第33卷 第10期 计 算 机 工 程 2007年5月
Vol.33 No.10 Computer Engineering May 2007
·博士论文· 文章编号：1000—3428(2007)10—0016—03 文献标识码：A 中图分类号：TP18

基于CRF的百科全书文本段落划分

许 勇[1]，宋 柔[2]
（1. 北京工业大学计算机科学学院，北京 100022；2. 北京语言大学计算机科学系，北京 100083）

摘 要： CRF模型是标注、切分序列数据的较新的概率模型，在信息抽取等文本处理领域广受关注。该文介绍了CRF方法，并将其应用到百科全书文本段落的划分上，利用CRF的特征表述机制加入了文本单元序列中的长距离约束，取得了比传统的隐马尔科夫方法更好的结果。

关键词： 文本段落划分；条件随机域模型；隐马尔科夫模型

Encyclopedia Text Topic Segmentation Based on CRF

XU Yong[1], SONG Rou[2]
(1. Institute of Computer Science, Beijing University of Technology, Beijing 100022;
2. Dept. of Computer Science, Beijing Language and Culture University, Beijing 100083)

【Abstract】 Conditional random field(CRF) is a newly proposed probabilistic model for segmenting and labeling sequence data, and has been successfully applied to many natural language processing tasks and information extraction. This paper introduces CRF model and applies it in encyclopedia text topic segmentation. With its long distance overlapping feature mechanism, the CRF model shows better performance than traditional HMM model on encyclopedia text segmentation task.

【Key words】 Topic segmentation; Conditional random fields(CRF); Hidden Markov model(HMM)

1 简介

文本段落划分的目标是将一篇文档划分为前后连接的几个段落，使得同一个段落中的文本内容比较单一集中，段落之间则相对独立。段落划分是一种较大颗粒度的文本加工手段，在信息检索、话题识别与跟踪(Topic Detection and Tracking)、文本知识获取等领域有重要应用。

文本段落划分大致可分为无段落类型判断的划分和有段落类型判断的划分。无类型的判断的划分仅对文档进行划分，不对段落作其类型的判断。这一类的划分中代表性的方法是Marti A. Hearst[2]等人提出的TextTiling算法，该算法依据候选划分点前后一定范围内文本窗口之间的相似度来划分段落，相似度是以向量空间方法度量的。有段落类型判断的划分则还需对段落做类别的判断，其中类别一般是预先给定的。这一类的划分一般应用于具有一定内容结构的文档集合。例如，常见问题回答文件(Frequently Asked Question)中问题和回答部分的切分；话题识别与跟踪中新闻播报语音识别文本流中的新闻单位的切分和类型的识别等。在这种划分中，一般采用基于马尔科夫性的概率模型，将段落划分的问题转化为文本处理单元(一般选为一个句子)顺序标注的问题，连续的相同标注即为一个段落，标注的变化即为段落的划分点。常见的有隐马尔科夫模型(HMM)[5]，最大熵马尔科夫模型(MEMM)[4]等。在这种方法中，求取的是对一个文本实例具有最大概率的状态序列(即标注序列)。

《中国大百科全书》中同题材的词条文本属于有比较清晰的内容结构的文档集合。针对百科全书的文本，为便于叙述，下文中将题材相关的内容成份称为该题材的知识点。利用文本段落划分技术，将百科全书同题材条目文本中的知识点划分成段落并判断其类型，是百科全书文本处理中的重要一环。通过这一步可以提供段落一级的知识点的直接检索；可以为后续的段落内部知识项目的发现与获取提供有利的条件。

本文采用条件随机域(Conditional Random Field, CRF)模型来处理百科全书同题材条目文本的有类型判断的知识点划分问题。CRF模型是较新的标注、切分序列数据的较新的概率模型[1]，也是基于马尔科夫性的模型，近年来比较受到关注，在自然语言处理、信息抽取等领域都有应用[3]，但文本段落划分方面尚未见到应用实例。下面简单介绍CRF模型。

2 CRF模型简单介绍

CRF模型定义了在序列数据$X(x_1x_2\ldots x_n)$上的状态变量序列$Y(y_1y_2\ldots y_n)$的条件概率$P(Y \mid X)$，其中每个位置上的状态变量y_i只依赖于序列数据X和与其相邻的状态变量。这是常用的线性链结构的CRF模型，CRF模型的完整的定义可参考文献[1]。条件概率$P(Y \mid X)$是以一组特征函数来计算的。局部特征函数是定义在序列数据X，序列中的位置i，及随机变量y_i、y_{i-1}上的函数$f(y,x,i)$，不加下标的y和x表示整体状态序列和数据。局部特征函数并不限定只考虑当前位置i上的数据，只是限定状态序列y为当前位置上的状态，或是当前位置上的转移。$\bar{f}(y,x,i)=<f^1,f^2,\ldots,f^K>$为由$K$个局部特征函数组成的特征函数向量，全局特征函数定义为其在序列数据上的累加：

$$F(x,y)=\sum_i \bar{f}(y,x,i)$$

基金项目： 国家自然科学基金资助项目(60272055)；国家“863”计划基金资助项目(2001AA110372-1)

作者简介： 许 勇(1975－)，男，博士生，主研方向：自然语言处理；宋 柔，教授、博导

收稿日期： 2006-06-20 **E-mail：** hopexy163@163.com

则对给定的序列数据 X，一个状态序列 Y 的概率为

$$P(Y \mid X, W) = \frac{\exp(W \cdot F(Y, X))}{Z(X)}$$

其中，W 为特征函数的权重向量。

$$Z(X) = \sum_{Y'} \exp(W \cdot F(Y', X))$$

为归一化项，保证所有可能的状态序列上的概率值之和为 1。规定一个 CRF 模型就是选定一组反映数据特点的特征函数。CRF 模型的训练过程实际上就是权重向量 W 的选择，使得训练集合上的对数似然函数取最大值。常用的训练算法有基于梯度的最优化算法和最大熵模型中的 IIS(Improved Iterative Scaling)算法。和 HMM 模型类似，对一个数据实例，CRF 模型计算对当前数据实例具有最大概率的状态序列作为结果，这一过程也可以用类似 Viterbi 算法的动态规划算法。

CRF 模型中的特征函数是非常有力的机制。它和最大熵模型中的特征函数一样，一般取为布尔函数形式的提示函数，如一个简单的状态特征函数：

$$f(x, y, i) = [x_i = a] \cdot [y_i = \alpha]$$

这里 $[c]$ 表示提示函数，即当 c 为真的时候 $[c]=1$，其他情况为 0。这样简单的状态特征函数经过训练之后，它的权重值大致上对应(不是等价)于 HMM 中的状态 α 条件下观察值 α 的概率 $P(a \mid \alpha)$，但是特征函数可以参考整个数据序列，从而引入更大范围的特征，如

$$f(x, y, i) = [x_i = a] \cdot [b \in \{x_j \mid j < i\}] \cdot [y_i = \alpha]$$

就检测当前位置之前是否出现了特定词汇 b。另外，特征函数可以是重叠的，如，一个特征可以检测当前位置上是否为某个词汇，另一个特征可以检测该位置的词是否以大写字母开始。这样，特征函数的表达能力就大大增加了。特征函数之所以有这样的能力是因为在CRF模型中，条件概率 $P(Y \mid X)$ 是直接使用特征函数的取值和权重向量计算而得的，但是在HMM中，上述概率的处理要通过贝叶斯公式转化为状态条件上的数据的概率 $P(X|Y)$，这样，对序列数据中的单元 $x_1, x_2, \ldots, x_n$ 之间互相条件独立的假设就不可避免，也就难以处理序列数据中的长距离的相互间的影响和重叠的特性。像HMM模型一样要借助于状态条件下数据的概率的模型称为产生型的模型(Generative Model)，而像CRF一类的模型称为条件型模型。

特征函数在实际应用中一般取为布尔型函数，但是理论上并没有这样的限制，特征函数可以取任意实数为值。

3 基于 CRF 的段落划分

段落划分试验是在《中国大百科全书》中国地理卷中的市、县一级的行政地名共 728 个条目文本上进行的。这一题材的条目文本中含有 17 个知识点类型，分别是概述、地处方位、面积人口、行政中心、下辖区县、历史沿革、地理环境、气候、农业、矿产资源、工业、城区概况、交通、文教、旅游、辖县情况、附加。其中下辖区县只包含辖区或辖县的名称和数目，辖县情况则会挑选一二个重要的具体加以介绍。附加这个知识点出现次数非常少，而且只在最后出现，且其内容芜杂，难以归类，故将这样的内容归为附加。下面是一些知识点类别的实例：

概述：河南省郑州市属县，河南工、农业发达县份。

行政中心：县府驻孝义镇。

工业：县内乡镇企业占全县工农业总产值 62.3%，居全省首位。

并不是每个条目文本都包含全部 17 个知识点类型。实际上，包含全部知识点类型的条目很少。一般一个条目文本包含 7~8 个知识点类型。

通过观察发现这一题材的条目文本中含有大量专名类型，如“ 油菜 ”、“ 花生 ”等各种农产品名称，“ 化工 ”、“ 机械 ”、“ 造船 ”等工业部门名称以及企业名等。如果这些词汇以词形本身计为特征，则由于频率过低，无法体现其类型相关性。因此设计了针对中国市县行政地名条目文本的小规模的词汇语义类型系统，归并了常见的专名类型和一部分同义词如‘ 降雨量 ”和“ 降水量 ”，以词汇类型标记替换了原来的词形。

在这个词汇集处理的基础上，我们随机选择了 143 个标注好段落的条目文本作为训练集，剩余的 585 个条目文本测试集上做了 3 个试验，分别是 HMM 模型段落划分，简单特征的 CRF 模型段落划分，扩充特征的 CRF 模型段落划分。这 3 个试验都是以句子为数据单位进行的，即段落的可能的转移限定为每个句子的末尾。HMM 中的参数平滑方法采用了简单的加一方法。简单特征的 CRF 模型，指的是特征函数只有提示函数形式的转移特征函数、状态词汇特征函数和开始状态特征函数，即

$$state_{ij}(x, y, i') = [w_i \in \{x_{i'}\}] \cdot [y_{i'} = s_j]$$

$$trans_{ij}(x, y, i') = [y_{i'-1} = s_i] \cdot [y_{i'} = s_j]$$

$$start_i(x, y, i') = [y_{i'} = s_i] \cdot [i' = 0]$$

其中，$\{x_{i'}\}$ 为 i' 处句子的词汇集合。这 3 种特征函数分别对应于 HMM 模型中的状态-观察指概率，状态转移概率，状态的初始分布概率。表 1 是试验结果。

表 1 HMM 段落划分和简单特征 CRF 段落划分结果

	段落个数	句子个数	CRF		HMM	
			正确率(%)	召回率(%)	正确率(%)	召回率(%)
概述	572	703	99.4	98.5	98.7	98.5
地处方位	544	600	95.2	94.5	97.2	96.3
面积人口	562	612	96.3	98.3	97.1	96.4
行政中心	213	217	98.5	96.7	98.8	82.4
下辖区县	134	134	92.3	99.2	95.6	97.7
历史沿革	536	2 137	93.5	97.1	95.2	96.6
地理环境	428	965	82.6	91.6	88.3	90.3
气候	231	374	92.9	88.2	94.5	88.4
农业	409	1 098	86.9	88.8	86.8	92.6
矿产资源	149	206	74.8	70.8	77.5	59.6
工业	464	1 166	87.1	85.8	81.1	88.1
城区概况	96	235	76.4	56.5	83.8	42.4
交通	339	758	85.1	88.1	79.3	93.4
文教	83	123	88.6	63.4	95.6	53.2
旅游	399	1 029	87.9	87.3	83.9	94.7
辖县情况	54	148	36.3	27.0	0	0
附加	43	59	0	0	0	0
合计	5 256	10 564	划分正确的句子数: 9 434 89.3%		划分正确的句子数: 9 455 89.5%	

从结果中可以看出，HMM 模型和只采用简单特征的 CRF 模型的效果是大致一样的。从表 1 中的知识点的正确率和召回率来看，最不理想的是“ 辖县情况 ”和“ 附加 ”两种类别。“ 附加 ”这种类型本身数量少，而且其内容为具体地名条目特有的附加性信息，内容零散，不具有一般性，暂时不做考虑。“ 辖县情况 ”这种知识点就不同了，其划分结果不理想的原因是这个知识点不具有较集中的词汇分布。这个知识点类型实际上可视为一种嵌套的结构。下面是一个典型的“ 辖县情况 ”的段落实例。

宁安县

……

13:宁安镇位于县境北部牡丹江畔，有牡图铁路通过。

14:人口近 6 万,是全县政治、经济、文化、交通中心。

15:东京城镇是进出境泊湖地区的门户和牡丹江上游的木材集散地。

16:东京城西约 3km 的渤海镇以产“响水大米”著名。

17:镇西北有渤海国都城上京龙泉府故城遗址，是全国重点文物保护单位。

18:县境山青水碧，景色秀丽，有镜泊湖、小北湖、桦树川水库火山口森林(俗称地下森林)，火山熔岩洞穴、宁安镇大石桥及清代抗俄名将萨布素将军墓等旅游地。

……

这个条目中编号 13~17 的句子为“辖县情况段落”，依序简要说明了宁安县的辖镇宁安镇和东京城镇的情况，其中涉及交通、人口、物产、旅游资源等内容，大致结构和宁安县本身的内容结构相似，但却是作为地名条目的一个内容段落出现的。亦可看出，这种类型不具有集中的内容相关的词汇分布，很难仅仅以词汇-知识点的关联程度作为判断依据。但是这种知识点的转入和转出的句子具有较明显的规律性，转入的句子中，将要说明的下级行政地名以一定模式出现在句首，如宁安镇位于……、市属叶县……县府驻地大良镇……，转出的句子在句首含有指称本条目行政级别(高于被说明的下级地名)的特征性词汇，如县境、市境，或者是本条目的地名。根据这个规律归纳了几个“辖县情况”的转入模式和转出模式，然后将转入模式和转出模式的顺序第 1 次出现作为*辖县情况*的特征窗口，增加了针对“辖县情况”类型的特征窗口位置特征和状态转入、连续、转出特征。设 s_s 为“辖县情况”对应的状态，*subwnd* 为“辖县情况”的特征窗口，则

$$s\_pos(x,y,i)=|\{x_i\}|\ y_i=s_s\ \text{且}\ i\text{位于}\ subwnd\text{内}$$

$$s\_in(x,y,i)=y_{i-1}\neq s_s\ \text{且}\ y_i=s_s\ \text{且}\ i\text{为}subwnd\text{内第1个位置}$$

$$s\_cont(x,y,i)=y_{i-1}=s_s\ \text{且}\ y_i=s_s\ \text{且}\ i\text{为}subwnd\text{内第1个以后的位置}$$

$$s\_out(x,y,i)=y_{i-1}=s_s\ \text{且}\ y_i\neq s_s\ \text{且}\ i\text{为}subwnd\text{后第1个位置}$$

s_pos 中 $|\{x_i\}|$ 为 i 处句子的词汇集合的大小。经过这样扩充后的 CRF 模型能够较好地划分“辖县情况”知识点。在“辖县情况”知识点上的召回率和正确率分别为 96.6%和 98.6%，且总的错误划分句子数减少了 131 个，正确划分的句子总数的百分比提高为 90.5%，比 HMM 模型高出了一个百分点。当然，在 HMM 模型中和 CRF 模型中也可以通过将词汇集扩展为“辖县情况”特征窗口内外两种词汇也可以减少“辖县情况”的错误，但这样的话，词汇集会翻倍增加，不利于进一步的扩充。

4 结论及今后工作

本文简单介绍了 CRF 模型以及基于 CRF 模型在百科全书文本段落划分上的应用实例。CRF 模型在表达序列数据内的长距离相关特性和重叠特性方面比 HMM 等产生型概率模型有优势。本文中的实验情况表明，虽然和 HMM 模型相比，性能的提高比较有限，但却可以利用 CRF 模型的特点有针对性地解决一类问题。

有类型判断的文本段落划分是比较复杂的问题。目前引入的特征还不足以表示段落作为整体的细微的内部连续和段落间的转移规律，今后需要在这方面多做些进一步的工作。

参考文献

1 Lafferty J, Callum A, Pereira F. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data[C]//Proceedings of the International Conference on Machine Learning. 2001.

2 Hearst M A. Multi-paragraph Segmentation of Expository Text[C] //Proceedings of the 32nd Meeting of the Association for Comput -ational Linguistics. 1994.

3 Sha F, Pereira F. Shallow Parsing with Conditional Random Fields[C]//Proceedings of HLT-NAACL. 2003.

4 Callum A, Freitag D, Pereira F. Maximum Entropy Markov Models for Information Extraction and Segmentation[C]//Proc. of ICML'00. 2000.

5 Yamron J, Carp I, Gillick L, et al. A Hidden Markov Model Approach to Text Segmentation and Event Tracking[C]//Proceedings of the IEEE ICASSP. 1998.

(上接第 15 页)

本文用MIT中的人脸数据库和一些多人脸图像对本方法的检测性能进行了测试。当眼睛完全睁开的情况下，其人脸检测成功率见表 3，与其它结果[7]的比较如图 5。

3.3 预警处理

采集的视频信号，通过处理加工后得到的数字图像，经人脸定位与检测，确实人脸位置，提取其描述特征值后，在本机图像特征数据库中进行搜索，若有满足设定条件值的对象，即进行报警处理，若有满足进一步搜索的条件，则到中心网络服务器上搜索与记录。同时，结合人工监视的反馈信息，来进行特征数据记录的更新与增加，并将更新部分反馈给各监视点的数据库。当有符合报警的情形时，当即进行报警处理，其方式有信息发布、安全警报、数据记录等。

4 结论

本系统是一种基于图像内容检索与识别的监控系统，它具有通过对图像定位、人脸检测和特征提取、分析后再进行安全决策的功能，在提取对象的特征值后，搜索本机或网络服务器的特征数据库来实现监控分析与报警处理。具有自动分析、记录和更新等新特点，还可以结合网络进行信息发布、数据更新等功能，操作方式更着重对象特征的处理上，增强了监控系统的辨别与识别能力。

参考文献

1 余洪山, 王耀南. 一种新型智能图像监控系统[J]. 信息与控制, 2002, 31(6): 529-533.

2 杨育彬, 李 宁, 陈世福, 等. 一种基于 Bayesian 学习的彩色肺癌图像语义描述模型[J]. 计算机研究与发展, 2002, 39(12): 1618-1624.

3 石跃祥. 基于图像内容的链码检索方法[J]. 计算机应用研究, 2001, 18(10): 49-50.

4 田 辉, 王伟明, 田 辉. 基于 DirectShow 实现视频图像捕捉的方法[J]. 微计算机信息, 2002, 18(11): 74-78.

5 柳 伟, 李国辉, 曹莉华. 一种基于内容的图像检索方法的实现[J]. 中国图像图形学报, 1998, 3(4): 304-308.

6 石跃祥, 蔡自兴, B.Benhabib, 等. 基于眼睛梯度对特征的人脸检测方法[J]. 计算机工程与应用, 2005, 41(26): 27-29.

7 Lu Xiaoguang, Wang Yunhong, Jain A K. Combining Classifiersfor Face Recognition[Z]. 2005-02-26. http//: Biometrics.cse.msu.edu/Lu_ICME03.pdf.