Topic modeling based sentiment analysis on social media for stock market prediction[C]// Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2015: 1354-1364.

[2] DONG G, LI R, YANG W, et al. Microblog burst keywords detection based on social trust and dynamics model[J]. Chinese Journal of Electronics, 2014, 23(4): 695-700.

[3] CHENG N C, HOU M, TENG Y L. Short text attitude analysis based on textual characteristic[J]. Journal of Chinese Information Processing, 2015, 29(2): 163-169.

[4] ZHAO Z B, SHI Y X, LI B Y. Newly-emerging domain word detection method based on syntactic analysis and term vector[J]. Computer Science, 2019, 46(6): 29-34.

[5] LIU Y T, WU B, XIE T, et al. New word detection in ancient Chinese corpus[J]. Journal of Chinese Information Processing, 2019, 33(1): 46-55.

[6] TUNG C H, LEE H J. Identification of unknown words from corpus[J]. Computational Proceedings of Chinese and Oriental Languages, 1994, 8: 131-145.

[7] CHURCH K, HANKS P. Word association norms, mutual information, and lexicography[J]. Computational Linguistics, 1990, 16(1): 22-29.

[8] FENG H, CHEN K, DENG X, et al. Accessor variety criteria for Chinese word extraction[J]. Computational Linguistics, 2004, 30(1): 75-93.

[9] BU F, ZHU X, LI M. Measuring the non-compositionality of multiword expressions[C]// Proceedings of the 23rd International Conference on Computational Linguistics. 2010: 116-124.

[10] DENG K, BOL P K, LI K J, et al. On the unsupervised analysis of domain-specific Chinese texts[C]// Proceedings of the National Academy of Sciences. 2016: 6154-6159.

[11] CHEN A, SUN M. Domain-specific new words detection in Chinese[C]// Proceedings of the 6th Joint Conference on Lexical and Computational Semantics(* SEM 2017). 2017: 44-53.

[12] PAN C Z, SUN M S, DENG K. TopWORDS-Seg: Simultaneous Text Segmentation and Word Discovery for Open-Domain Chinese Texts via Bayesian Inference[C]// Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2022: 158-169.

[13] TIAN Y, SONG Y, XIA F, et al. Improving Chinese word segmentation with wordhood memory networks[C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 8274-8285.

[14] QIU X P, PEI H Z, YAN H, et al. A concise model for multi-criteria Chinese word segmentation with transformer encoder [C]// Findings of the Association for Computational Linguistics: EMNLP 2020. 2020: 2887-2897.

[15] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in Neural Information Processing Systems, 2017, 30: 5998-6008.

[16] EMERSON T. The second internationalChinese word segmentation bakeoff[C]// Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing. 2005.

[17] JIN G, CHEN X. The fourth international Chinese language processing bakeoff: Chinese word segmentation, named entity recognition and Chinese pos tagging[C]// Proceedings of the Sixth SIGHAN Workshop on Chinese Language Processing. 2008.

[18] AUNAK V, GUPTA V, METZE F. Effective dimensionality reduction for word embeddings [C] // Proceedings of the 4th Workshop on Representation Learning for NLP. 2019: 235-243.

[19] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [C]// Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2018: 4171-4186.

ZHU Yuying, born in 1997, master. Her main research interests include NLP and machine learning.

GUO Yan, born in 1981, lecturer. Her main research interests include information security, NLP and blockchain.

（责任编辑：何杨）