基于统计学习框架的中文新词检测方法

张海军[1,2] 栾 静[1] 李 勇[1] 齐向伟[1]
(新疆师范大学计算机科学技术学院 乌鲁木齐 830054)[1]
(中国科学技术大学计算机科学与技术学院 合肥 230027)[2]

摘 要 新词自动检测是中文信息处理的重要基础,但中文字符极强的构词能力给新词检测带来了巨大困难。提出一种新词检测的形式化描述模型,用以建立特征和新词检测结果之间的统计联系。在此基础上提出应用统计学习模型作为框架来整合不同类型的可用特征,以充分发挥特征之间的组合作用,进一步改善新词检测效果。实验表明,统计框架方法的性能明显地优于特征的简单叠加,能有效提高新词检测效果,开放实验和封闭实验的 F 值分别为49.72%和69.83%,达到了目前的较好水平。

关键词 统计框架,新词检测,重复模式,语言知识特征,统计特征
中图法分类号 TP391 **文献标识码** A

Method of New Chinese Word Detection Based on Statistical Learning Framework

ZHANG Hai-jun[1,2] LUAN Jing[1] LI Yong[1] QI Xiang-wei[1]
(School of Computer Science and Technology, Xinjiang Normal University, Urumqi 830054, China)[1]
(School of Computer Science and Technology, University of Science and Technology of China, Hefei 230027, China)[2]

Abstract Automatic detection of new words is an important foundation in Chinese information processing, but Chinese has an extremely strong word-building ability, which brings great difficulties for new Chinese word detection. This paper put forward a formal model for new word detection, through which the relations between features and detection effects can be constructed. On this basis, this paper also proposed to employ high-effective statistical learning model as a framework to integrate diffirent kinds of available features, which can make full use of the combination of features to further improve the effects of new word detection. Experiments show that the performance of statistical framework is much better than that of simple sum of single features and the method of this paper can effectively improve the result of new word detection. F value in open and closed experiment is 49.72% and 69.83% respectively, which reaches a better level among current studies.

Keywords Statistical framework, New words detection, Repeats, Linguistic knowledge feature, Statistical feature

1 引言

词语是语言信息自动处理的基本单位。为使处理过程顺利地进行,必须对大量产生的新词进行检测和识别。新词检测在句法分析、词典编纂、机器翻译以及舆情监测等领域都有着重要应用。与印欧语言不同,中文没有特定符号来表示词语边界,因此任何相邻中文字符都有构词的可能性;且书面语中没有字符形态变化,这都给中文新词自动检测带来了巨大障碍。目前在中文新词自动检测的研究中,主要有基于单字散串和基于高频重复模式方法。因后者具有能有效地识别新造词、对语料依赖程度小、适应能力强以及召回率高等特点[1],近年来受到了广泛关注,也取得了较多研究成果。

2 相关研究

基于高频重复模式的新词检测包含2个基本步骤,即高频重复模式的提取和候选新词的过滤。前者从语料中提取重复模式,构造候选新词集合;后者对候选新词集合中的非词垃圾字串进行过滤,以提取新词。目前大量研究都集中在候选新词的过滤方面。刘挺等[2]使用滑动窗口来提取局部重复模式,构造候选词集合,然后应用经验函数来检测新词,研究表明,该方法能有效提高中文分词效果;郑家恒等[3]使用递增的n-gram模型提取重复模式,在此基础上使用手工编制的提取和过滤规则(包括常用构词规则、特殊构词规则和互斥性字串过滤规则)从互联网语料中提取新词;邹刚等[4]在文献[3]方法的基础上,使用正则表达式来表示过滤规则,实现任意长度

到稿日期:2011-03-24 返修日期:2011-06-26 本文受国家自然科学基金(61163045),新疆师范大学博士博士后科研启动基金(XJNUBS1111)资助。

张海军(1973—),男,博士,讲师,主要研究方向为自然语言处理、新词识别技术,E-mail:ustczhj@mail.ustc.edu.cn;栾 静(1962—),女,博士,教授,主要研究方向为自然语言处理、舆情监测;李 勇(1983—),男,硕士,讲师,主要研究方向为自然语言处理、知识库构建技术;齐向伟(1981—),男,硕士,讲师,主要研究方向为自然语言处理。

新词的检测；崔世起等[5]将新词结构分成了不同的组成形式，如1+1、1+2、2+1、1+1+1等(其中1+1表示由2个单字构成的2字词，1+2表示由1个单字和1个双字词构成的3字词)，并针对不同的组成结构采用特定的处理方法。Luo Shengfen等[6]针对2字串，将多种字串的内部统计特征，包括出现频率、互信息、色子系数等9种特征组成了一个加权词语抽取模型，配合左右熵来进行词语抽取；罗智勇等[7]以支持向量机(SVM)为统计模型，使用左右熵、似然比和相关频率比作为特征进行武侠小说中新词的检测；贺敏[8,9]在重复模式提取的基础上，应用外部环境和内部特征相结合的方法来检测新词，研究中主要使用了上下文邻接分析、位置成词概率和双字耦合度，达到了较好的新词检测效果。

目前的新词检测研究主要集中在新特征的挖掘和使用上，但因没有可靠模型的指导，特征选择还存在一定的盲目性；对特征的使用一般也仅限于单个特征或类型相似特征的简单组合，尚未考虑将语言知识特征和统计特征等不同类型特征进行有效整合，以实现组合特征的综合作用和更好的新词检测效果。本文在候选新词集合基础上，根据概率论的相关原理，提出一种新词检测的形式化描述模型，用于建立特征和新词检测结果之间的有效联系，并提出在新词检测中应用统计模型作为框架，以有效地整合新词的语言知识和统计这两种不同类型的特征，改进和提高新词检测效果。

3 基于统计学习框架的新词检测方法

3.1 新词检测的形式化描述

在已经取得重复模式(候选新词)集合的前提下，新词检测的任务就转化为以重复模式的各种有效特征作为判别标准。判断其是否是新词的过程，实际是在可用特征的基础上对候选新词进行标注的过程。根据概率论的相关理论，候选新词标记 $\hat{t}$ 的最大似然估计可表示为

$$\hat{t}=\underset{t\in\{新词,非新词\}}{\operatorname{argmax}} P(t|候选新词) \tag{1}$$

其中候选新词的标记结果集合为{新词，非新词}。该式可进一步转化为

$$\hat{t}=\underset{t\in\{新词,非新词\}}{\operatorname{argmax}} \frac{P(候选新词|t)}{P(候选新词)}=\underset{t\in\{新词,非新词\}}{\operatorname{argmax}} P(候选新词|t) \tag{2}$$

新词本身具有未知性，候选新词本身和标记之间没有先验知识，也就是说候选新词同标记 t 之间的条件概率是未知的。如果是已知的，那么这个词就不能称之为“候选新词”了。为了解决这个问题，考虑对候选新词进行本质特征的分解，这样即可通过本质特征与标记 t 之间的关系来求解候选新词整体与标记之间的关系。但前提是，分解出来的特征要能充分体现候选新词的本质特性。在具体处理时，可用本质特征的集合来代表候选新词。这样候选新词与标记 t 之间的关系就转化为本质特征与标记 t 之间的关系，实际上是在候选新词的本质特征与标记 t 之间建立起了有效的联系。根据以上的论述，新词检测过程可进一步描述为

$$\hat{t}=\underset{t\in\{新词,非新词\}}{\operatorname{argmax}} P(F_S|t) \tag{3}$$

式中，F_S 表示能代表候选新词的本质特征集合。若根据上式对候选新词的特征和标记进行训练，因特征之间关系复杂，难以直接进行特征的训练和标注。为了解决特征之间的独立性问题，考虑应用有效的统计框架处理以上模型。鉴于条件随机域模型(CRF)在自然语言处理领域的广泛应用，并且不要求所用特征之间具有独立性，因而非常适合以上模型的求解，可用之有效地整合能代表候选新词的各类本质特征。

3.2 条件随机域模型(CRF)

条件随机域是一种无向图模型，对于确定结点的输入值，它能够计算该结点输出值上的条件概率，其训练目标是使得条件概率最大化。设 $x=x_1\cdots x_T$ 为给定的输入观察值数据序列，也就是无向图模型中 T 个输入结点上的数据，比如某个候选重复模式的所有特征所组成的数据序列；定义 Y 为有限状态机的状态集合，每个状态可以对应一个标记；设 $y=y_1\cdots y_T$ 为一个长度与 x 相等的状态序列，即无向图模型中 T 个输出值。在带有参数的线性链条件随机域模型的作用下，从给定输入序列 x 得到的输出序列 y 的条件概率表示为

$$P_\Lambda(y|x)=\frac{1}{Z_\Lambda(x)}\exp\left(\sum_{t=1}^{T}\sum_k\lambda_k f_k(y_{t-1},y_t,x,t)\right) \tag{4}$$

式中，$Z_\Lambda(x)$ 是一个规范化系数，它确保在给定输入上所有可能的状态序列的概率之和为1。规范化系数 $Z_\Lambda(x)$ 的计算涉及到的状态序列数目非常巨大，一般呈指数级增长。但在线性链模型中，状态结点间没有闭合路径，可通过动态规划算法便捷地计算规范化系数，且寻找最可能状态序列的问题也可用动态规划方法加以解决。上式中的 $f_k(y_{t-1},y_t,x,t)$ 表示一个特征函数，其值一般为布尔类型，满足特定条件时为1，否则为0。比如在新词检测中，当所给特征满足新词的条件时，该函数的值为1，不满足时为0。λ_k 是在训练中得到的、与每个特征函数 f_k 相关的权重参数。如果它为较大的正数，则事件更可能发生；如果为较大负数，则事件倾向于不发生[10]。

条件随机域模型的主要优点是：(1)能够综合利用字、词、词性等多层次资源，能更好地使用领域知识和标记之间的依赖，充分利用各种语言知识特征和统计特征；(2)该模型对特征没有独立性要求，在使用时无需考虑特征之间是否相互独立，因此可将多个代表候选新词的本质特征放入CRF框架中，以实现各类特征的综合作用，改善新词检测效果。

根据CRF模型的特点，可不用考虑特征之间的关系，将之直接加入到CRF框架中，测试特征对新词检测效果所做的独立或组合贡献，以确定能代表候选新词的本质特征集合，提高新词检测效果。

3.3 新词检测所用特征集合

在CRF框架下，训练和解码所选用的语言知识特征包括前缀、后缀、串长、命名实体后缀；统计特征包括候选模式的出现频率、互信息、色子系数和左右熵。上述特征并不复杂，应用CRF统计模型可充分利用各类特征，实现更有效的新词检测。其中，前缀、后缀、串长是用于词语检测的基本语言特征，命名实体后缀用于识别新词中的命名实体，是首次应用在CRF模型中的语言特征，该特征在使用时根据候选字串所具有的命名实体后缀的长度来构造；其他的统计特征，如互信息和左右熵等，用于衡量新词结构的独立性和在上下文中使用的灵活性。

对其中的数值型特征，因CRF模型在训练和解码时将特征值作为字符串来处理，所以需对连续的数值型特征进行离散化并转化成字串特征，从而将无限量的连续数值特征量转化为有限的离散字串特征量，提高了训练和解码效率，并可有效地改善新词检测效果。为方便后续分析处理，对以上特征进行编号，具体见表1所列。

表 1　特征编号对照表

编号	1	2	3	4	5	6	7	8	9	10	11
特征名称	前缀	双字前缀	后缀	双字后缀	命名实体后缀	串长	串频	互信息	色子系数	左熵	右熵

3.4　基于统计框架的新词标注

基于重复模式新词检测的基本步骤是，首先在语料中提取满足阈值约束的重复字串，构造候选新词集合，然后根据候选新词的相关特征，比如出现频率、前缀、后缀以及其它信息来判断和标注候选新词集合中的条目。为有效提高多类特征的组合作用，本文采用具有更强包容能力的 CRF 模型作为统计框架整合多类特征，以实现更好的新词检测效果。在具体实施新词检测时，首先提取候选新词的各类特征（见表 1），然后根据特征值，应用标注语料对 CRF 模型进行训练，最后使用 CRF 模型标注从测试语料中提取的候选新词。标注的结果只有两个："是新词"与"非新词"。

4　实验

4.1　实验及数据分析

为验证本文方法的性能，进行了如下试验：实验所用的 CRF 工具采用日本 Kudo 教授所提供的开源工具"CRF++ 0.52"[11]，训练语料采用北京大学计算语言研究所提供的 1998 年 1 月的标注语料，测试语料采用兰开斯特大学标注的汉语平衡语料库。实验时首先对训练语料和测试语料抽取重复模式，构造候选新词集合（提取重复模式所用的阈值为 2），然后使用统计框架对其中字串进行标记。为检验新词的开放实验效果，需要确保训练语料与测试语料所提取的候选新词集合没有交集。实验中所用性能的评价标准为准确率、召回率和 F 值，F 值中所用的调和因子为 1，开放实验结果参见表 2。

表 2　多特征组合新词检测实验数据表

编号	特征集合	准确率(%)	召回率(%)	F 值(%)
1	1—7	33.44	59.46	42.81
2	1—8	37.68	61.16	46.63
3	1—9	39.71	59.5	47.63
4	1—11	41.4	62.23	49.72

注：特征集合"1—7"表示将编号从 1 到 7 的特征组合在一起构成特征集合，作用于条件随机域模型。其它的也类似。

从实验数据可见，随着有效特征的加入，新词检测的效果在稳步提高。说明在条件随机域框架下，多特征组合可以实现更好的新词检测。这也进一步说明，多特征组合可以发挥特征之间的综合作用，比特征的简单组合具有更好的新词检测效果，实验 4（特征集合 1—11）已很好地说明了这个问题。为研究本文方法的效果与串长之间的关系，现对编号为 3 和 4 的试验数据做进一步分析，绘制串长和新词检测效果图，如图 1 所示。

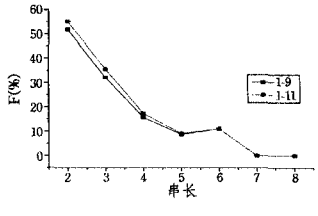

图 1　新词检测的串长-效果关系图

从图中可见，无论使用哪种特征组合，串长和新词检测效果之间关系的变化趋势都是一致的：随着串长的增长，新词检测的效果在逐渐降低，短串具有更好的新词检测和提取效果。长串效果较差的主要原因是组成长串的字符较多，组合情况更加复杂；而短串的组合情况相对较少，其更适合在组合特征的作用下进行标注。可见，要改善新词检测效果，应从长串着手进行研究和改进。

4.2　最大熵框架下的新词检测

最大熵（ME）模型也是一个重要的统计框架，同 CRF 相似，都属于判别型模型。二者在处理标注问题时具有很多共同的优点，主要表现在 ME 模型也对特征没有独立性要求，也可作为统计学习框架对候选新词实施过滤。为了进一步验证本文所提出的新词检测方法的效果，对最大熵模型进行新词检测实验是非常必要的。实验时采用相同的实验语料和条件，使用特征组合 1—9 和 1—11（在 CRF 框架中取得较好效果的特征组合），结果见表 3。

表 3　ME 框架下多特征组合新词检测结果

编号	特征集合	准确率(%)	召回率(%)	F 值(%)
5	1—9	37.89	58.41	45.96
6	1—11	39.93	59.47	47.78

从表中可见，在最大熵统计框架下实验 6 的效果较实验 5 好些，也证明增加有效特征会提高新词检测效果；从横向上看，实验 5 和实验 6 分别比试验 3 和实验 4 的效果稍差，说明条件随机域模型比最大熵模型具有更好的新词检测性能，其主要原因是 CRF 模型是一种全局最优模型，且具有更强的特征融合能力。当然，如果有更好的统计模型出现，也许会取得比 CRF 更好的效果。同单个特征相比，CRF 和 ME 都取得了较好的特征组合效果，可见，应用统计框架来进行特征整合是一个很有前途的新词检测研究方向。

4.3　模型比较

文献[8]对新词检测特征进行了深入探索并进行了较全面的实验，取得了相对较好的检测效果，但其采用的是封闭实验。为加强可比性，本文也采用封闭实验环境重新进行试验（所用特征组合为 1—11，统计框架为 CRF 模型和 ME 模型），新词检测效果对比见表 4。

表 4　不同方法新词检测效果对比表

新词检测方法	准确率	召回率	F 值
文献[8]方法	45.96%	71.19%	55.86%
本文方法(CRF)	69.15%	70.53%	69.83%
本文方法(ME)	66.49%	69.33%	67.88%

从以上对比数据可见，文献[8]方法的召回率比本文方法高，本文方法在准确率和总体性能（F 值）方面具有一定优势。文献[8]中采用的实验语料规模比本文中的要小得多，这会导致其中所用新词检测特征不能完全发挥作用，所以会在一定程度上影响其新词检测效果。但从理论上讲，本文所用的统计框架方法（无论是 CRF 模型还是 ME 模型），能有效地整合不同类型的多个特征，体现特征之间的合力作用，可实现更加有效的新词检测，是新词检测研究的发展方向。

结束语　采用统计模型作为框架，实现对新词检测特征的有效整合，以获得更好的新词检测效果。实验表明，本方法能充分发挥多特征的组合作用，随着特征的加入，新词检测效果在逐步提高。最终开放实验和封闭试验的 F 值分别为 49.72%和 69.83%，达到了较好的新词检测效果，证明使用统计模型作为框架整合有效特征是一种非常有前途的新词检测研究方法。由于使用相似的处理步骤，本文方法可方便地扩展到基于重复模式的命名实体、有意义串的识别研究中，以获得更好的识别效果。

本研究下一步工作是充分挖掘有效的新词检测特征并将

其放入统计框架中，以进一步改进新词检测效果，为基于海量语料的机器翻译和舆情热点发现提供支持。

参 考 文 献

[1] 张海军，史树敏，朱朝勇，等. 中文新词识别技术综述[J]. 计算机科学，2010，37(3)：6-10

[2] 刘挺，吴岩，王开铸. 串频统计和词形匹配相结合的汉语自动分词系统[J]. 中文信息学报，1998，12(1)：17-25

[3] 郑家恒，李文花. 基于构词法的网络新词自动识别初探[J]. 山西大学学报：自然科学版，2002，25(2)：115-119

[4] 邹纲，刘洋，刘群，等. 面向 Internet 的中文新词语检测[J]. 中文信息学报，2004，18(6)：1-9

[5] 崔世起，刘群，孟遥，等. 基于大规模语料库的新词检测[J]. 计算机研究与发展，2006，43(5)：927-932

[6] Luo S，Sun M. Two-character Chinese word extraction based on hybrid of internal and contextual measures[C]// Proceedings of the Second SIGHAN Workshop on Chinese Language. Sapporo，Japan，2003：24-30

[7] 罗智勇，宋柔. 基于多特征的自适应新词识别[J]. 北京工业大学学报，2007，33(7)：718-725

[8] 贺敏，龚才春，张华平，等. 一种基于大规模语料的新词识别方法[J]. 计算机工程与应用，2007，43(21)：157-159

[9] 贺敏. 面向互联网的中文有意义串挖掘[D]. 北京：中国科学院研究生院，2007

[10] Sutton C，McCallum A. An Introduction to Conditional Random Fields for Relational Learning[M]. Cambridge MA：MIT Press，2006

[11] CRF++：Yet Another CRF toolkit[EB/OL]. http://chasen.org/~taku/software/CRF++，2009-05-01

(上接第 205 页)

4.2 讨论

1)收敛性：实验表明，无论参变量如何变化，和初始分布不同，算法都能收敛。即最终磁化率大多在经过短时间($t=6$)后，其波动幅度明显减少，逐渐趋向一个稳定值。这与舆情传播的趋势相仿，表明系统能从非平衡态趋向平衡态。仿真表明舆情传播起始 4 个周期内变化梯度较大，其逐步递减，第 4 至第 6 周期变化趋势平缓。

2) K 的影响(见图 2、图 3)：当 $h=0$，即不考虑偏好度的影响时，无论正反向初始分布有何不同，k 值在$[0,+1]$范围，最终磁化趋向零，说明环境影响较小时，舆情总体分布“正向”“负向”比例相近，但当 K 值继续增大到$[2,8]$时，环境影响度使得舆情状态向初始分布中多数人意见方向转变。

3) h 的影响(见图 4－图 7)：偏好度 h 为正(见图 4、图 5)代表网络舆情主体对“正向”本体的偏好程度。无论初始分布“正向”概率大或是“负向”概率大，随着时间增加，磁化率迅速从初始值(0.40 或－0.40)分别上升到与 h 值大致相等的稳定值，说明 h 对舆情传播起着主导作用，最终稳定值与初始分布无关。偏好度 h 为负值(见图 6、图 7)代表网络舆情主体对“负向”本体态度的偏好程度，与上述相仿，最终磁化率稳定在 h 值附近。

4) 自转率 v 的影响：由式(11)和式(12)分析，当 $k=0$，$h=0$时，$P=v$，即状态转移概率就是自转率，此时网络舆情空间状态概率仅与自转率和初始分布状态概率相关，v 值影响磁化率起始值的大小。

5) 初始分布的影响：当初始“正向”概率大于“负向”概率时，序参量 k 在$[0,8]$，h 在$[0,1]$范围时，磁化率始终在$[0,1]$的“正向”范围内变化，显示向多数意见靠拢的趋势。当初始时主体持“正向”本体概率小于“负向”概率时，k 在$[0,1]$的变化只引起磁化率在$[0,-1]$“负向”范围内变化，同样显示向“负向”多数意见靠拢的趋势。但 $k=0$ 时，无论 h 为正还是负，最终分布与初始分布无关，强烈趋向靠拢 h 值大小的分布概率。

结束语 本文研究网络舆情传播预测的计算机模型、算法及其仿真。首先，对 IPO 传播过程马尔科夫链引入状态一步协同转移概率，提出一个协同-马尔科夫模型，设计了算法并进行了计算机仿真。同时对仿真的结果做了详细的讨论。

1)当 k 值增大时，环境影响使得网络舆情主体从众心理加大，传播向初始分布的多数意见聚拢。

2)当 h 值增大时，偏好使网络舆情主体根据对本体状态偏好度的正或负分别向磁化率正或负聚拢。

3)网络舆情开始 1～4 次传播变化比较剧烈，需加以注意。同时密切关注偏好度影响起的“正向、负向”放大作用，此时不受初始多数分布的倾向影响。

4)该算法计算复杂度为 $O(t)$。

由于协同-马尔科夫模型对舆情空间整体状态进行处理，因此计算耗时较少，且因舆情空间整体协同影响强烈，故适合预测 IPO 的传播。未来将研究协同-马尔科夫模型与协同元胞自动机的不同比较，并研究参变量在现实 IPO 中的物理特性和数学表达，以便将仿真结果与 IPO 实际传播数据进行比对，进一步分析模型误差及其原因，提高预测精度。

参 考 文 献

[1] 刘建明. 舆论传播[M]. 北京：清华大学出版社，2001

[2] 张玉峰，王志芳. 基于内容相似性的论坛用户社会网络挖掘[J]. 情报杂志，2010，29(8)：125-130

[3] 章栋兵. 互联网舆情分析关键技术的研究和实现[D]. 武汉：武汉理工大学，2010

[4] Zeng J P，Zhang S Y，Wu C R，et al. Modelling Topic Propagation over the Internet[J]. Mathematical and Computer Modelling of Dynamic Systems，2009，15(1)：83-93

[5] Zeng J P，Zhang S Y，Wu C R，et al. Predictive Model for Internet Public Opinion[C]// Proceedings of Fourth Conference on Fuzzy System and Knowledge Discovery. 2007

[6] Alves S G，Oliveira Neto N M，Martins M L. Electoral Surveys' Influence on the Voting Processes：A Cellular Automata Model [J]. Physica A：Statistical Mechanics and Its Applications，2002，316(1-4)：601-614

[7] 刘慕仁，邓敏艺，孔令江. 舆论传播的元胞自动机模型(I)[J]. 广西师范大学学报：自然科学版，2002，20(2)：1-3

[8] 曾祥平，方勇，袁媛，等. 基于元胞自动机的网络舆论激励模型[J]. 计算机应用，2007，27(11)：2686-2688

[9] 方薇，何留进，等. 采用元胞自动机的网络舆情传播模型研究[J]. 计算机应用，2010，30(3)：751-755

[10] 曾显葵. 基于多数规则和协同规则的元胞自动机舆论传播模型研究[D]. 南宁：广西师范大学，2007

[11] 哈肯 H. 协同学导论[M]. 张纪岳，等译. 西安：西北大学出版社，1981

[12] 沈小峰，胡刚，江璐，等. 耗散结构论[M]. 上海：上海人民出版社，1987