计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2020,29(6):181−188 [doi: 10.15888/j.cnki.csa.007434]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于改进 PMI 和最小邻接熵结合策略的未登录词识别①

徐豪杰[1], 吴新丽[1], 杨文珍[1], 潘志庚[2]

[1](浙江理工大学 虚拟现实实验室, 杭州 310018)
[2](杭州师范大学 数字媒体与人机交互研究中心, 杭州 311121)
通讯作者: 杨文珍, E-mail: ywz@zstu.edu.cn

摘　要: 中文分词是中文自然语言处理的重要任务, 其目前存在的一个重大瓶颈是未登录词识别问题. 该文提出一种非监督的基于改进 PMI 和最小邻接熵结合策略的未登录词识别方法. 滤除文本中无关识别的标点符号和特殊字符后, 此方法先运用改进 PMI 算法识别出文本中凝聚程度较强的字符串, 并通过停用词词表和核心词库的筛选过滤, 得到候选未登录词; 然后, 计算候选未登录词的最小邻接熵, 并依据词频-最小邻接熵判定阈值, 确定出文本中的未登录词. 通过理论及实验分析, 此方法对不同的文本, 在不需要长时间学习训练调整参数的情况下, 即可生成个性化的未登录词词典, 应用于中文分词系统后, 其分词正确率、召回率分别达到 81.49%、80.30%.

关键词: 中文分词; 未登录词识别; 改进 PMI 算法; 邻接熵

引用格式: 徐豪杰,吴新丽,杨文珍,潘志庚.基于改进 PMI 和最小邻接熵结合策略的未登录词识别.计算机系统应用,2020,29(6):181–188. http://www.c-s-a.org.cn/1003-3254/7434.html

Out-of-Vocabulary Detection Based on Combination Strategy of Improved PMI and Minimum Branch Entropy

XU Hao-Jie[1], WU Xin-Li[1], YANG Wen-Zhen[1], PAN Zhi-Geng[2]

[1](Virtual Reality Laboratory, Zhejiang Sci-Tech University, Hangzhou 310018, China)
[2](Digital Media & Human-Computer Interaction Research Center, Hangzhou Normal University, Hangzhou 311121, China)

Abstract: Chinese word segmentation is an important task in Chinese natural language processing. One of bottleneck problems in Chinese word segmentation is Out-Of-Vocabulary (OOV) detection. This study proposes an unsupervised OOV detection method based on improved PMI algorithm and minimum branch entropy combining strategy. Firstly, the punctuation marks and special characters which are not related in the text are removed. The improved PMI algorithm recognizes the string with strong cohesion in the text, and gets the candidate OOV through the filtering of the stop word list and the core vocabulary. Then the minimum branch entropy of candidate OOV is calculated, when the term frequency-minimum branch entropy threshold is met, the output is the OOV. Through theoretical and experimental analysis, the algorithm can generate a personalized OOV dictionary for different texts, and does not require long-term learning and training to adjust parameters, and has a certain improvement in the accuracy and recall rate of detection.

Key words: Chinese word segmentation; out-of-vocabulary detection; improved PMI algorithm; branch entropy

① 基金项目: 国家重点研发计划 (2018YFB1004901); 浙江省自然科学基金 (LQ19F020012); 浙江省基础公益研究计划 (LGF19E050005)
Foundation item: National Key Research and Development Program of China (2018YFB1004901); Natural Science Foundation of Zhejiang Province (LQ19F020012); Basic Research Plan for Public Welfare of Zhejiang Province (LGF19E050005)
收稿时间: 2019-11-24; 修改时间: 2019-11-28; 采用时间: 2019-12-05; csa 在线出版时间: 2020-06-10

1 前言

中文分词存在两个重要挑战: 歧义问题和未登录词问题[1]. 其中 60% 的分词错误是由未登录词导致的, 故如何高效且正确地识别未登录词是中文自然语言处理研究的重点和难点[2,3]. 前人的相关研究有新词和未登录词两个概念, 一定程度上, 新词可归属于未登录词范畴, 一般情况下并不对这两个概念作明确区分.

未登录词是指在词典中不存在的、未被及时收录的词, 包括: 中外人名、地名、机构组织名、事件名、缩略语、派生词、各领域术语以及没有固定生产机制的网络新词[4]. 例如, 缩略词“高数”、“抵京”、“发改委”、“音协办”, 2019 年网络新词“种草”、“萌萌哒”、“人鱼线”、“雨女无瓜”等等. 未登录词随着社会发展不断涌现, 本质上是不可穷尽收集登录的, 对分词系统而言, 词表不能无限扩大, 那么对未登录词的自动识别就显得愈发重要.

互信息 (Pointwise Mutual Information, PMI) 是概率统计学领域的一个重要概念. 本文将改进 PMI 算法和最小邻接熵结合, 为改进 PMI 算法产生的垃圾串提供了新的约束, 提出了一种基于生语料文本本身的未登录词识别模型, 利用非监督学习方法, 在凝聚强度较弱的字符之间切割出字符串, 过滤掉词库中存在的词语得到候选未登录词, 统计出候选未登录词的最小邻接熵, 输出满足阈值要求的独立词, 从而提高了未登录词识别的正确率.

2 相关工作

未登录词的识别方法主要可以分为两类: 基于规则的未登录词识别和基于统计的未登录词识别[4,5]. 前者利用有规律可循的构词学原理, 配合语义和词性信息所构造的规则模板, 通过匹配规则来发现未登录词, 通常根据汉语构词法建立规则知识库, 过滤掉不符合规则或者符合相关典型错误构词规则的垃圾字串, 留下的即是候选未登录词. 基于规则的识别方法针对性强, 但规则库的制定需要根据特定文本进行修改以适应文本的多样性, 且该方法难以应对毫无规则的网络新词; 统计的方法往往需要大规模语料库, 通过文本当中某个统计量的固有特征进行统计, 计算出成词概率高的字串. 基于统计的未登录词识别方法灵活通用, 移植性好, 但是面临着数据稀疏和正确率不高等问题. 目前多数研究者都采用统计与规则相结合的方法, 发挥组合优势, 设置多重过滤手段, 从而提升未登录词识别效果.

Pecina 等[6]在计算了 50 余种量化指标之后, 证明了 PMI 指标是衡量字串间相关度最好的指标之一. 但单纯的 PMI 指标缺点是容易过高估计低频且总是相邻出现的字串. 例如“鸳鸯”、“憔悴”两词, 几乎难以在用语习惯中找到其他搭配, 这样的字串 PMI 值非常高, 包含这些字串的垃圾串 PMI 值也非常高, 算法容易出现误判, 产生例如“对鸳鸯”、“憔悴的”等错误. 针对此类问题, 有学者提出了将 PMI 与 log-likelihood 方法相结合的手段进行未登录词识别[7,8], 使得该类错误得到一定改善. 张峰等[9]利用 PMI 算法计算字间相关度, 设置了相关的构词规则, 人工建立了普通词语搭配前缀、后缀库, 过滤掉类似“十分兴奋”、“非常开心”这样的复合词. 梁颖红等[10]则将 PMI 和 NC-value 相结合, 提高了 3 字以上未登录词的识别率, 但是 PMI 和 NC-value 存在着一定的推导耦合关系, 两者并不能够独立地约束过滤相关垃圾串. 天荣朋等[11]利用元递增算法 (N-Gram) 提取未登录词候选项, 对提取出来的候选未登录词使用频率和停用字、PMI 和邻接熵 (Branch Entropy, BE)、相应词典等进行多重筛选, 取得了不错的实验效果. 何婷婷等[12]分析了中文术语构成特点, 提出了一种基于质子串分解的术语自动抽取方法, 该方法对特定的中文文本有较好的效果. Pazienza 等[13]在 PMI 算法的基础上, 提出了 PMI^n 算法, 当向 PMI 指标中引入 3 个及以上的联合概率因子时, PMI^n 方法能够克服单纯 PMI 算法的缺点. 国内学者杜丽萍等[14]则通过理论和实验证明, 当向 PMI 方法中引进 3 个及以上的联合概率因子时, 能够克服单纯 PMI 指标的缺点, 其中 PMI^3 的效率最高, 但依然缺少对上下文左右邻接字的考量和联系.

中文的成词规则和习惯表明, 词是较为独立的, 能够在文本中不同位置自由搭配使用. 本文结合表征凝聚强度的互信息, 为改进 PMI 算法引入表征词语在文本字符串中自由程度的最小邻接熵, 利用改进的 n 阶互信息 PMI 算法, 自动筛选出可能成词的片段, 再通过候选未登录词的词频和最小邻接熵融合判定阈值过滤第一轮当中存在的垃圾串, 确定出未登录词. 该方法充分发挥了最小邻接熵在表征独立词在文本中自由灵活搭配程度和改进 PMI 算法在词内部凝聚强度两方面的优势, 两者相互约束配合, 能进一步提高成词正确率.

3 改进 PMI 和最小邻接熵的结合策略

本文采取基于改进 PMI 和最小邻接熵结合策略的方法, 对不同文本, 能在较小时间开销下学习训练调整参数, 生成个性化的未登录词词典, 并提升了现有分词系统的性能. 首先, 我们对生语料文本进行预处理, 去除干扰未登录词识别的标点符号、数字、特殊符号、英文字母、URL 链接等, 在预处理之后, 系统先用 PMI^3 算法摘选出 5 字及 5 字以下的所有可能成词的字符串, 经过 35 万结巴词库过滤, 得到候选的未登录词, 候选词进入判定环节, 符合词频-最小邻接熵 (*TF-Entropy*) 融合指标阈值的会被写入未登录词输出文本当中. 因为词频指标依然一定程度上能够反映字符串的成词率, 本文为了不误过滤低频未登录词, 故利用数据融合手段, 将词频指标和最小邻接熵指标融合. 图 1 为本文未登录词的识别流程.

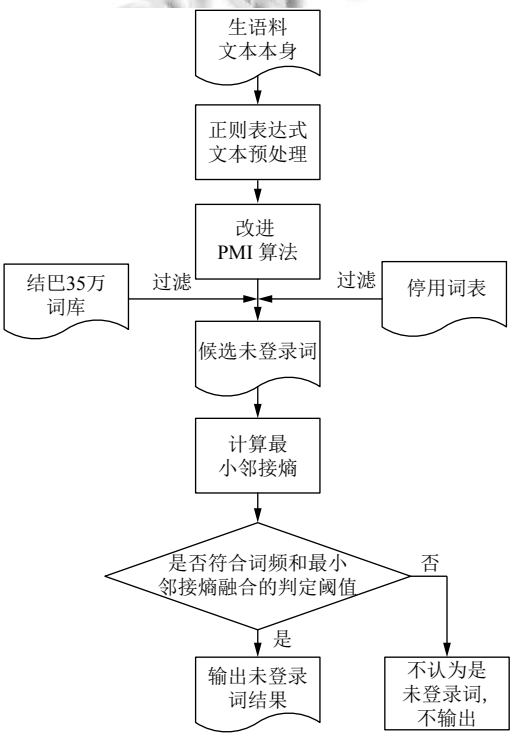

图 1 改进 PMI 和最小邻接熵结合策略的未登录词识别

3.1 独立词的成词维度和指标设计

借鉴信息论中的互信息概念, 两事件x和y之间的互信息计算公式如下:

$$R_{\mathrm{PMI}}(x,y)=\log\frac{p(x,y)}{p(x)p(y)} \tag{1}$$

在本文中, $p(x)$和$p(y)$分别表示字串x和y的频率, $p(x,y)$表示字符串xy在文本当中的频率. 可以利用互信息表征一个字串内部的凝聚强度, 互信息值越高, x和y之间的聚合强度越大, xy组成一个独立词的概率越高, 其中存在独立词边界的可能性就越低.

Bouma[15]对传统的 PMI 算法做出了相关改进, 引入 n 个联合概率因子, 提出了 PMI^n 算法. PMI^n 算法定义如下:

$$R_{\mathrm{PMI}^n}(x,y)=\log\frac{p^n(x,y)}{p(x)p(y)},n\in N^+ \tag{2}$$

$R_{\mathrm{PMI}^n}(x,y)$表示字串x和y相关度, 也称 PMI^n 值. x与y之间的聚合联结越紧密, 他们成词可能性越大, 那么xy的组合越可能是对独立词. 当$n=1$时, PMI^n 算法即 PMI 算法. PMI^n 算法识别候选未登录词是先计算出两个字之间的凝聚强度, 确定 2 元待扩展种子, 再利用 PMI^n 方法计算 2 元待扩展种子分别和其左邻接字、右邻接字的凝聚强度, 判断 2 元字符是否能扩展成 3 元, 如此迭代, 扩展出最大词长为 5 元的独立词.

杜丽萍等[14]从理论和实验上证明了 PMI^n 算法比单纯的 PMI 具有更高的精度, 且$n=3$时算法效率最高. 综合考虑计算资源和效率, 本文选定 PMI^n 作为本文第一步提取候选未登录词的算法.

从中文句法规则和成词规律当中分析得出, 如果一个字符串片段能够成为独立词, 它应有较为丰富的左邻字集合和右邻字集合. 以香农提出的信息熵 (entropy) 为基础, 衍生出最小邻接熵的概念, 用来表征字符串片段的自由运用程度.

信息熵的定义如下: 如果某事件的发生概率为p, 当该事件发生时, 此事件包含的信息量为$\log(p)$. 候选未登录词左右邻接熵的计算公式如下:

$$\begin{cases} H_{\mathrm{left}}=\sum_{i=1}^{n}-p_i\times\log(p_i) \\ H_{\mathrm{right}}=\sum_{j=1}^{m}-p_j\times\log(p_j) \end{cases} \tag{3}$$

其中, i表示左邻接字的编号, j表示右邻接字的编号, n 和 m 分别表示不重复左右邻接字的总数, H_{left}表示候选字串左邻接字i的出现频率, H_{left}表示候选未登录词左邻接熵, H_{right}表示候选未登录词右邻接熵. 同时为了保证独立词左右边界的清晰性, 防止出现类似“辈子”、“后遗”、“鹅卵”等单边邻接熵较高的垃圾串, 我们应该取左右邻接熵的较小值, 故最小邻接熵如下:

$$Entropy = Min\{H_{\text{left}}, H_{\text{right}}\} \tag{4}$$

基于统计的思想认为, 一个字符串搭配如果在语料中的次数越多, 那么该字符串是一个独立词的可能性越高. 通过实验发现, 未登录词与登录词在词频 TF 较高时的分布有一定差异, 故引入词频特征能一定程度上提升算法效果. 为更加科学地考虑词频 TF 指标, 也为方便比较加权和计算, 本文在阈值过滤环节构造了一个 *TF-Entropy* 参数, 即词频-最小邻接熵判定阈值. 该指标的计算方法是分别对词频和最小邻接熵做归一化, 把候选未登录词的词频和最小邻接熵变成 (0, 1) 之间的小数, 消除两者的量纲, 成为纯量, 再根据两者在独立词成词的贡献度设置融合权重.

$$X' = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \tag{5}$$

X 为某一个变量, $X_{\max}$ 表示这组变量中的最大值, $X_{\min}$ 表示变量中的最小值, X' 是归一化之后的纯量.

$$TF - Entropy = i \times F' + (1 - i) \times Entropy' \tag{6}$$

单纯 PMI 方法需要整定词频、PMI、邻接熵等多个参数, 无法集中精力调参优化. 本文方法在保证判定维度不变的情况下, 结合了内部聚合程度、外部自由程度, 用于对未登录词的识别, 同时设计了一个综合的衡量指标, 方便针对特定文本优化调整, 在较小时间开销下学习训练调整参数, 生成个性化的未登录词词典.

3.2 独立词判定算法

词是字的组合, 相邻的字同时出现的次数越多, 就越有可能构成一个词. 字与字相邻共现的频率能够反映成词的可信度, PMI^3 算法能够很好的表征词内部字与字之间凝聚强度. 在文本当中, 所有的字符可以看做构成了一个字符串, 字符串当中嵌入了多个独立词, 未登录词识别的目标是能够抽取字符串当中存在的独立词. 信息熵是信息论当中的概念, 能很好的运用在表征灵活自由度的场合中. 因此, 本文为 PMI^3 算法引入最小邻接熵的约束, 结合凝聚、自由两个维度, 进行独立词的判定. 将文本进行数据清洗预处理, 去除所有标点符号, 得到纯净的文本 CORPUS, 其中 STOPWORDS 和 Jieba_Vocabulary 分别代表停用词词表和结巴过滤词库, NEW_CANDIDATES 代表算法运行过程中新扩展的候选未登录词, OOV_CANDIDATES 表示候选未登录词列表集合, 具体步骤如下所示:

GET_OOV_WORDS(CORPUS, STOPWORDS, Jieba_Vocabulary)

Step 1. 遍历 CORPUS, 计算中间二元字串和左边二元字串 PMI^3 值的平均值, 计算中间二元字串和右边二元字串 PMI^3 值的平均值.

Step 2. 如果中间二元字串的 PMI^3 值分别大于 3 倍左、右侧二元字串的 PMI^3 值, 则将中间二元字串组合后的词加入待扩展种子, 得到待扩展种子列表.

Step 3. 在种子左右各取一元, 分别计算三元 PMI^3 值, 如果左侧三元字串的 PMI^3 值大于右侧, 则①往左扩展; 否则②往右扩展.

① 左侧三元字串 PMI^3 值大于等于 1/3 种子本身 PMI^3 值, 则将左扩展三元字串加入 NEW_CANDIDATES; 否则, 终止扩展输出种子字串.

② 右侧三元字串 PMI^3 值大于等于 1/3 种子本身 PMI^3 值, 则将右扩展三元字串加入 NEW_CANDIDATES; 否则, 终止扩展输出种子字串.

Step 4. 依次迭代, 直至达到 5 元字串, 终止扩展.

Step 5. 利用 STOPWORDS 和 Jieba_Vocabulary 对 NEW_CANDIDATES 进行过滤, 得到 OOV_CANDIDATES.

Step 6. 分别计算候选未登录词 OOV_CANDIDATES 的左右邻接熵, 求出 OOV_CANDIDATES 的最小邻接熵.

Step 7. 从内存中查取候选未登录词 OOV_CANDIDATES 的词频 *TF*.

Step 8. 将词频和最小邻接熵分别归一化去量纲得到纯量.

Step 9. 对统计得到的词频和最小邻接熵的做数值融合处理, 方便下一步判断.

Step 10. OOV_CANDIDATES 中符合词频-最小邻接熵阈值要求的, 输出为最终的未登录词 OOV_WORDS.

4 实验与分析

4.1 实验数据

1) 由于不同文本存在个性化差异, 甚至文本之间的风格、写作逻辑也存在很大不同, 本文讨论的算法旨在于分析一种通用的未登录词识别方案, 故选取了数据稀疏程度较高的互联网文本, 其中包括 4 个类别, 分别为新闻数据、微博数据、汽车论坛数据、餐饮点评数据, 约 14 万字, 共 79 226 个词. 实验结果的评测标

准语料, 按照北大现代汉语基本加工规范进行处理, 以作为未登录词识别和中文分词性能实验的标准答案.

2) 符号过滤表: 利用正则表达式从语料中筛选的标点符号和特殊符号.

3) 停用词词典: 包含 702 个停用词 (选自哈尔滨工业大学停用词表), 用于过滤候选未登录词中的垃圾串.

4) 结巴词典: 共收集 354 895 个词语, 是目前较为规范的词典之一, 用于过滤候选未登录词中的已登录词, 以便得到词库中未登录的独立词.

4.2 实验过程

4.2.1 文本预处理

由于互联网论坛语料极不规范, 预处理的目的是将其中的 URL 链接、标点符号等干扰项过滤掉, 得到较为纯净的实验文本, 以保证未登录词自动识别过程中, 算法不受干扰. 利用正则表达式从语料中筛选过滤掉标点等符号, 例句的预处理结果如表 1 所示.

表 1 文本预处理前后对比

原文本内容	预处理后的文本内容
2012 年比赛时间为 1 月 7 日 (星期六). 比赛项目: 男、女: 马拉松 (42.195 公里)、半程马拉松 (21.0975 公里)、10 公里、5 公里. 同时举办海峡两岸城市马拉松邀请赛, 轮椅半程马拉松, 12 公里轮滑比赛. 规模约 8 万人, 每个项目约 2 万人, 报满为止. 相关比赛办法请参阅各项目竞赛规程详情进厦门国际马拉松官网: http://www.xmim.org/cn/ 网上报名已经开始. 有想法的朋友赶快啊!	2012 年比赛时间为 1 月 7 日星期六比赛项目男女马拉松 42195 公里半程马拉松 210975 公里 10 公里 5 公里同时举办海峡两岸城市马拉松邀请赛轮椅半程马拉松 12 公里轮滑比赛规模约 8 万人每个项目约 2 万人报满为止相关比赛办法请参阅各项目竞赛规程详情进厦门国际马拉松官网网上报名已经开始有想法的朋友赶快啊

4.2.2 未登录词部分识别结果

Chen 等[1]经研究指出, 99% 的独立词是在 5 字及 5 字以下, 所以算法设置最大抽取词长为 5. 实验中词频纯量 *TF'*的融合权重为 0.2, 词频-最小邻接熵 *TF-Entropy* 设置为 0.7. 表 2 列举了部分基于 PMI^3 算法和最小邻接熵结合策略的未登录词识别结果. 从表中统计的的数据可知, 该算法对 2 字词和 3 字词识别数目占总识别到的词数的 78.6%, 对 3 字以上的多字词识别占比为 60.9%, 验证了结合策略算法对多字词有一定识别能力. 结果中有一些错误未登录词, 例如“据了解”、“接到报警后”等, 主要原因是在特定主题的互联网语料中, 同样的表述反复出现和使用, 导致这些字串的凝聚和自由程度较高, 作为垃圾串没有被很好的识别过滤.

表 2 部分排序靠前的 N-Gram 未登录词识别结果

N-Gram 数量	比例 (%)	排序靠前的未登录词									
2-Gram 4836	38.9	何炅	骏捷	异响	吐槽	韩庚	朗逸	宏光	设计	吴邪	芋圆
		破拆	宝来	挖鼻	唐狮	微博	生煎	试驾	储物	启悦	前脸
3-Gram 4928	39.7	科鲁兹	邓华德	胡文静	姚姗姗	王哲林	米查姆	**据了解**	团购券	同级别	博尔特
		六合村	指南者	柯兰多	跑高速	市音协	三厢版	侧气囊	张起灵	包裹性	合资车
4-Gram 1959	15.7	其它描述	京东商城	专项检查	指哪打哪	真皮座椅	编辑点评	指向精准	网易体育	独立悬挂	汽车之家
		皮质包裹	十二星座	游戏演示	小微企业	后排座椅	万达公馆	路感清晰	**油耗目前**	华夏母亲	应急救援
5-Gram 683	5.5	梅德韦杰夫 索尼爱立信 伦敦奥运会 日间行车灯 液压扩张器 鹰潭市政府 无线客户端 **商品房销售** 科技金羊网 海南省军区									
		中国青年网 韩国国奥队 小额便利贷 奥迪旅行版 英格兰女超 中关村在线 **接到报警后** 大气上档次 领取准考证 四六级考试									

4.2.3 未登录词识别算法对比

评测未登录词识别和中文分词系统性能, 一般采用正确率 *P*、召回率 *R*、*F* 值来衡量, 其计算公式如下所示, 其中 *TP*、*FP*、*FN* 分别表示正样本识别正确的个数, 正样本识别错误的个数, 负样本识别错误的个数.

$$P = \frac{TP}{TP+FP} \tag{7}$$

$$R = \frac{TP}{TP+FN} \tag{8}$$

$$F = \frac{2\times P\times R}{P+R}\times 100\% \tag{9}$$

针对未登录词识别系统, *TP* 表示系统识别到的正确未登录词词数, *TP* + *FP* 表示系统识别出的所有未登录词词数, *TP* + *FN* 表示标准实验文本中所有未登录词

词数. 将本文方法与常见未登录词识别代表方法, 进行未登录词识别对比实验, 结果见表 3.

表 3 未登录词识别对比试验 (%)

序号	方法	*P*	*R*	*F* 值
1	传统 PMI 方法	23.67	14.31	17.84
2	传统 N-Gram 方法	24.33	15.65	19.05
3	基于 N-Gram+互信息	30.09	17.63	22.23
4	改进 PMI 方法 (3 阶)	36.45	18.12	24.21
5	本文方法	**39.42**	**20.66**	**27.11**

方法 1 利用 PMI 算法与基本过滤算法结合, 从语料中识别未登录词. 方法 2 使用 N 元递增算法 (N-Gram) 完成对未登录词的识别, 并加入了少量的简单规则过滤方法, 有效地提高了未登录词识别的效果, 但是对三字词以上的未登录词识别效果较差, 所以其正确率较低. 方法 3 在传统 N-Gram 的基础上引入互信息 (PMI) 来进行未登录词的抽取, 实验证明了该方法的有效性. 方法 4 在单纯 PMI 算法基础上引入 3 阶联合概率因子, 较好克服了 PMI 方法精度不高的缺点. 但方法 3、4 均缺少对未登录词上下文的联系, 没有考虑词语的外部成词概率. 单纯的 PMI^3 算法, 虽然相对 PMI 算法, 正确率有很大的提高, 但单以词语内部凝聚强度的大小作为词语的边界, 缺少考虑词语外部自由程度反映的成词概率. 所以方法 5 在基于改进 PMI^3 方法 (3 阶) 的基础上引入表征未登录词所在文本中自由灵活程度的最小邻接熵指标, 通过互信息和邻接熵的互相约束提高未登录词发现精度. 该方法有效解决了未登录词的边界界定问题, 精准确定未登录词的前后边界, 最终得到一个系统效率较高, 实验结果更精确的未登录词识别结果.

采用 PMI^3 和最小邻接熵相结合的未登录词识别方法, 在能较好识别多字词的基础上, 同时考虑了词语的内、外部成词概率, 所以其总体识别效果进一步提高. 融合表征词内部字间凝聚强度的 PMI^3 算法和词外部的灵活自由程度的最小邻接熵的未登录词识别系统在未登录词识别中取得了不错的效果, 其正确率、召回率、*F* 值相对于其他算法均有一定的提高.

4.2.4 改进分词系统实验

本文将 PMI^3 算法和代表着独立词外部自由程度的邻接熵相结合, 能够更好的提取出文本当中个性化词串, 帮助识别未登录词, 最后与核心词典库加载融合, 生成个性化的文本词典. 改进的 Jieba 分词系统切分文本主要分为 3 个步骤: (1) 基于生语料文本本身进行未登录词识别和个性化候选字串的提取; (2) 将识别结果编纂成针对文本本身的用户词典, 加载融合到 Jieba 分词系统中; (3) 对语料进行分词.

将本文结合策略算法识别到的未登录词作为个性化用户词典, 引入关闭自身未登录词发现功能的 Jieba 分词工具, 与 Jieba 自带的未登录词识别功能作对比. 实验设计分组如下. 实验 1: 关闭未登录词识别功能的 Jieba 分词系统, 该系统完全依赖 Jieba 自带的核心词库, 不具有未登录词识别能力; 实验 2: 开启未登录词识别功能的 Jieba 分词系统, 该系统依赖核心词库的同时, 利用隐马尔科夫模型 (Hidden Markov Model, HMM) 思想识别未登录词; 实验 3: 加载个性化用户词典的 Jieba 分词系统, 本方法基于 Jieba 系统自带的核心词库, 并加载融合了结合策略算法识别得到的个性化词典. 表 4 例举了语料中例句在实验 1、实验 2 和实验 3 中的结果.

表 4 分词实验结果举例

实验编号	分词实验结果举例
实验 1	|最近|有|特惠|套餐|除|早餐|时段|**双层**|**堡**|加|饮料|15|元|~|在|卖|的|**三国**|**杀**|优惠卡|觉得|有点|无用|唉|, |如果|爱|吃|**麦**|**旋风**|的话|就|还|可以|, |饮品|的话|, |套餐|都|有|的|嘛|, |单点|下午茶|呗|.
实验 2	|最近|有|特惠|套餐|除|早餐|时段|**双层**|**堡**|加|饮料|15|元|~|在|卖|的|**三国**|**杀**|优惠卡|觉得|有点|无用|唉|, |如果|爱吃|**麦**|**旋风**|的话|就|还|可以|, |饮品|的话|, |套餐|都|有|的|嘛|, |单点|下午茶|呗|.
实验 3	|最近|有|特惠|套餐|除|早餐|时段|**双层堡**|加|饮料|15|元|~|在|卖|的|**三国杀**|优惠卡|觉得|有点|无用|唉|, |如果|爱|吃|**麦旋风**|的话|就|还|可以|, |饮品|的话|, |套餐|都|有|的|嘛|, |单点|下午茶|呗|.

例句: 最近有特惠套餐除早餐时段双层堡加饮料 15 元~在卖的三国杀优惠卡觉得有点无用唉, 如果爱吃麦旋风的话就还可以, 饮品的话, 套餐都有的嘛, 单点下午茶呗.

表 4 中, 针对例句, 关闭和开启未登录词识别功能的 Jieba 分词系统均把未登录词“双层堡”、“三国杀”切分为“双层|堡”、“三国|杀”; Jieba 分词系统加载个性化用户词典 (词典中包含未登录词“双层堡”、“三国杀”) 后, 分词系统把未登录词“双层堡”、“三国杀”切分为一个词. 关闭未登录词识别功能的 Jieba 分词系统分词把“麦旋风”切分为“麦|旋风”, 开启未登录词识别功能的 Jieba 分词系统把未登录词“麦旋风”切分, 将“麦

旋风”中的“麦”和它前面的“爱吃”结合起来切分为“爱吃麦”和“旋风”, 结果为“爱吃麦|旋风”; 加载个性化用户词典的 Jieba 分词系统 (词典中包含未登录词“麦旋风”) 把未登录词“麦旋风”切分为一个独立词. 从互联网测试语料的分词结果来看, 主要有 2 种情况: ① 关闭和开启未登录词识别功能的 Jieba 分词系统在遇到未登录词时, 大多情况下均是将未登录词切分为多个“散串”, 如“双层|堡”、“三国|杀”, Jieba 分词系统加载包含这些未登录词的个性化用户词典后, 均能被正确切分; ② 开启未登录词识别功能的 Jieba 分词系统自动识别出的未登录词不正确, 导致句中邻近词的错分, 如例句中把“爱吃麦”当做一个词, “麦旋风”后的“旋风”单独成词. 实验表明通过加载个性化用户词典改进分词系统是一种可靠有效的方法.

为进一步验证该未登录词识别方法的有效性, 以及识别生成的个性化词典应用到分词系统中的整体效果, 对数据稀疏程度较高的论坛语料, 统计上述 3 组分词系统的中文分词正确率、召回率和 F 值, 针对分词系统, 式 (7)~式 (9) 中的 TP 表示分词系统切分正确的词数, $TP + FP$ 表示分词系统切分出的总词数, $TP + FN$ 表示标准实验文本中包含的总词数, 结果如表 5 所示.

表 5 中文分词系统对比

实验编号	切分出的总词数	识别到的未登录词数目	正确率 (%)	召回率 (%)	F 值 (%)
实验 1	83 125	0	79.72	73.28	76.36
实验 2	78 159	3906	81.30	80.12	80.71
实验 3	**71 412**	**4563**	**81.49**	**80.30**	**81.29**

从表 5 可见, Jieba 加载用户词典后, 分词系统识别出的未登录词数目达 4563 个, 对互联网语料的分词效果有明显提升, 精确率、召回率、F 值也相对加载前分别提高 1.77%、7.02% 和 4.93%. 相对开启未登录词识别功能的 Jieba 分词系统, Jieba 加载用户词典后分词系统识别出的未登录词数目增加 657 个, 精确率、召回率和 F 值也分别提高 0.19%、0.18% 和 0.58%. 通过将本文结合策略算法识别到的未登录词构成个性化用户词典, 可以提高原分词系统的分词性能, 尤其对于网络新词效果显著, 纠正了针对未登录词的分词错误. 得益于构建的个性化未登录词词典, 本文方法在分词结果中有着最大的 F 值, 验证了该方法在网络新词识别方面具有较高的正确率和召回率.

5 结论与展望

本文在前人研究基础上, 针对未登录词识别问题, 基于互联网论坛文本, 将 PMI[3] 算法和独立词外部邻接熵相结合, 加强过滤条件, 将词频、最小邻接熵按照成词机理, 科学融合成一个综合指标, 形成 *TF-Entropy* 过滤参数, 很好地抑制了垃圾串的出现. 相比其他算法, 该方法在 P、R、F 值上均带来了一定的提升, 且能在较小时间开销下学习训练调整参数. 本文算法能够很好地针对特定文档构建个性化用户词典, 加载融合常用核心词库后, 提升了现有分词系统性能.

该算法在数据稀疏, 训练数据标记不规范, 缺少大规模语料的情况下, 有着较好的表现. 下一步工作是针对特定文本, 研究自适应整定词频-最小邻接熵参数的方法, 引入更多专家经验, 在保证分词速度的同时能够提高对未登录词的识别率, 最终实现正确分词, 进一步提高在数据稀疏的情况下, 分词系统对文本本身的自动化处理能力.

参考文献

1 Chen XC, Shi Z, Qiu XP, *et al*. Adversarial multi-criteria learning for Chinese word segmentation. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver, BC, Canada. 2017. 1193–1203.

2 Lileikytė R, Fraga-Silva T, Lamel L, *et al*. Effective keyword search for low-resourced conversational speech. Proceedings of 2017 IEEE International Conference on Acoustics, Speech and Signal Processing. New Orleans, LA, USA. 2017. 5785–5789.

3 Sheikh I, Fohr D, Illina I, *et al*. Modelling semantic context of OOV words in large vocabulary continuous speech recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2017, 25(3): 598–610. [doi: 10.1109/TASLP.2017.2651361]

4 Li XQ, Zhang JJ, Zong CQ. Towards zero unknown word in neural machine translation. Proceedings of the 25th International Joint Conference on Artificial Intelligence. New York City, NY, USA. 2016. 2852–2858.

5 Van Heerden C, Karakos D, Narasimhan K, *et al*. Constructing sub-word units for spoken term detection. Proceedings of 2017 IEEE International Conference on Acoustics, Speech and Signal Processing. New Orleans, LA, USA. 2017. 5780–5784.

6 Pecina P, Schlesinger P. Combining association measures for

collocation extraction. Proceedings of COLING/ACL on Main Conference Poster Sessions. Sydney, Australia. 2006. 651–658.
7 Pantel P, Lin DK. A statistical corpus-based term extractor. Proceedings of the 14th Biennial Conference of the Canadian Society for Computational Studies of Intelligence. Ottawa, ON, Canada. 2001. 36–46.
8 韩艳, 林煜熙, 姚建民. 基于统计信息的未登录词的扩展识别方法. 中文信息学报, 2009, 23(3): 24–30, 50.
9 张锋, 许云, 侯艳, 等. 基于互信息的中文术语抽取系统. 计算机应用研究, 2005, 22(5): 72–73, 77.
10 梁颖红, 张文静, 周德富. 基于混合策略的高精度长术语自动抽取. 中文信息学报, 2009, 23(6): 26–30.
11 天荣朋, 许国艳, 宋健. 基于改进互信息和邻接熵的微博新词发现方法. 计算机应用, 2016, 36(10): 2772–2776.
12 何婷婷, 张勇. 基于质子串分解的中文术语自动抽取. 计算机工程, 2006, 32(23): 188–190.
13 Pazienza MT, Pennacchiotti M, Zanzotto FM. Terminology extraction: An analysis of linguistic and statistical approaches. In: Sirmakessis S, ed. Knowledge Mining. Berlin, Heidelberg: Springer, 2005. 255–279.
14 杜丽萍, 李晓戈, 于根, 等. 基于互信息改进算法的新词发现对中文分词系统改进. 北京大学学报 (自然科学版), 2016, 52(1): 35–40.
15 Bouma G. Normalized (pointwise) mutual information in collocation extraction. From form to meaning: Processing texts automatically. Proceedings of the Biennial GSCL Conference. Potsdam, Germany. 2009. 31–40.