软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2013,24(5):1051−1060 [doi: 10.3724/SP.J.1001.2013.04254]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

基于条件随机场方法的开放领域新词发现*

陈 飞[1,2,3], 刘奕群[1,2,3], 魏 超[1,2,3], 张云亮[3], 张 敏[1,2,3], 马少平[1,2,3]

[1](智能技术与系统国家重点实验室(清华大学),北京 100084)

[2](清华大学 清华信息科学与技术国家实验室(清华大学)(筹),北京 100084)

[3](清华大学 计算机科学与技术系,北京 100084)

通讯作者: 陈飞, E-mail: chenfei27@gmail.com, http://www.csai.tsinghua.edu.cn/

摘 要: 开放领域新词发现研究对于中文自然语言处理的性能提升有着重要的意义.利用条件随机场(condition random field,简称 CRF)可对序列输入标注的特点,将新词发现问题转化为预测已分词词语边界是否为新词边界的问题.在对海量规模中文互联网语料进行分析挖掘的基础上,提出了一系列区分新词边界的统计特征,并采用 CRF 方法综合这些特征实现了开放领域新词发现的算法,同时比较了 *K*-Means 聚类、等频率、基于信息增益这 3 种离散化方法对新词发现结果的影响.通过在 SogouT 大规模中文语料库上的新词发现实验,验证了所提出的方法有较好的效果.

关键词: 新词发现; condition random field(CRF);中文分词

中图法分类号: TP391 **文献标识码**: A

中文引用格式: 陈飞,刘奕群,魏超,张云亮,张敏,马少平.基于条件随机场方法的开放领域新词发现.软件学报,2013,24(5):1051−1060. http://www.jos.org.cn/1000-9825/4254.htm

英文引用格式: Chen F, Liu YQ, Wei C, Zhang YL, Zhang M, Ma SP. Open domain new word detection using condition random field method. Ruan Jian Xue Bao/Journal of Software, 2013,24(5):1051−1060 (in Chinese). http://www.jos.org.cn/1000-9825/4254.htm

Open Domain New Word Detection Using Condition Random Field Method

CHEN Fei[1,2,3], LIU Yi-Qun[1,2,3], WEI Chao[1,2,3], ZHANG Yun-Liang[3], ZHANG Min[1,2,3], MA Shao-Ping[1,2,3]

[1](State Key Laboratory of Intelligent Technology and Systems (Tsinghua University), Beijing 100084, China)

[2](Tsinghua National Laboratory for Information Science and Technology (Tsinghua University), Beijing 100084, China)

[3](Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China)

Corresponding author: CHEN Fei, E-mail: chenfei27@gmail.com, http://www.csai.tsinghua.edu.cn/

Abstract: Open domain new word detection is vital for Chinese natural language processing research. This paper proposes a novel detection algorithm based condition random field (CRF), which treats the new word detection problem as a classification problem. In this algorithm, the study tries to separate boundaries of new words from existing words with both the CRF method and a serial of statistical features extracted from large scale corpus. The effectiveness of three different discretization strategies are also compared including *K*-means, equal-frequency, and information gain. Experimental results on a large-scale Web corpus named SogouT show the effectiveness of the proposed algorithms.

Key words: new word detection; conditional random field; Chinese word segmentation

新词发现是中文自然语言处理领域一个非常重要的研究内容.它与分词过程密切相关[1].由于在中文信息处理中,不像英文等西方语言,词与词之间有固定的分隔符,所以分词通常作为中文信息处理任务最开始的一个

* 基金项目: 国家自然科学基金(60903107, 61073071); 国家高技术研究发展计划(863)(2011AA01A205)
收稿时间: 2011-09-20; 修改时间: 2012-03-19; 定稿时间: 2012-04-23

必要步骤[2].文献[3]中提到,分词任务中所遇到的分词工具字典未包含的词(未登录词,本文所指新词属于未登录词)会显著影响分词的性能.因此,新词发现对于提高分词,以致后续工作都有重要的意义.然而近年来,个人博客、个性签名、微博等 Web 2.0 应用的出现,允许用户自己生成网页内容,导致类似于“神马”、“超女”等等新词汇大量出现,并以非常快的速度更新,使得新词发现面临更大的挑战.因而,目前关于新词发现的研究主要集中在人名[4,5]、地名[6,7]、翻译缩写[8,9]或者某几个领域术语(如军事[10]、财经等领域)的自动提取.在文献[11]中,按照新词发现任务的范围将其分为 3 类:(1) one-for-one,这类研究主要解决某个特殊问题的新词发现问题,如人名、地名等;(2) one-for-several,主要解决几个特定类或者领域的新词发现问题;(3) one-for-all,面向所有问题,即开放领域内的新词发现问题,并且指出,这类问题在文献[11]之前还没有可应用的算法被提出来.按照方法,主要存在 3 类:(1) 基于 *n*-gram 语言模型[12];(2) 依赖于某个分词工具分词之后,进行新词发现[11];(3) 将新词发现与分词工具进行结合,在分词的同时,检测新词[2].此外,文献[1]提出了以分析用户行为、采用协同过滤的方法进行新词发现.文献[11]中还提出了将新词发现分为基于语言规则和基于统计机器学习的方法,这也是目前主流的分类方式.

由于互联网等交流工具的快速发展,新词也以极快的速度产生,并且变化多样,很难用模板规则匹配;即使生成了规则模板,但由于新词更新十分迅速,且构成规则多样化,也会使模板很快失效.另一方面,规则的提出与维护需要由语言学家进行,不仅耗费时间、金钱,而且不可扩展.因此,目前的新词发现工作主要集中在某个或者某几个特殊领域下按照统计机器学习的方法,引入领域知识(如军事领域特有的“战斗机”、“坦克”,财经领域特有的“股票”、“上市”等)进行特定领域的新词发现.依靠领域知识和该领域新词的标注信息,机器学习的方法可以更好地学到适合该领域新词发现的模型,即便是基于规则匹配的方法也能够提出更多特定于该领域构词法的规则,从而提高识别的准确度和召回率.而对于开放领域新词发现,其新词识别过程是面向所有可能的领域,甚至是像“神马”这种不针对任何特定领域的新词,因此当前的研究相比之下较少.其最大的难点在于,无论是规则匹配还是统计机器学习,都没有可利用的领域知识来针对新词发现进行优化.因为领域的划分及其可能的数量是不确定的,即使存在某种比较全面的领域划分规则,判断待发现新词属于某个领域也是十分困难的,这些都无疑增加了开放领域新词识别的难度.本文提出先将文档正文进行分词,并以词语之间的语言统计信息作为特征,利用 CRF 能够对序列输入上下文相关标注的特点,将新词发现问题转化为预测已分词词语边界是否为新词边界的机器学习问题,实现 one-for-all 开放领域新词发现.因为所有特征的计算、模型的学习都是程序自动离线完成的,因此维护方便,能够适应新词构成方式的快速变化;另外,由于采用 CRF 方法进行学习、预测,能够充分利用其特点,更好地描述新词发现与输入序列特征以及上一次预测结果之间非独立性的关系.在本文实验部分,还对特征数值的离散化过程进行了等频率、*K*-Means 聚类和信息增益这 3 种方法的比较.实验结果表明,本文的方法对新词发现有比较好的效果,其中以等频率进行离散化的方法最优.

本文第 1 节介绍相关工作及 CRF 原理.第 2 节介绍本文所用的特征及分析.第 3 节介绍实验.第 4 节给出结论并对未来的工作进行介绍.

1 相关工作

1.1 新词发现

新词发现作为中文信息处理领域的一个重要的步骤,近年来出现了很多相关的研究.文献[1]提出了基于用户行为的方法,主要依赖于用户的输入法字典与用户自身维护的各个领域字典.通过分析多个用户维护的领域字典,计算词语代表性度量并以此在字典中发现与领域相关的关键词,按照协同过滤的方式,利用这些关键词发现该领域的专家用户,最终以这些专家用户的输入法字典进行新词发现.文献[13]中通过对 Sinica corpus 统计得到被分词工具错分的新词中有绝大多数被分为单音节词的结论,从而仅仅对分词之后的单音节词进行上下文相关的分析并认为,如果一个单音节词在其上下文符合语义或者语法的独立性,则其是一个词,否则可能是一个新词的一部分;通过学习的方式得到语义或语法上的规则,并通过规则的精确度阈值调整来进行规则选择,达到控制最终的精确度和召回率的目的,并进行新词的检测.文献[14]先对文本进行分词,然后再统计其中 2-gram~

8-gram 的搭配,选择其中出现频率高的搭配作为新词.在文献[2]中,将分词与新词发现结合在一起.通过计算分词之后的可信度,选取位于阈值之上的词语或者位于那些可信度在阈值之上的词语之间的词语作为新词.对于领域术语的发现,文献[10]中使用了分词之后的词本身、词性、左信息熵、右信息熵、互信息、TF/IDF 作为特征,用 CRF 学习进行领域术语的识别.文中提到,利用词本身作为一个特征,可以加强该方法的领域相关性.其原因在于,每个领域都有其自身特有的词汇,如文献[10]中所述军事领域,枪、飞机、舰、弹等为军事领域所特有词语.同时,文献[10]中使用了 *K*-Mean 聚类的方法离散化特征取值.在文献[15]中,首先通过计算语料库中分词后的词语之间静态联合率(static association rate)提取出具有紧密关系的词语,再用语法规则、领域特征对上述词语进行过滤,最终选择具有最高置信度(the highest confidence)的词语作为领域术语.文献[16]使用当前待处理的字与其前后相邻两字的组合,使用 CRF 学习进行未登录中文命名实体的识别.以上方法都是以特定领域为前提进行的领域相关的新词发现.而对于开放领域的新词发现,文献[11]中提出了对分词之后的词语,按照词语所在上下文进行其在句子中可能的角色标注,最终通过其被标注的角色,利用隐马尔可夫模 $p(y|x)$进行分解,而不是直接分解联合概率分布 $p(x,y)$来避免对输入的概率分布 $p(x)$进行建模、计算[17−19].因为输入 x 通常具有比较复杂的关系,对 $p(x)$的计算往往需要对输入带有某种独立性假设.而 CRF 通过直接分解条件概率分布 $p(y|x)$,避免了对 $p(x)$的计算,因此它能更好地学习到数据集中的关系.

在线性链 CRF 中,将 $p(y|x)$做如下的分解:

$$p(y\mid x;\omega)=\frac{1}{Z(x,\omega)}\exp\sum_{i=1}^{N}\sum_{j}\omega_j f_j(y_{i-1},y_i,x,i),$$

$$Z(x,y)=\sum_{Z_{1,N}}\exp\sum_{i=1}^{N}\sum_{j}\omega_j f_j(y_{i-1},y_i,x,i),$$

其中,ω为待估计参数;f_j 为特征函数;i 表示对输入序列 x 在特征函数 f_j 上求和,这样可以保证对于变长的输入 f_j 有估计 j 数目的特征函数值.虽然在理论上来说,特征函数 f_j 中可以与所有的 x 产生关系,但是在实际使用时,考虑到复杂性以及实际问题中输入之间关系的特征,可能选择的仅仅是当前输入以及前后一两个输入作为该特征函数的自变量.CRF 的一个优点在于,不用假设输入 x 之间的独立性关系就能计算 $p(y|x;\omega)$.而输入与输出之间的关系是通过 CRF 的使用者在特定的任务中指定的特征函数 f_j 以及 CRF 自动学习的参数 ω_j 来体现.线性链 CRF 则对 CRF 有一定的条件限制:当前输出 y_i 除了与 x 有函数关系以外,只能与前一个输出 y_{i-1} 有关.在本文的新词发现任务中,需要预测当前词与邻近词能否构成新词(即输出 y_i),其结果不仅依赖于这几个词的特征取值(即输入 x),而且依赖于对上一个词的预测结果(即 y_{i-1}),因为上一个词是否被预测为新词会影响当前词的预测,这与线性链 CRF 的模型正好吻合.因此,本文采用线性链 CRF 作为机器学习的方法进行新词发现.

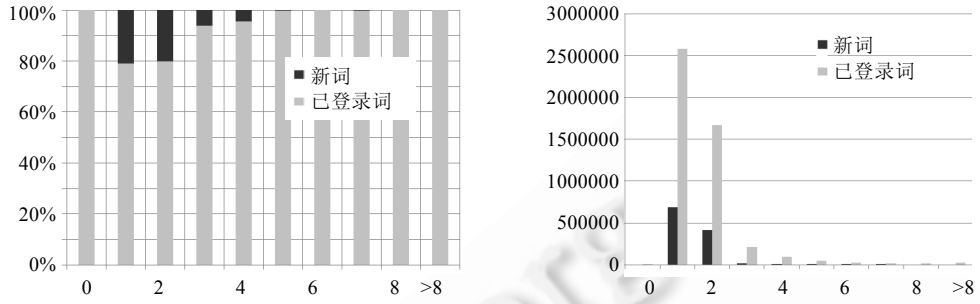
2 特征介绍

CRF 的学习与预测是在样本的多个特征上进行的,本节列出本文所用特征.在计算特征取值时,先用分词工具对语料库正文进行分词,得到具有词性标注的独立词语,再计算每个词的各个特征取值.由于 CRF 在学习与预测时要求这些特征取值为离散变量,因此本文对部分取值连续的特征进行了离散化,而对取值有限的特征未进行离散化.在对特征取值进行统计时,将分词工具正确分出的词统计为“已登录词”,将分词工具未正确分出但被标注为新词的词统计为“新词”.

(1) 本词长度 L_0:计算分词之后每个词所包含的单字个数.L_0 衡量该词所包含的字的数量,在中文语料库中,即使是新词,其长度也不会很长.因此,L_0 可以衡量该词作为新词一部分的可能性大小.因为如果本词长度值已经很大,那么它与其他词组合并能构成新词的概率就小.由于长度是由分词工具切词后再计算的,故 L_0 只可能存在有限的取值,因此未对 L_0 进行离散化处理.

从图 1 中可以看出,作为构成新词的词语,其长度主要分布于 1~2 之间.长度超过 4 的词,基本不可能与其他词组合从而构成新词.从图 1(a)我们还可以发现,新词与已登录词之比很小,这是由于新词的数量本来就比已登录词少很多,这一点同样适用于在其他特征上所得到的比例图.而图 1(a)只是新词与已登录词在 L_0 的各个取值

上的比例,其在各个值上的分布如图 1(b)所示.可以看到,大部分的词语其 L_0 取值都在 5 以下.至于图 1 中的 L_0 有少数取值为 0 的情况,是因为在预处理过程中对标点和空格等非汉字也进行了分词,而在计算它们的各个特征时并没有进行计算,直接取值为 0.但这并不会干扰新词发现的效果,因为这些非汉字字符在被标注时绝不会被标记为新词.同样地,L_0 存在超过 8 的情况,由图 1(b)可知,其数量仍然非常少,这是因为网页中存在像连续上百个"—"这样的特殊字符,而在预处理时又不可能完全将其排除,但实际上在分词的过程中它们是被作为一个单元来处理的.同样地,它们在标注时仍然被标注为非新词,因此对新词发现的结果并不会有影响.



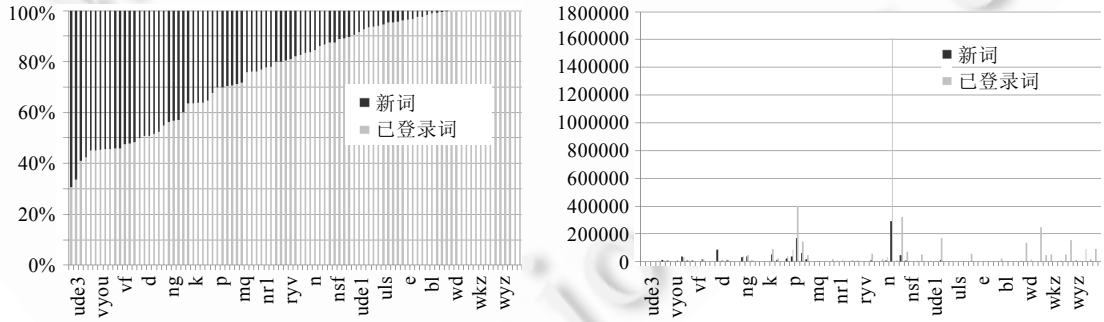
(a) 新词与已登录词在 L_0 各个取值上的比例 (b) 新词与已登录词在 L_0 各个取值上的分布

Fig.1 Proportion and distribution of both new word and word in dictionary on L_0 values

图 1 新词与已登录词在 L_0 各个取值上的比例及分布

(2) 本词词性 POS_0:通过分词工具进行分词之后得到的词性标注.在利用规则的新词发现方法中[15],往往由语言学家总结与维护一些常见的新词词性结构,如"n+v","v+v+n"等规则,并以此来对分词之后的未登录词进行发现.采用这种方法,其优点在于语言学家总结的规则对新词发现具有较高的准确率,但是总结和维护这些规则需要的时间和成本是非常巨大的.本文引入本词词性作为特征,利用 CRF 在学习过程中可包含本词及前后词语词性的特性,自动归纳出可能组合为新词的各种语言结构,从而有助于判断多个词语的组合能否构成新词.

如图 2 所示,对于词性,如果该词为"ude3"(助词"得")、"vyou"(动词"有")或者"vf"(趋向动词)等等,则其成为新词的一部分的可能性很大.同样地,图 2(a)说明的是新词在 POS_0 的各个取值上的比例,而图 2(b)是其分布.



(a) 新词与已登录词在 POS_0 各个取值上的比例 (b) 新词与已登录词在 POS_0 各个取值上的分布

Fig.2 Proportion and distribution of both new word and word in dictionary on POS_0 values

图 2 新词与已登录词在 POS_0 各个取值上的比例及分布

(3) 本词的左信息熵 LE_0:其定义为

$$LE(w)=-\frac{1}{n}\sum_{a\in A}C(a,w)\log\frac{C(a,w)}{n},$$

其中,w 表示本词,A 为语料库中位于 w 左边的词的集合,$C(a,w)$表示语料库中词语 a 与 w 同时出现的次数.

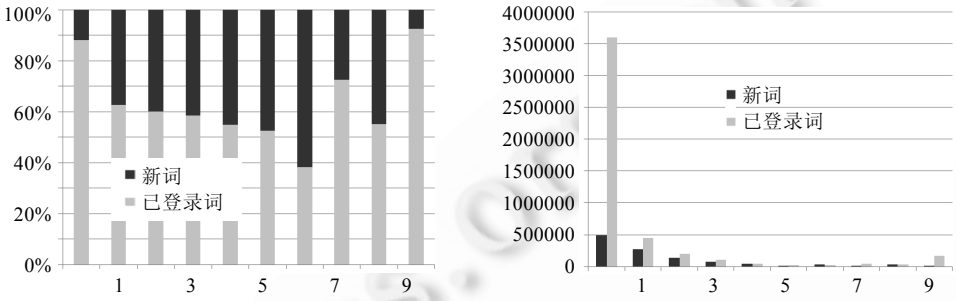
(4) 本词的右信息熵 RE_0:同左信息熵,其定义为

$$RE(w)=-\frac{1}{n}\sum_{a\in B}C(w,a)\log\frac{C(w,a)}{n},$$

其中,w 表示本词,B 为语料库中位于 w 右边的词的集合,$C(w,a)$表示语料库中词语 w 与 a 同时出现的次数.

词语的左、右信息熵分别衡量语料库中作为词语左、右侧邻近词语的固定层度.词语之间的组合出现得越固定,其熵值就越大.而如果语料库中在词语 w 左侧(或者右侧)固定地出现词语 w_1,组合 w_1w(或者 ww_1)就可能是一个词语.因此,可通过计算语料库中词语左、右信息熵来判断词语之间组合为新词的可能性.

(5) 本词全文词频 TF_0:计算本词在整个语料库中出现的次数.由于词语在语料库中出现的次数取值范围很大,因此对该特征值做了 10 个等级的离散化处理.如图 3 所示为经过等频率方法离散化之后特征取值的分布.



(a) 新词与已登录词在 TF_0 各个取值上的比例 (b) 新词与已登录词在 TF_0 各个取值上的分布

Fig.3 Proportion and distribution of both new word and word in dictionary on TF_0 values

图 3 新词与已登录词在 TF_0 各个取值上的比例及分布

由图 3(a)我们可以看到,离散化后等级在 0 或者 9 的,很大一部分都是已登录词,而取值在 4~6 之间的词则有可能是新词.

(6) 本词 IDF_0:定义为

$$IDF_w=\log\left(\frac{D}{|D_w|}+0.01\right),$$

其中,D 表示文档总数,$|D_w|$表示包含有词 w 的文档总数.

(7) 本词与前一词共同出现频率 IFA_0:计算本词与前一词在语料库中同时出现的次数.

TF 与 *IDF* 是衡量词语在文档中重要程度的重要指标,在本文中,也将其作为 CRF 学习的特征.以上的特征都是在整个语料库上进行计算.此外,本文还提出了在各个文档上计算的局部特征:

(8) 本词本文词频 TFD_0:计算本词在当前文档中出现的次数.

(9) 本词本文左信息熵 LED_0:以当前文档为统计对象计算本词的左信息熵.

(10) 本词本文右信息熵 RED_0:以当前文档为统计对象计算本词的右信息熵.

(11) 本词平均左信息熵 $LEDM_0$:定义为

$$LEDM_w=\frac{\sum_{D_w}LED_w}{|D_w|},$$

其中,D_w 为包含词 w 的文档集合,LED_w 表示 w 在其所在文档内计算的左信息熵.

(12) 本词平均右信息熵 $REDM_0$:定义为

$$REDM_w=\frac{\sum_{D_w}RED_w}{|D_w|},$$

其中,D_w 为包含词 w 的文档集合,RED_w 表示 w 在其所在文档内计算的右信息熵.

由于没有像文献[10]中那样将本词自身作为一个特征用于 CRF 学习,因此,当如上所述的全局特征被离散化之后,将会出现特征取值相同,但是该词实际所属类别却不同的情况.这在 CRF 学习过程中将会影响参数估计,因此,本文提出计算以上的局部特征,即在每个文档内计算这些特征的取值,而不是在整个语料库中.由于每个文档中包含的词语不可能完全一样,因此对于不同的词,在不同的文档中,其特征取值不会完全相同.

(13) 互信息 M_0:定义为

$$M(w_1,w_2)=\frac{p(w_1,w_2)}{p(w_1)\cdot p(w_2)},$$

其中,$p(w)$为词 w 出现的概率.

在 CRF 进行学习的过程中,除了直接指定各个特征以外,还可以将特征进行组合.在本文中,对如下特征进行组合:

A) 前一词与本词长度、本词与下一词长度:L_{-1}/L_0 以及 L_0/L_1;

B) 前一词与本词词性、本词与下一词词性:POS_{-1}/POS_0 以及 POS_0/POS_1;

C) 本词右信息熵与下一词左信息熵:RE_0/RE_1;

D) 上一词与本词全文词频、本词与下一词全文词频:TF_{-1}/TF_0 以及 TF_0/TF_1;

E) 上一词与本词本文词频、本词与下一词本文词频:TFD_{-1}/TFD_0 以及 TFD_0/TFD_1;

F) 本词本文右信息熵与下一词本文左信息熵:RED_0/RED_1;

G) 本词平均右信息熵与下一词平均右信息熵:$REDM_0/REDM_1$.

由于所选特征较多,我们将特征分为 3 类:第 1 类为特征 1~特征 7,其计算都是基于整个语料库进行,称为全局特征,对应的组合特征为 A~D;第 2 类为特征 8~特征 12,其计算是按照当前文档进行,称为文档局部相关特征,对应的组合特征为 E~G;第 3 类为特征 13,即互信息.

3 实 验

本文将新词发现的问题转化为预测分词之后词语边界是否为新词词语边界的问题.如“神马”这个新词经过分词之后变成了“神/n 马/nr1”.即分词工具认为“神”右边是一个词的边界,“马”右边也是一个词语边界.而实际上,“神”右边不是真正的词语边界.本文在 CRF 的学习与预测中,对于被分词工具分开但其实不是真正边界的情况标记为“0”;如果是真正边界,则标记为“1”.因此,“神马”经分词并标注后形式为“神 0 马 1”.加上离散化后的特征取值,其最终形式见表 1.其中,最后的一位“0”表示“神”字在新词中不是一个边界,而“马”末尾的“1”表示其是一个边界.这样就表示“神马”是一个未被识别的新词.

Table 1 Instance labeled by Sogou high-frequency words

表 1 Sogou 高频词标注后实例

Word	Values of features																Label
神	n	1	2	3	1	0	4	4	4	2	1	3	1	1	0	1	0
马	nr1	1	2	7	1	0	6	5	4	2	2	4	1	0	1	4	1

3.1 实验数据集

SogouT[20]

在本文的验证实验部分,数据集使用搜狗实验室于 2006 年抓取的互联网网页文档(SogouT 2006,共计约 4 000 万张互联网网页,压缩后大小约为 130GB),并随机抽取其中 20 000 个文档作为该实验的语料库.如表 2 所示,每个文档的格式为:一个唯一的文档 ID(docid)、抓取该文档的 URL(url)、文档正文内容(content).

SogouCA[21]

在本文的对比实验部分,数据集使用搜狗实验室发布的若干新闻站点的体育、IT、国内、国际等 18 个频道的新闻数据(SogouCA,共计约 100 万张新闻网页,压缩后大小约为 450MB).同样地,在该实验中随机抽取其中 20 000 个文档作为语料库.每个文档的格式与 SogouT 中的文档相同.

Table 2 Data format of SogouT 2006 and SogouCA

表 2 SogouT 2006 以及 SogouCA 的数据格式

Item	Description	Content
url	网页 URL 字符	〈url〉http://flash.17173.com/〈/url〉
docid	Sogou 为每个抓取网页分配的唯一 ID 号	〈/docid〉170c299974a1b540-8350bb9a81017190〈/docid〉
content	网页正文内容	〈content〉每周游戏排行榜…〈/content〉

SogouW[22]

实验中对应的新词,使用搜狗实验室发布的互联网词库(SogouW).其来自于对 SOGOU 搜索引擎所索引到的中文互联网语料的统计分析(进行统计的时间为 2006 年 10 月),共计 15 万条高频词.除了标出这部分词条的词频信息之外,还标出了常用的词性信息.

3.2 验证实验

实验首先将 SogouT 2006 中随机选取的 20 000 个文档,以中国科学院计算技术研究所的 ICTCLAS 分词器进行分词,并在其上计算各词的特征取值.然后以 SogouW 作为新词,对分词后的网页正文进行上述词语分界“0”,“1”标注,最后按照 95%用于 CRF 学习、5%用于测试(采用 95%训练与 5%测试的比例,是为了与之后的对比实验保持一致),进行本文所提出的基于 CRF 新词识别方法的验证实验.

验证实验共分为两个部分:

第 1 部分实验验证各个特征的有效性,因为特征较多,我们按照前述分类之后的特征进行实验.即先以第 1 类特征进行 CRF 学习,得到测试结果作为一组基线结果(Baseline).然后分别将第 1 类与第 2 类特征组合、第 1 类与第 3 类特征组合以及所有特征组合进行对比实验.

第 2 部分实验是关于特征取值离散化.由于 CRF 的特征取值只能是离散化的值,因此在学习之前必须将连续特征值进行离散化.在文献[10]中,用 *K*-Means 对每个特征聚 10 类进行 10 等级离散化.本文中,同时对等频率、*K*-Means 与信息增益这 3 种离散化方法进行了对比实验,并同样取 10 个离散等级.

以上实验的结果见表 3.

Table 3 Experimental results of the open field new word detection based on CRF (SogouT data set)

表 3 基于 CRF 的开放领域新词发现实验结果(SogouT 数据集)

Features	Discretization strategy	Precision	Recall	*F*-Value
第 1 类特征 (Baseline)	等频率	0.916	0.932	0.924
	***K*-Means**	**0.917**	**0.933**	**0.925**
	信息增益	0.916	0.932	0.924
第 1 类特征+第 2 类特征	**等频率**	**0.918**	**0.933**	**0.926**
	K-Means	0.917	0.933	0.925
	信息增益	0.917	0.933	0.925
第 1 类特征+第 3 类特征	**等频率**	**0.933**	**0.942**	**0.938**
	K-Means	0.932	0.941	0.937
	信息增益	0.931	0.940	0.936
第 1 类特征+第 2 类特征+第 3 类特征	**等频率**	**0.934**	**0. 942**	**0.938**
	***K*-Means**	**0.934**	**0.942**	**0.938**
	信息增益	0.933	0.942	0.937

3.3 结果及分析

从表 3 中可以看到,Baseline 的正确率已经比较高,而在其上增加的第 2 类特征对正确率与召回率并没有太大的提升.而对于互信息的加入,正确率与召回率都有较大的提高.我们之所以考虑加入第 2 类特征,是因为第 1 类特征都是全局特征,且本文的特征中并没有将当前词本身作为一个特征.那么有可能在整个语料库中,有某些词在计算特征取值并经过离散化之后会有相同的值,但它们是不同的词,并且一些是新词的边界,而另一些却不是,那么 CRF 仅仅依靠这些特征,就会将这样的词语边界都预测为新词边界或者非新词边界.因此,希望通过引

入第 2 类局部特征来帮助 CRF 对这类问题进行预测.而第 2 类特征的加入对实验结果提高不大,说明上述情况在语料库中发生的几率较小.

另外,通过第 2 部分的实验可以看到,等频率离散化方法在正确率、召回率以及 *F* 值上具有稳定的、较好的效果.

3.4 对比实验

因为 SogouT 为互联网网页数据,其用词不规范、噪声大、领域广,为新词发现带来了非常大的挑战.为了对比说明这一点,我们进行了上述方法在新闻语料 SogouCA 上进行新词发现的对比实验.由于中国科学院分词器 ICTCLAS 是在新闻语料上进行的训练,因此选用新闻语料做对比实验,可以排除分词器所带来的偏置.另外,新闻语料与 SogouT 相比具有用词更规范、领域更集中的特点,可以对比验证开放领域新词发现,具有更大的挑战.按照验证实验相同的步骤,在 SogouCA 数据集上得到的结果见表 4.

Table 4 Experimental results of the open field new word detection based on CRF (SogouCA data set)

表 4 基于 CRF 的开放领域新词发现实验结果(SogouCA 数据集)

Features	Discretization strategy	Precision	Recall	*F*-Value
第 1 类特征 (Baseline)	**等频率**	**0.942**	**0.952**	**0.947**
	K-Means	0.941	0.952	0.947
	信息增益	0.942	0.952	0.947
第 1 类特征+第 2 类特征	**等频率**	**0.944**	**0.953**	**0.948**
	K-Means	0.941	0.952	0.946
	信息增益	0.943	0.953	0.948
第 1 类特征+第 3 类特征	**等频率**	**0.948**	**0.956**	**0.952**
	K-Means	0.945	0.955	0.950
	信息增益	0.948	0.956	0.952
第 1 类特征+第 2 类特征+第 3 类特征	**等频率**	**0.949**	**0. 957**	**0.953**
	K-Means	0.946	0.955	0.951
	信息增益	0.949	0.956	0.952

从表 4 以及与表 3 对比中,我们可以得到以下结论:

(1) 等频率离散化方法仍然比其他离散化方法要好;

(2) 表 4 中无论是准确率、召回率还是 *F* 值都比表 3 中的好,这验证了开放领域新词发现比较指定领域(这里是新闻领域)的新词发现更困难.

为了与其他新词发现方法进行对比,我们实现了文献[23]中所述方法.因为该文献提出的方法和本文一样,也是与领域无关的,并且其在 SIGHAN 测评的数据集(PK corpus)上取得了较好的结果.我们在 SogouT 2006 与 SogouCA 的文档中分别计算该文献中提出的 IWP,Analogy to New Words,Anti-word 列表(由于原文没有提供 Anti-word 列表以及获取该列表的方法,我们构造由停词连接的词语作为该列表)、词语在文档中频率等特征.但由于在文献[23]中所提出的方法只能对两个字的词(NW11)以及由一个两个字的词加上一个单字组成的三字词(NW21)进行新词发现,因此我们在 SogouT 与 SogouCA 语料库上也分别进行这两类实验.为了与该文中的训练、测试比例保持一致,我们将语料库分为 95%用于 SVMLight 学习,5%用于测试,得到的结果见表 5.

Table 5 Experimental results of comparison experiment on SogouT 2006 and SogouCA data sets

表 5 SogouT 2006 与 SogouCA 数据集上对比实验结果

Data set	Model	Precision	Recall	*F*-Value
SogouT	NW11	0.892	0.146	0.251
	NW21	0.662	0.670	0.666
SogouCA	NW11	0.793	0.847	0.819
	NW21	0.679	0.709	0.693

从表 5 中我们可以看到,文献[23]中所述用于与本文对比的方法,在 SogouT 上的 NW11 类型新词识别上性能较差,其 Recall 和 *F* 值取值都很低.分析其原因可能是:在文献[23]中,其训练语料 PK corpus(4.7MB 训练数据,

89KB 的测试数据)为 1998 年《人民日报》的新闻文章,相对于 SogouT 的互联网文档,前者无论是汉语用词还是文章结构,都比后者正式、规范,这无论是对于预处理中的分词还是其采用的 IWP 与 Analogy to New Words 两特征,都有很强的偏置.对 NW11 的新词识别使用了这两个特征,其结果大幅度下降,而没有使用这些特征的 NW21 下降幅度略小;结合在 SogouCA 新闻语料上的实验结果与文献[23]得到的结果一致,就直接验证了这一点.同时,也再次验证了开放领域新词发现相对于指定领域新词发现的挑战性.在此,我们对比文献[23]中的方法在 SogouT 和 SogouCA 上的下降程度与本文提出的基于 CRF 的新词识别方法在这两个数据集上的下降程度,可以得出本文提出的方法在正确率、召回率和 F 值上,不仅比对比实验的方法在指定领域(这里是新闻领域)有更好的结果,而且当新词识别扩展到开放领域上时,能够得到比对比实验方法更稳定的结果(对比表 3 与表 4,本文所提出的方法在准确率、召回率、F 值上性能降低皆在 3%以内).而且本文提出的新词发现方法对词语组成(NW11,NW21)没有任何限制.

4 结论与未来的工作

本文利用分词器对语料库文档进行分词,再计算这些词的各种语言统计特征,并用 CRF 进行训练,最终识别新词.在没有引入词本身以及语言规则作为特征的条件下,仅以统计特征进行训练与预测,仍然得到了较好的效果,从而克服了现有的新词发现方法与领域相关或者难以维护规则等困难.并比较了 3 类特征以及等频率、K-Means 聚类、信息增益这 3 种离散化方法对新词发现的影响,并且得到等频率离散化方法效果最好的结论.

由于目前的新词发现方法需要依赖于分词器,因此将本文所提出的方法与分词器进行结合,使得分词器可以直接发现新词,这是未来一方面的工作.另外,可以利用上述方法对一字、二字词等进行统计学习来实现新词发现,从而不依赖于分词器,这也是未来可进一步做的工作.

致谢 本文是作者陈飞在“人民搜索”实习期间工作的一项深入研究.“人民搜索”的熊锦华老师和方小娟给予了陈飞很大的帮助,在此向他们表示感谢.

References:

[1] Zheng YB, Liu ZY, Sun MS, Ru LY, Zhang Y. Incorporating user behaviors in new word detection. In: Proc. of the IJCAI 2009. San Francisco: Morgan Kaufmann Publishers, 2009. 2101−2106.

[2] Peng FC, Feng FF, McCallum A. Chinese segmentation and new word detection using conditional random fields. In: Proc. of the 20th Int'l Conf. on Computational Linguistics (COLING 2004). Stroudsburg: Association for Computational Linguistics, 2004. Article No.562. [doi: 10.3115/1220355. 1220436]

[3] Sproat R, Emerson T. First Int'l Chinese word segmentation bakeoff. In: Proc. of the 2nd SIGHAN Workshop on Chinese Language Processing. Sapporo, 2003. 133−143.

[4] Ji H, Luo ZS. A Chinese name identifying system based on inverse name frequency model and rules. In: Proc. of the Natural Language Understanding and Machine Translation. Beijing: Tsinghua University Press, 2001. 123−128.

[5] Zhang HP, Liu Q. Automatic recognition of Chinese personal name based on role tagging. Chinese Journal of Computers, 2004, 27(1):85−91 (in Chinese with English abstract).

[6] Li LS, Huang DG, Chen CR, Yang YS. Identifying Chinese place names based on support vector machines and rules. Journal of Chinese Information Processing, 2006,20(5):51−57 (in Chinese with English abstract).

[7] Li LS, Huang DG, Chen CR, Yang YS. Research on method of automatic recognition of Chinese place names based on support vector machines. Mini-Micro Systems, 2005,26(8):1416−1419 (in Chinese with English abstract).

[8] Li ZF, Yarowsky D. Unsupervised translation induction for Chinese abbreviations using monolingual corpora. In: Proc. of the ACL 2008. Columbus: Association for Computer Linguistics, 2008. 425−433.

[9] Yang D, Pan YC, Furui S. Automatic Chinese abbreviation generation using conditional random field. In: Proc. of the HLT-NAACL. Stroudsburg: Association for Computational Linguistics, 2009. 273−276.

[10] Jia MY, Yang BR, Zhen DQ, Yang J. Research on automatic military intelligence term extraction using CRF model. Computer Engineering and Applications, 2009,45(32):126−129 (in Chinese with English abstract).

[11] Zhang HP, Liu Q, Zhang H, Cheng XQ. Automatic recognition of Chinese unknown words based on roles tagging. In: Proc. of the 1st SIGHAN Workshop on Chinese Language Processing. Stroudsburg: Association for Computational Linguistics, 2002. 1−7. [doi: 10.3115/1118824.1118841]

[12] Nie JY, Hannan ML, Jin WY. Unknown word detection and segmentation of Chinese using statistical and heuristic knowledge. Communications of COLIPS, 1995,5(1-2):47−57.

[13] Chen KJ, Bai MH. Unknown word detection for Chinese by a corpus-based learning method. Int'l Journal of Computational Linguistics and Chinese Language Processing, 1998,3(1):27−44.

[14] Wang M, Huang C, Chen K. The identification and classification of unknown words in Chinese: An *N*-grams-based approach. In: Proc. of the 1994 Kyoto Conf.: A Festschrift for Professor Akira Ikeya. Tokyo: Logico-Linguistic Society of Japan, 1995. 113−123.

[15] Sui ZF, Chen YR, Hu JF, Wu YF, Yu SW. The research on the automatic term extraction in the domain of information science and technology. In: Proc. of the 5th East Asia Forum of the Terminology. Haikou: CNIS Press, 2002. 444−451.

[16] Chen AT, Peng FC, Shan Y, Sun G. Chinese named entity recognition with conditional probabilistic models. In: Proc. of the 5th SIGHAN Workshop on Chinese Language Processing. Sydney: Association for Computational Linguistics, 2006. 173−176.

[17] Sutton C, McCallum A. An introduction to conditional random fields for relational learning. In: Getoor L, Taskar B, eds. Introduction to Statistical Relational Learning. MIT Press, 2006.

[18] Wallach HM. Conditional random fields: An introduction. Technical Report, MS-CIS-04-21, Philly: Department of Computer and Information Science, University of Pennsylvania, 2004.

[19] Wallach HM. Efficient training of conditional random fields [MS. Thesis]. Scotland: University of Edinburgh, 2002.

[20] http://www.sogou.com/labs/dl/w.html

[21] http://www.sogou.com/labs/dl/ca.html

[22] http://www.sogou.com/labs/dl/w.html

[23] Li HQ, Huang CN, Gao JF, Fan XZ. The use of SVM for Chinese new word identification. In: Proc. of the 1st Int'l Joint Conf. on Natural Language Processing (IJCNLP 2004). 2005. 723−732. [doi: 10.1007/978-3-540-30211-7_76]

附中文参考文献:

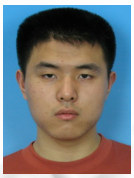
[5] 张华平,刘群.基于角色标注的中国人名自动识别研究.计算机学报,2004,27(1):85−91.

[6] 李丽双,黄德根,陈春荣,杨元生.SVM 与规则相结合的中文地名自动识别.中文信息学报,2006,20(5):51−57.

[7] 李丽双,黄德根,陈春荣,杨元生.用支持向量机进行中文地名识别的研究.小型微型计算机系统,2005,26(8):1416−1419.

[10] 贾美英,杨炳儒,郑德权,杨靖.采用 CRF 技术的军事情报术语自动抽取研究.计算机工程与应用,2009,45(32):126−129.

陈飞(1987－),男,重庆人,博士,主要研究领域为信息检索.
E-mail: chenfei27@gmail.com



张云亮(1986－),男,博士,CCF 学生会员,主要研究领域为网络安全.
E-mail: zhangyunliang1986@gmail.com

刘奕群(1981－),男,博士生,助理研究员,CCF 会员,主要研究领域为信息检索.
E-mail: yiqunliu@tsinghua.edu.cn

张敏(1977－),女,博士,副教授,CCF 高级会员,主要研究领域为信息检索.
E-mail: z-m@tsinghua.edu.cn

魏超(1987－),男,硕士,主要研究领域为信息检索.
E-mail: weichao053825@163.com

马少平(1961－),男,博士,教授,博士生导师,主要研究领域为知识工程,信息检索,汉字识别与后处理,中文古籍数字化.
E-mail: msp@tsinghua.edu.cn