第34卷 第2期 北京工业大学学报 Vol.34 No.2
2008年 2月 JOURNAL OF BEIJING UNIVERSITY OF TECHNOLOGY Feb. 2008

基于半CRF模型的百科全书文本段落划分

许 勇[1], 宋 柔[2]
(1. 北京工业大学 计算机学院, 北京 100022; 2. 北京语言大学 计算机系, 北京 100083)

摘 要: 介绍了基于半条件随机域(semi-Markov conditional random fields,简称 semi-CRFs)模型的百科全书文本段落划分方法. 为了克服单纯的HMM模型和CRF模型的段落类型重复问题,以经过整理的HMM模型状态的后验分布为基本依据,使用了基于词汇语义本体知识库的段落开始特征以及针对特定段落类型的提示性特征来进一步适应目标文本的特点. 实验结果表明,该划分方法可以综合利用各种不同类型的信息,比较适合百科全书文本的段落结构,可以取得比单纯的HMM模型和CRF模型更好的性能.

关键词: 自然语言处理; 机器学习; 隐马尔科夫模型; 文本段落划分; 半条件随机域模型
中图分类号: TP 391 **文献标识码**: A **文章编号**: 0254-0037(2008)02-0204-07

文本的线性段落划分的目标是将一篇文档按照文中内容的变化划分为前后连接且不重叠的几个段落. 段落划分以文本单元为单位,文本单元一般选择句子、自然段落或者对齐到句子边界的定长文本块. 段落划分在信息检索、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[1].

文本段落划分大致可分为领域无关的段落划分和领域相关的段落划分2种类型. 领域无关的段落划分的代表性的方法是TextTiling算法[2],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模型(vector space model)[3]方法度量的. 领域相关的段落划分则由于限定了领域,因此比领域无关的段落划分具有更具体的划分要求,除了确定段落边界之外,一般还要求判断每个段落的类别. 例如,常见问题回答文件(frequently asked question)的段落划分中的段落有提问和回答2种类别. 在这种划分中,一般采用基于马尔科夫性(Markov property)的概率模型,将段落划分的问题转化为对文本单元的顺序标注的问题,连续相同的一段标记即为一个段落,如HMM模型方法[4]、最大熵马尔可夫模型方法等[5].

领域无关的段落划分一般只能使用段落内词汇重复(word repeat)等通用的依据,领域相关的段落划分中,除了词汇重复之外,还有段落类别之间的转移规律可作参考. 另外,一些难以由词汇分布的变化体现的提示性依据(如英语的大小写、可以识别的特定句式等)对确定断落很有帮助,在利用这些特征的问题上,最大熵马尔科夫模型、条件随机域(conditional random field, 简称CRF)等模型比传统的HMM模型更有优势.

《中国大百科全书》中同题材条目文本的段落划分属于领域相关的段落划分. 所谓同题材是指具有比较确定的内容成分,并且内容成分之间有一定顺序规律的条目文本集合. 如,分布于各个卷目的人物类条目中大都会含有人物的国籍、生卒年、生平事迹等内容成分;行政地名条目大都含有其行政辖属关系、面积人口、地理环境等内容成分. 从段落划分的角度看,内容成分即为段落类别. 为便于叙述,下文中将百科全书文本中题材相关的段落类别称为该题材的知识点. 将百科全书同题材条目文本按知识点划分并判断其类别,可以提供知识点的直接检索;另外,可以为后续的知识点内部知识项目的发现与获取提供有利的条件. 和一般的领域相关段落划分相比,百科全书同题材条目文本的段落有自己的特点,传统的隐马尔科夫模型等方法不太适合于同题材条目文本的段落结构. 本文采用semi-CRF模型为框架,以面向领域的词汇语义本体知识库为基础,综合利用各种不同的划分依据处理百科全书同题材条目文本的段落划分问题.

收稿日期: 2006-11-10.
基金项目: 国家自然科学基金资助项目(60272055); 国家"八六三"计划资助项目(2001AA114111).
作者简介: 许 勇(1975-), 男, 吉林延吉人, 博士研究生.

以下介绍《中国大百科全书》的《中国地理》卷目中市县一级行政地名题材上条目文本的段落划分实验情况.

1　市县行政地名文本的知识点

《中国大百科全书》中国地理卷中的市、县一级的行政地名共有 718 个条目. 这一题材的条目文本大致含有 14 个知识点,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、自然环境、农业、工业、城区概况、交通、文教、旅游、辖域情况. 其中下辖区县只包含辖区或辖县的名称和数目,辖域情况则会挑选 1～2 个重要的区县加以具体介绍,并不是每个条目文本都包含全部 14 个知识点,一般 1 个条目文本包含 7～8 个知识点.

《中国大百科全书》的文本通常不很长,许多条目只有 1 个自然段,因此本文的研究中以句子作为文本单元. 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.

工业:县内乡镇企业占全县工农业总产值 62.3%,居全省首位.

2　词汇语义知识体系

百科全书条目文本中包含大量的和一定概念相关的术语和专名. 在《中国大百科全书》的《中国地理》卷目中市县一级行政地名条目文本中有"油菜"、"花生"等各种农产品名称;"化工"、"机械"等工业部门名称;"黄河"、"长白山"等自然地理专名;"陇海铁路"、"京广线"等交通线专名等等. 如果这些词汇以词形本身进入统计,则因出现次数过低,统计数据过于稀疏,无法体现其知识点相关性. 针对这个问题,通常采用已有的词汇语义资源,如《同义词词林》. 但是经过调查,这类资源由于词汇陈旧、收词不全、兼类严重等问题不适合于百科全书段落划分,因此从百科全书段落划分、知识提取的角度出发,建立了一个专用的百科全书地理词语语义知识库. 这个库包含 2.2 万词语,其中约 3 000 个领域特征词语标注了 120 个语义属性. 大部分语义属性面向的是词汇的语义所隶属的概念,另有小部分是一些同义词的归并,如"降水量"、"降雨量"统一标为[降水量]语义属性. 标注语义属性之后,以这些属性代替原有词形计入统计. 语义属性和相应词汇举例如下(加有[]的是语义属性记):

[农产品名称]:烟叶,鳗鱼,杨梅…

[工业产品名称]:车胎,合成氨,沥青…

特征词汇具有知识点的领域属性. 也就是说一部分特征词汇是体现知识点内容的内容词,并且有抽象能力的高低之分. 作者在标注语义属性的基础上区分了抽象程度级别分别为 1 级和 2 级,1 级高于 2 级. 抽象程度的区分有助于揭示知识点的段落特性. 下面是部分举例:

农业1　级:农业,农作物,…

　　2　级:[农产品名称],耕地,林场,…

工业1　级:工业,重工业,企业,工厂,…

　　2　级:[工业产品名称],…

3　基于 HMM 和 CRF 的基本划分方法

通过观察可以发现,市县行政地名条目文本的段落结构大体上也符合一般的段落划分的依据,即词汇频次的变化和段落类型之间的转移. 为便于叙述,将这 2 个基本依据分别简记为 WR 和 ST. 基本的划分方法采用了单纯的 HMM 模型和 CRF 模型,这 2 个模型都可以直接使用 WR 和 ST 2 种依据. CRF 模型是标注序列数据的基于马尔科夫性的概率模型[6],同时也是条件型概率模型,因此可以像最大熵模型一样利用重叠的提示性特征,近年来比较受到关注,在自然语言处理、信息抽取(information extraction)等领

域都有应用[7]. CRF 模型定义了在序列数据 $X=\langle x_1, x_2, \cdots, x_L\rangle$ 上的状态变量序列 $Y=\langle y_1, y_2, \cdots, y_L\rangle$ 的条件概率 $P(Y|X)$, $P(Y|X)$是以一组定义在 X,位置 n,及 y_n、y_{n-1}上的特征函数 $f(Y, X, n)$计算的. 随机选择了 143 个标注了段落的条目文本为训练集,将剩余的 575 个条目文本为测试集,进行了 HMM 模型和 CRF 模型的基本段落划分实验.

表 1 训练集与测试集

Table 1 Train data and test data

类型	文件	段落	句子
训练集	143	1 183	2 645
测试集	575	4 817	10 563

HMM 模型的状态对应于 14 个知识点类别,参数是在训练集上估计的初始状态分布和状态转移概率和每个状态上的词汇分布,参数平滑采用了简单的加一法(add one). CRF 模型的特征函数选为提示函数形式的开始状态特征函数组 f^0、转移特征函数组 f^1、状态词汇特征函数组 f^2

$$f^0_{(i)}(X, Y, n)=[\![y_n=i]\!][\![n=1]\!]$$

$$f^1_{(i,j)}(X, Y, n)=[\![y_{n-1}=i]\!][\![y_n=j]\!]$$

$$f^2_{(w,i)}(X, Y, n)=[\![w\in\{x_n\}]\!][\![y_n=i]\!]$$

式中$[\![c]\!]$表示提示函数,即当 c 为真时$[\![c]\!]=1$,其他情况为 0;$\{x_n\}$为 n 处句子的词汇集合;w 为词汇. 实验结果的评测指标有 2 种,一种是具体到每个知识点的性能指标,另一种是整体性能指标. 某知识点 i 的性能指标用召回率和准确率衡量

$$\text{召回率}(i)=\frac{\text{测试集中知识点 } i \text{ 的划分正确的句子数}}{\text{测试集中知识点 } i \text{ 的句子数}}$$

$$\text{准确率}(i)=\frac{\text{测试集中知识点 } i \text{ 的划分正确的句子数}}{\text{测试集中被算法划分为知识点 } i \text{ 的句子数}}$$

整体性能指标为划分正确率

$$\text{划分正确率}=\frac{\text{实验集中划分正确的句子数}}{\text{实验集中句子总数}}$$

HMM 模型与 CRF 模型的划分正确率分别为 90.09% 和 89.72%. 从实验结果中分析出 3 个方面的问题. 第 1 是基本的划分依据 WR 和 ST 在总体上是有效的,可以作为基础依据,但对有些知识点不起作用,例如辖域情况. HMM 模型在这个知识点上的正确率和召回率分别为 21.95% 和 32.35%. CRF 模型在该知识点上的正确率和召回率分别为 6.33% 和 23.23%. 第 2 个问题是在 HMM 和 CRF 模型中使用 WR 和 ST 依据的方式. 实验中 2 个模型都是以各自的方式"直接"使用了 WR 和 ST 依据,这样导致的问题是 2 个模型的划分结果中都有不少重复知识点. 重复知识点是划分结果中一个知识点出现 2 次以上,HMM 模型和 CRF 模型在全部测试集上的重复次数分别为 201、167. 知识点不重复是百科全书同题材文本的段落划分特有的要求. 百科全书的编写比较精炼,关于一个条目,介绍完一方面的内容之后不会再去介绍已经介绍过的内容类型. 实验发现,如果仅将路径的搜索限制在无重复路径的范围,则会带来一定程度的性能下降. 这说明在 HMM 模型和 CRF 模型中直接使用 WR 和 ST 依据的方式不适于百科全书文本的段落特点. 第 3 个问题是未能使用体现段落特性的划分依据. 知识点作为段落,从开始、持续到结束是有一定内部结构的. 划分方法中缺少对这一特性的支持,就难以真正适合百科全书文本的特点.

4 Semi-Markov CRF 段落划分方法

semi-markov CRF(简称 semi-CRF)[8]是 CRF 模型的扩展. CRF 模型的特征函数是定义在序列数据 X 的单个位置 n 处,而 semi-CRF 的特征函数是定义在序列数据 X 的划分 s 中的某个段落上,形如 $g^k(m, X, s)=g'^k(y_m, y_{m-1}, X, t_m, u_m)$,其中 m 是段落序号,t_m、u_m 分别是第 m 个段落的开始和结束位置. Semi-CRF 的表达能力稍强于 CRF 模型,可以使用具有段落特性的特征,更适合于面向段落的处理. 作者选择了经过整理的 HMM 模型状态的后验分布、段落转移、开始状态特征以及一些提示性特征为 semi-CRF 的特征函数组,进行了段落划分实验. 这其中有些特征适用于 CRF 和 semi-CRF 模型,有些只适用于

semi-CRF 模型.

4.1 HMM 模型的状态后验分布

为了克服在模型中直接使用 WR 和 ST 依据带来的重复知识点问题，使用了经过整理的 HMM 模型的状态后验分布作为依据. 对于给定参数为 μ 的 HMM 模型和 1 个序列数据实例 $X=\langle x_1,\cdots,x_L\rangle$，在位置 $n(1\leqslant n\leqslant L)$处于状态 i 的概率为

$$\gamma_i(n)=P(y_n=i\mid X,n,\mu)=\frac{\alpha_i(n)\beta_i(n)}{\sum_{q\in Q}\alpha_q(n)\beta_q(n)}$$

式中 α、β 分别为 HMM 模型中前向算法与后向算法中的前向、后向变量；Q 为与知识点对应的模型状态集合. 该式定义了对序列数据实例 X，n 位置处(在文本段落划分的情况下，为第 n 个句子)状态的后验分布. 由此分布可计算 X 上任意划分 $s=\langle s_1,\cdots,s_p\rangle$ 中任意段落 $s_m=\langle t_m,u_m\rangle$ 上的“平均”状态分布：设 r 为均匀分布的随机变量，取值为 t_m 到 u_m 的位置，则 s_m 上的状态分布为

$$\gamma'_i(m,X,s)=P(y_m=i\mid s_m)=\sum_{\gamma=t_m}^{u_m}\gamma_i(r)P(r)$$

也可以计算 s_m 中每个位置相对于状态的条件分布

$$\delta_i(n,m,X,s)=P(r=n\mid y_m=i)=\frac{\gamma_i(n)P(r=n)}{\gamma'_i(m,X,s)}$$

如果 s_m 是整个 X 构成的段落，即 $s_m=\langle 1,L\rangle$，则相应的 $\delta_i(n,m,X,s)$为在整个 X 范围内的位置相对于状态 i 的分布，将此分布记为 $\delta'_i(n,X)$，$i\in Q$，$1\leqslant n\leqslant L$.

从状态转移上看，重复知识点就是离开某个状态之后经过若干其他状态又回到了该状态. 若要满足不重复的要求，在模型的参数上应该取消重复“回路”的概率. 状态 i 的 $\delta'_i(n,X)$分布的期望值 $E[\delta'_i(n,X)]=\sum_{n=1,2,\cdots L}\delta'_i(n,X)n$，大体说明这个状态在数据实例 X 上所处的平均位置，依该值由小到大可以确定 X 上状态之间的“前后”顺序. 根据这个序在每个数据实例上重新调整了参数，以避免状态的重复. 调整的方法是：如果最大概率状态路径确定的划分中不包含重复状态，则参数保持不变，否则对每个出现在最大概率路径中的状态，将转移到位于其“前”的状态的转移概率设置为 0，再行归一化. 根据调整后的参数计算的状态分布 $\gamma'_i(m,X,s)$，$i\in Q$ 可以用作 semi-CRF 的特征

$$g^1_{(i)}(m,X,s)=\gamma'_i(m,X,s)[\![\,y_m=i\,]\!]$$

这样，14 个后验状态分布特征函数代替了原先大量的词汇-状态对形式的特征函数.

4.2 辖域情况知识点的提示性特征

辖域情况知识点在词汇分布上和面积人口、农业、工业、交通、文教、旅游等知识点重叠得比较严重，而且这个知识点的出现次数不多，因此很难仅仅以词汇分布的差异和知识点的转移作为判断依据. 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，将要说明的行政地名连同行政级别(低于本条目地名的级别)以一定模式出现在句首，如宁安镇位于…、市属叶县…，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域情况的转入模式和转出模式. 在 1 个条目文本上，可以根据转入和转出模式的顺序第 1 次出现的位置确定辖域情况的特征窗口，设特征窗口的开始和结束位置分别为 sw_b、sw_e，针对辖域情况类型的窗口特征函数为

$$g^2(m,X,s)=[\![\,sw\_b=t_m\wedge sw\_e=u_m\,]\!][\![\,y_m=\text{辖域情况}\,]\!]$$

4.3 基于内容词抽象等级的段落开始特征

自然环境、农业等知识点的开始句子大都包含有相应内容词，但是对相应知识点抽象程度高的词汇更容易出现在开始句，在整体文本内的分布规律对知识点的开始位置有更高的提示性作用. 根据这个特点，

使用了2种针对自然环境、农业、工业、城区概况、交通、文教、旅游这7个知识点的段落开始特征. 设 V_{cw} 为上述7个知识点的集合.

第1种段落开始特征根据当前段落开始句子中的内容词界定知识点的范围. 设 $s_m=\langle t_m, u_m\rangle$ 为当前段落, $CW(t_m)$ 为 t_m 处句子中出现的内容词集合, $c(w)$ 为内容词 w 的抽象度等级, $i(w)$ 为其对应的知识点. V_{cw} 中的知识点 i 的内容词集合为 $IW(i)$, i 在 t_m 处的等级 $R(t_m,i)=\min\{c(w)\mid w\in CW(t_m)\cap IW(i)\}$, 第1种段落开始特征函数为

$$g^3_{(r)}(m,X,s)=[\![\, y_m\in V_{cw}\wedge R(t_m,y_m)=r \,]\!],\ r\in\{1,2\}$$

这种特征有2个特征函数,分别界定级别为1的知识点范围与级别为2的知识点范围.

第2种段落开始特征力求更"精确"一些. 设 $L(t_m)=\{i(w)\mid w\in CW(t_m)\}$, $L'(t_m)=L(t_m)-L(t_m-1)$, 符号含义同上. 对 $L'(t_m)$ 中的元素进行筛选的步骤为:

1) 若 $[L(t_m-2)\cup L(t_m-1)]\cap[L(t_m)\cup L(t_m+1)]\neq\phi$, 置 $L'(t_m)=L'(t_m)\cap L(t_m+1)$

2) 置 $L'(t_m)=\{i\mid i\in L'(t_m)\wedge t_m<E(\delta'_i)\}$

上述步骤中第1步主要是在前后1个句子的范围内通过比较获得更可靠的段落开始特征;第2步是使知识点的开始位置符合其 δ' 分布,因为大于 δ'_i 分布的期望值的位置不大可能是知识点 i 的开始位置. 经过筛选之后,第2种特征函数为:

$$g^4_{(i,j,r)}(m,X,s)=[\![\, y_m=i\wedge i\in V_{cw}]\!][\![j\in L'(t_m)\wedge R(t_m,j)=r \,]\!]$$

第2种特征函数共有 $7\times2\times7=98$ 个.

5 Semi-CRF 段落划分实验结果

Semi-CRF 模型中采用的特征有开始状态特征、状态转移特征、经过整理的 HMM 状态的后验分布、针对辖域情况的特征、内容词段落开始特征. 这几个特征函数可以以表2中的几个方面来说明其性质. 为了下文的表述方便,表2中亦给出了每个特征的符号命名.

表2 特征的符号命名以及适用模型

Table 2 Symbols for the features and suitable models

依据	基础依据	适用模型	符号命名
开始状态特征	ST	HMM, CRF, semi-CRF	ST
状态转移特征	ST	HMM, CRF, semi-CRF	ST
整理过的 HMM 后验状态分布	ST + WR	HMM, CRF, semi-CRF	HMM-WR/ST
辖域情况特征	提示性依据	CRF, semi-CRF	DI-SUB
内容词段落开始特征	提示性依据	Semi-CRF	DI-CB

经过整理的 HMM 状态的后验分布实际上是基本依据 ST 和 WR 结合在一起的一种表达方式. 与此相对比,在第3节中 HMM 和 CRF 模型的基本划分实验中直接使用 WR 的方式记为 Naïve-WR. 针对辖域情况的特征既适用于在 CRF 模型,也适用于 semi-CRF 模型. 基于内容词的段落开始特征则相反,难以在 CRF 模型中使用. 表3是不同的依据和使用方式的组合在 HMM、CRF、semi-CRF 3种模型中的划分正确率. 其中, HMM 和 ST + HMM-WR/ST 的组合是使用调整过的参数来计算最大概率路径所得划分的结果. 通过表3的结果可以看出, HMM-WR/ST、DI-SUB、DI-CB 3种特征都可以获得性能的提高, HMM-WR/ST 是通过避免知识点重复, DI-SUB、DI-CB 2个特征则体现了提示性依据的有效性. 但由于模型限制, HMM 不能使用 DI-SUB 和 DI-CB. CRF 和 ST + HMM-WR/ST + DI-SUB + DI-CB 的组合中使用的是"模拟"的 DI-CB, 即在每个句子上都产生 DI-CB 特征, 而不是在段落的开始句子. CRF 和 ST + HMM-WR/ST + DI-SUB + DI-CB 组合的结果表明, DI-CB 特征不适用于 CRF 模型, 也说明这个特征体现的是知识点的段落特性.

表 3　依据的不同组合在不同模型中的划分正确率

Table 3　Performances of clue combinations on each model

特征组合	HMM	CRF	Semi CRF
ST + Naïve-WR	90.0	89.7	
ST + Naïve-WR + DI-SUB		90.9	
ST + HMM-WR/ST	91.2	91.0	91.4
ST + HMM-WR/ST + DI-SUB		91.8	92.3
ST + HMM-WR/ST + DI-SUB + DI-CB		92.2	93.6

表 4 是 semi-CRF 和 ST + HMM-WR/ST + DI-SUB + DI-CB 组合的各个知识点的正确率和召回率，总的划分正确率为 93.6%.

表 4　各个知识点的正确率和召回率

Table 4　The precision and recall of each topic

知识点	段落数	句子数	正确率/%	召回率/%	知识点	段落数	句子数	正确率/%	召回率/%
概述	571	703	99.28	99.43	农业	388	1 110	90.54	90.54
地处方位	543	600	96.70	97.83	工业	475	1 380	91.88	89.92
面积人口	561	612	99.17	97.87	城区概况	95	244	78.48	76.29
行政中心	213	218	97.63	95.87	交通	338	775	87.45	92.32
下辖区县	134	134	97.08	99.25	文教	79	110	89.65	70.91
历史沿革	535	2 140	98.11	97.19	旅游	401	1 063	91.94	95.21
自然环境	431	1 332	91.82	95.34	辖域情况	53	142	93.65	83.09

6　结束语

介绍了基于 semi-CRF 模型的百科全书同题材条目文本的段落划分方法，实验中的具体题材是中国地理卷目中市县一级行政地名条目. 为了有效处理文本中的大量专名和术语带来的数据稀疏问题，建立了百科全书地理词语语义知识库，标注了大部分专名和术语的语义属性. 在此基础上，使用了 HMM 模型状态的后验分布特征. 这个分布是经过整理过的，目的是为了克服单纯 HMM 模型和 CRF 模型中重复知识点问题. 实验结果表明经过整理的 HMM 模型的状态后验分布可以避免重复的知识点. 在百科全书地理词语知识库中，还区分了词汇及语义属性相对于部分知识点的抽象程度，并据此使用了提示性的段落开始特征. 实验中使用的另一个提示性特征是针对辖域情况知识点的窗口特征. 这些提示性依据的使用都收到了效果，提高了整体划分性能. 实验表明，上述方法较适合于百科全书的同题材条目文本的段落划分.

目前 HMM 模型状态的后验分布的整理方法是经验性的，其有效性与数据集合的特点和段落划分标准都有关系. 今后工作中需要就这一方面在更多的题材范围内进行更为深入的研究，寻找更为合理、更有理论依据的整理方法.

参考文献：

[1] REYNAR J C. Topic segmentation: algorithms and applications [D]. Philadelphia, USA: University of Pennsylvania, 1998: 130-151.

[2] MARTI A H. Multi-paragraph segmentation of expository text[C]// Proceedings of the 32nd Annual Meeting of the Associa-

tion for Computational Linguistics. Las Cruces, New Mexico: Association for Computational Linguistics, 1994: 9-16.

[3] CHRISTOPHER D M, HINRICH S. Foundations of statistical natural language processing[M]. Cambridge, Massachusetts: MIT Press, 1999: 539-544.

[4] 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. Seattle, Washington: Institute of Electrical and Electronics Engineers Signal Processing Society, 1998: 333-336.

[5] MCCOLLUM A, FREITAG D, PEREIRA F. Maximum entropy markov models for information extraction and segmentation[C]// Proceedings of ICML 2000. Stanford, California: Morgan Kaufmann Publishers Inc, 2000: 591-598.

[6] JOHN L, ANDREW M, FERNANDO P. Conditional random fields: probabilistic models for segmenting and labeling sequence data [C]// Proceedings of the International Conference on Machine Learning (ICML-2001). MA: Morgan Kaufmann Publishers Inc, 2001: 282-289.

[7] FEI S, FERNANDO P. Shallow parsing with conditional random fields [C]// Proceedings of HLT-NAACL. Edmonton, Canada: Association for Computational Linguistics, 2003: 134-141.

[8] SUNITA S, WILLIAM W C. Semi-markov conditional random fields for information extraction[C/OL]// Proceedings of the Eighteenth Annual Conference on Neural Information Processing Systems. Vancouver: MIT Press, 2004. http://citeseer.ist.psu.edu/653054.html

A Semi-Markov CRF Model Approach to Encyclopedia Text Topic Segmentation

XU Yong[1], SONG Rou[2]

(1. College of Computer Science, Beijing University of Technology, Beijing 100022, China;
2. College of Information Science, Beijing Language and Culture University, Beijing 100083, China)

Abstract: This paper introduced the semi-markov Conditional Random Fields (semi-CRFs) model based method for Chinese Encyclopedia text topic segmentation. The authors adopted HMM model state posterior as the basic segmentation clue which was adjusted to each text instance to overcome the topic duplication problem of fully connected state HMM model and CRF model. The authors also used several segment level word semantic features derived from domain thesaurus, and additional topic specific cue phrases to make the method more adapted to target domain. The experiment result showed that this method was suitable for Chinese Encyclopedia text topic structure and achieved better performance than HMM model and CRF model.

Key words: natural language processing systems; machine learning; hidden markov models; topic segmentation; semi-Markov CRF