第34卷 第22期 计 算 机 工 程 2008年11月
Vol.34 No.22 Computer Engineering November 2008
·软件技术与数据库· 文章编号：1000—3428(2008)22—0043—03 文献标识码：A 中图分类号：TP311

面向概括性小文本的文本分割算法

陈 源，陈 蓉，胡俊锋，林 霖，张靖波，于中华
(四川大学计算机学院，成都 610064)

摘 要：文本分割是自然语言文本处理的一项重要研究内容。该文针对现有模型无法有效分割概括性小文本的不足，提出基于隐马尔可夫模型的统计算法。该算法利用小文本中各结构块的长度及词汇信息，对概括性小文本进行同一主题不同论述侧面的分割。对发射概率设计了基于句群和基于分割点2种不同的计算方法。以Medline摘要为样本进行的实验表明，该算法对概括性小文本分割是有效的，明显好于经典的TextTiling算法。

关键词：文本分割；概括性小文本；隐马尔可夫模型；边界识别；相似性度量

Text Segmentation Algorithm Oriented to Small General-text

CHEN Yuan, CHEN Rong, HU Jun-feng, LIN Lin, ZHANG Jing-bo, YU Zhong-hua
(School of Computer Science, Sichuan University, Chengdu 610064)

【Abstract】 Text segmentation is an important filed in the area of natural language processing. However, there is a defect that the existing models cannot effectively segment small general-text. For the reason, an algorithm based on Hidden Markov Model(HMM) is proposed in this paper. The algorithm segments a small general-text with a single topic into its different aspects of discussion using the length distribution of every structure block and the terms. Two methods are designed for computing symbol emission probabilities of the HMM, one of them is based on sentence group while the other is based on segmentation point. Experiments on Medline abstracts show that the effect of the algorithm proposed is much better than the TextTiling algorithm.

【Key words】 text segmentation; small general-text; Hidden Markov Model(HMM); boundary recognition; similarity metric

1 概述

当前正处在一个被大量信息“淹没”的时代，如何快速、准确地获取所需要的信息，将信息变成决策所需要的知识，已经成为知识经济面临的重要课题。传统的信息检索系统，如搜索引擎等，在处理自然语言文本时，简单地将文本看成词、短语或句子的集合，没有充分利用文本的结构信息，对关键词在文本的每次出现都一视同仁，从而导致搜索的准确率不高，很难满足用户的需求。文本结构分析是进一步提高搜索引擎可用性的关键。此外，文本结构分析还对自然语言其他方面的处理，如机器翻译、自动文摘、信息抽取(Information Extraction)、文本挖掘等，具有重要意义。

人们已经对文本结构分析进行了大量的研究工作，提出了一系列有效的模型和算法。但是，已有的工作主要针对具有段落结构的长文本，分割的依据是论述焦点和主题的转换，基于的文本特征是反映不同主题的实词及其聚集。而概括性小文本，如Medline摘要，没有段落划分，整篇文本围绕一个主题从不同侧面进行概括性论述，论述的侧面比较固定(如Medline摘要一般包括背景、方法、结果、结论等)。对概括性小文本自动分割不同的论述侧面，目前尚未见相关的研究报道。为此，本文以Medline摘要为例，对概括性小文本的自动分割问题进行了研究，提出了基于隐马尔可夫模型(HMM)的分割算法，并对算法进行了充分的实验验证。

2 相关工作

2.1 相关算法介绍

文本分割(子主题分割)的主要任务是分析文档结构，自动识别具有独立意义的单元并标记出单元边界。目前，研究者已经提出了一系列的模型和算法，来解决文本的自动分割问题。依据分割所基于的信息类型，可以将现有方法归结为3大类[1]：基于词汇分布，基于特定的语言现象和基于概率统计。基于词汇分布的分割算法假设分割点两侧的单词分布不同，分割点的确定与其两侧不同单词的出现有关。如果有大量不同的实词出现在某候选分割点的两侧，则认为该候选成为真正分割点的可能性较大，反之则较小。基于词汇的分割算法大多属于段落层次上的无监督学习方法，直观、易于实现，并且不需要额外的训练数据。但是其缺点也比较突出。以Texttiling为例[2]，由于该算法工作在段落层次上，因此只有在长度较大、段落长度较均衡的文本中才能达到较好的分割效果。基于特定语言现象的分割算法的主要代表是Passoneau and Litman设计的决策树模型[3]，该模型广泛应用于语音文本流的边界分割。模型假设特定的语言现象与片断首尾隐含某种必然联系，通过建立语义知识库对段落间的相似度进行判断。语义知识库集成了短语复用、语义重复及tf-idf等多种特征以及深层的语义特征(如边界两侧的名词短语是否语义相背)。这类算法的缺点是建立语义知识库需要大量的训练语料，并需要针对不同的特殊文本进行修改和完善。

基金项目：国家自然科学基金资助项目(60473071)；高等学校博士学科点专项科研基金资助项目(20020610007)；四川大学计算机学院青年基金资助项目

作者简介：陈 源(1983－)，男，硕士研究生，主研方向：数据挖掘，自然语言处理；陈 蓉，工程师；胡俊锋、林 霖、张靖波，硕士研究生；于中华，副教授、博士

收稿日期：2008-05-25 **E-mail：**yuzhonghua@cs.scu.edu.cn

基于概率统计模型的分割算法认为，合适的概率统计模型能够为边界的确定提供可靠的依据[4-5]。这类分割方法属于有监督学习算法，需要有标注好的训练语料，在对训练语料学习的基础上，基于候选分割点的局部前后文环境，来判断候选分割点为真的可能性。

2.2 概括性小文本的特殊性

概括性小文本与一般文本不同。概括性小文本规模较小，全文通成一段，大多由 5~20 个句子组成。概括性小文本叙述的主题单一，针对主题的不同侧面(如历史，现状，发展等)组织内容，而且不同侧面的出现顺序及长度规律性强。

现有的文本分割算法无法适应概括性小文本自动分割的需要，主要存在的问题包括：(1)传统的文本分割是基于主题及主题的迁移来进行的，决策的依据是候选分割点前后表达不同主题的实词分布，而概括性小文本分割是在单一主题的短文本内进行的，目的是确定同一主题的不同论述侧面，因此，反映不同主题的实词无法作为有效的分割标志，虚词(尤其是表达转折关系的虚词)应该是主要的分割特征词；(2)由于概括性小文本篇幅较小，无法获取如 Beeferman 模型所要求的那样丰富的前后文(取候选分割点前后 K=500 个单词)；(3)概括性小文本，如 Medline 摘要，不同侧面的出现顺序及长度规律性较强，利用这些信息，可以弥补直接利用传统算法带来的实词缺乏区分能力、前后文窗口过小的困难。

本文提出了基于 HMM 的统计算法。该算法利用小文本中各结构块的长度及词汇信息(主要是虚词)，对概括性小文本进行同一主题不同论述侧面的分割。对发射概率分别设计了基于句群和基于分割点 2 种不同的计算方法。以 Medline 摘要为样本进行的实验表明，本文所提出的算法对概括性小文本分割是有效的，明显好于经典的 TextTiling 算法。

3 面向概括性小文本的分割算法

这里提出的面向概括性小文本的算法基于 HMM，并且假设文本所包含的论述侧面及其顺序固定，任务是确定这些论述侧面之间的分割点。

假设文本是一个句子序列$W=w_1,w_2,\cdots,w_n$，其中，n是文本所包含的句子数，论述侧面序列为$K=k_1, k_2,\cdots,k_c$，确定c个分割点，每个分割点的候选取值为 1~n，即句子的序号。

算法基于HMM，把$K=k_1,k_2,\cdots, k_c$看成是观测序列，假定状态集为$S=\{s_1,s_2,\cdots,s_n\}$，其中，$s_i=i$表示第i个句子的结尾为一个候选分割点。这样，确定最可能分割点的问题就归结为在给定观测序列的前提下，寻找最优状态序列$s_{i1},s_{i2},\cdots,s_{ic}$的问题。即

$$s_{i1},s_{i2},\cdots,s_{ic}=\underset{\forall i1,i2,\cdots,ic\in\{1,2,\cdots,n\}}{\operatorname{argmax}} P(s_{i1},s_{i2},\cdots,s_{ic}/k_1,k_2,\cdots,k_c)$$

上述问题属于典型的 HMM 解码问题。图 1 给出了面向概括性小文本分割的 HMM 示意图。

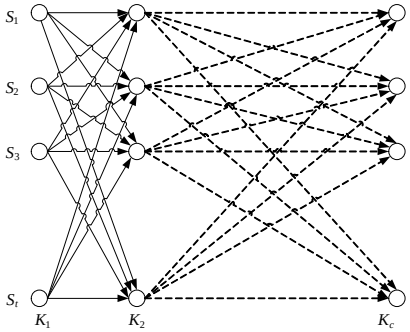

图 1 面向概括性小文本分割的 HMM

构建HMM的关键是训练确定模型的 2 个参数集($\boldsymbol{A}$, $\boldsymbol{B}$)，其中，$\boldsymbol{A}$表示状态转移概率矩阵$\{a_{ij}\}$；$\boldsymbol{B}$表示符号发射概率矩阵$\{b_{ijk}\}$。

3.1 符号发射概率计算

在HMM中，符号发射概率b_{ijk}定义为在第t个状态为s_i，t+1个状态为s_j的情况下，t时刻发射出观测符号k_m的概率，可表示为

$$P(O_t=k_m \mid X_t=s_i, X_{t+1}=s_j)=b_{ijk}$$

本文考虑了 2 种发射概率的计算方法：

第 1 种基于句群，即把任意 2 个状态s_i和s_j间的所有句子作为句群T_{ij}构成特征文本，b_{ijk}定义为T_{ij}属于k_m的概率$P(k_m|T_{ij})$。

第 2 种方法基于分割点，取t时刻的分割点s_i前后各一个句子构成句子的二元组作为特征文本，b_{ijk}定义为

$$P(O_{t-1}=k_q, O_t=k_p|T_i)$$

其中，T_i为潜在分割点s_i前后各一个句子构成的特征文本。

上述 2 种方法都基于向量空间模型来表达特征文本，向量的每一维表示一个特征词的权重，特征词选择的依据是各个单词在训练语料中的信息熵大小，信息熵值最小的前 N 个词被选择来作为特征词，这样，每个特征文本都被表示成 N 维向量的形式。

对于任意的特征词w_i和论述侧面k_j，w_i对k_j的权重WG_{ij}定义为

$$WG_{ij}=round\left(10\times\frac{1+\mathrm{lb}(tf_{ij})}{1+\mathrm{lb}(l_c)}\right)$$

其中，tf_{ij}为单词w_i在论述侧面k_j的特征文本中出现的次数；l_c为特征词的总数。

3.1.1 基于句群的发射概率计算

为了计算$P(k_m|T_{ij})$，首先对T_{ij}向量化，然后利用KNN算法的思想确定T_{ij}属于k_m的概率。算法流程如下：

(1)在训练语料中寻找最相似的前 K 个训练样本，构成集合 $\boldsymbol{A}$。相似度的计算采用向量夹角余弦的形式，即假设 $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 为任意 2 个句群向量，则它们之间的相似度为

$$\cos(\boldsymbol{x},\boldsymbol{y})=\frac{\boldsymbol{x}\cdot\boldsymbol{y}}{|\boldsymbol{x}||\boldsymbol{y}|}=\frac{\sum_{i=1}^{n}x_i y_i}{\sqrt{\sum_{i=1}^{n}x_i^2}\sqrt{\sum_{i=1}^{n}y_i^2}}$$

(2)假设$n_1, n_2,\cdots,n_c$为集合$\boldsymbol{A}$中分别属于c个论述侧面的不同元素的个数。据此条件概率$P(k_m|T_{ij})$计算如下：

$$P(k_m \mid T_{ij})=\frac{n_m}{\sum_{p=1}^{c}n_p}$$

3.1.2 基于分割点的发射概率计算

计算 $P(O_{t-1}=k_q,O_t=k_p|T_i)$ 时，首先对T_i向量化，然后确定T_i属于论述侧面k_q和k_p之间分割点特征文本的概率。算法流程如下：

(1)在训练语料中寻找最相似的前 K 个训练样本，构成集合 $\boldsymbol{A}$。相似度的计算仍然采用向量夹角余弦的形式。

(2)假设$n_1,n_2,\cdots,n_{c\times c}$为集合$\boldsymbol{A}$中分别属于$c\times c$个论述侧面序偶$<k_u, k_v>$(表示这 2 个论述侧面$k_u$和$k_v$之间的分割点)的不同元素的个数。据此条件概率 $P(O_{t-1}=k_q, O_t=k_p|T_i)$ 计算如下：

$$P(O_{t-1}=k_q, O_t=k_p|T_i)=\frac{n_t}{\sum_{r}^{c\times c}n_r}$$

其中，n_t为$\boldsymbol{A}$中属于序偶$<k_q, k_p>$的元素的个数。

3.2 状态转移概率计算

状态转移概率即文本中 2 个候选分割点之间构成一个完整论述侧面的概率。由于每篇文档长度不同，因此需要进行归一化处理。考虑到小文本的特点，文档长度不超过 20 句，归一化公式把所有文本长度都规范为 10，即每个文档都有 10 个状态，每个真实状态w_i都归一化为对应的 $\hat{w}_i$ ：

$$\hat{w}_i = round\left(\frac{w_{\max} - w_i}{w_{\max} - w_{\min}} \times 10\right)$$

其中，w_i为分割点标号；$\hat{w}_i$ 为分割点归一化后的标号；$w_{\max}$和$w_{\min}$分别为最大标号和最小标号。

状态转移概率基于最大似然估计加一法得到：

$$P(X_i, X_j) = \frac{C(L(X_i, X_j)) + 1}{N + B}$$

其中，X_i和X_j是任意的 2 个状态(候选分割点)；$L(X_i, X_j)$为X_i和X_j之间文本的长度，以句子个数为单位；$C(\bullet)$为训练语料中相应长度的文本作为单独论述侧面的次数；N为训练语料中论述侧面的总个数；B是论述侧面的不同长度数。

3.3 确定最佳分割点序列

确定最佳分割点序列采用经典的HMM解码算法——Viterbi算法[6]。首先定义

$$\delta_j(t) = \max_{X_1, X_2, \cdots, X_{t-1}} P(X_1, X_2, \cdots, X_{t-1}, o_t, \cdots, o_{t-1}, X_t = j \mid \mu)$$

Viterbi 算法基于动态规划的思想，采用如下方法计算最佳的状态序列：

(1)初始化

$$\delta_j(1) = \pi_j, 1 \le j \le N$$

(2)推导

$$\delta_j(t+1) = \max_{1 \le i \le N} \delta_i(t) a_{ij} b_{ijo_t}, 1 \le j \le N$$

$$\psi_j(t+1) = \arg\max_{1 \le i \le N} \delta_i(t) a_{ij} b_{ijo_t}, 1 \le j \le N$$

(3)回溯最可能的状态序列

$$\hat{X}_{T+1} = \arg\max_{1 \le i \le N} \delta_i(T+1), \quad \hat{X}_t = \psi_{\hat{X}_{t+1}}(t+1)$$

4 实验

4.1 实验语料

目前还没有针对概括性小文本的文本分割评测语料。考虑到 Medline 摘要属于典型的概括性小文本，并且其中包含大量在写法上已经自然分割好的摘要。因此，本文从美国国立生物技术信息中心(www.ncbi.nlm.nih.gov)网站上下载了 678 篇这种自然分割好的摘要作为训练语料，另外下载 87 篇这种摘要，去掉自然分割标记(BACKGROUND 等)作为测试语料。实验中分割的论述侧面序列为 BACKGROUND, METHOD, RESULTS, CONCLUSION。

表 1 给出了自动抽取出的部分特征词及其信息熵。

表 1　部分特征词及信息熵

特征词	信息熵	特征词	信息熵
Been	0.288 583	including	0.691 544
Primary	0.549 019	management	0.462 135
Infection	0.638 995	provide	0.493 551
That	0.693 855	experience	0.598 290
However	0.712 175	malignant	0.585 954

从表 1 中可以看到，对概括性小文本分割起重要作用的往往是虚词，如 been 等。这是由于分割在单一主题的小文本内进行，分割的目标是确定同一主题的不同论述侧面，此时反映不同主题的实词已经失去区分作用。进一步的分析发现，在 Medline 摘要的各论述侧面中，语法有着明显的区别，在介绍 METHOD 和 RESULTS 时，经常使用被动态，而其他论述侧面则大多使用主动态。

4.2 实验结果

为了对算法的性能进行深入的评价，使用相同的训练和测试语料，对算法在不同情况下的输出结果进行了统计分析，并与 TextTiling 方法进行了对比。表 2 给出了算法输出的部分实例。表 3 为实验结果。

表 2　部分分割结果实例

编号	基于句群	基于分割点	正确划分
21	{0,1,4,8}	{0,1,3,6}	{0,1,3,6}
22	{0,4,8,11}	{0,4,7,12}	{0,3,6,12}
23	{0,2,10,12}	{0,3,8,12}	{0,1,4,11}
24	{0,1,3,5}	{0,1,3,6}	{0,1,3,6}
25	{0,1,2,5}	{0,1,3,5}	{0,1,2,5}
26	{0,4,8,13}	{0,3,7,11}	{0,1,7,11}

表 3　实验结果　(%)

算法	分割点 1 准确率	分割点 2 准确率	分割点 3 准确率
TextTiling 方法	24	20	30
未考虑论述侧面长度	40	30	39
基于句群	44	42	49
基于分割点	56	43	70

从表 3 的实验结果可以看出，当未考虑论述侧面的长度时，算法的准确率虽然高于 TextTiling，但明显低于考虑长度的情况，说明长度对分割点的确定具有重要作用。同时，基于分割点计算发射概率的方法优于基于句群的方法，准确率提高了 8%左右，这说明分割点前后的局部前后文对分割起主要作用，远离分割点的文本片断作用较弱，对分割点的确定产生干扰。

5 结束语

本文针对现有模型无法有效分割概括性小文本的不足，提出了基于 HMM 的统计算法，该算法利用小文本中各结构块的长度及词汇信息，对概括性小文本进行同一主题不同论述侧面的分割。对发射概率分别设计了基于句群和基于分割点 2 种不同的计算方法。以 Medline 摘要为样本进行的实验表明，本文所提出的算法对概括性小文本分割是有效的，明显好于经典的 TextTiling 算法。提出的算法假设文本所包含的论述侧面及其顺序固定，任务是确定这些论述侧面之间的分割点。作为下一步的工作，除了进一步提高算法分割概括性小文本的准确率外，还要对算法进行扩展，使之可以处理论述侧面不固定的情况。

参考文献

[1] Beeferman D, Berger A, Lafferty J. Statistical Models for Text Segmentation[J]. Machine Leraning, 1999, 34(1-3): 177-210.

[2] Hearst M A. Text Tiling: Segmenting Text into Multi-paragraph Subtopic Passages[J]. Computational Linguistics, 1997, 23(1): 33-64.

[3] Passoneau R J, Litman D J. Discourse Segmentation by Human and Automated Means[J]. Computational Linguistics, 1997, 23(1): 103-139.

[4] 石　晶, 戴国忠. 基于PLSA模型的文本分割[J]. 计算机研究与发展, 2007, 44(2): 242-248.

[5] 朱靖波, 叶　娜, 罗海涛. 基于多元判别分析的文本分割模型[J]. Journal of Software, 2007, 18(3): 555-564.

[6] Fragkou P, Petridis V, Kehagias A. A Dynamic Programming Algorithm for the Segmentation of Greek Texts[C]//Proceedings of the CONSOLE XII Conference. [S. 1.]: IEEE Press, 2003.