ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.18, No.3, March 2007, pp.555−564
DOI: 10.1360/jos180555

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

基于多元判别分析的文本分割模型*

朱靖波+, 叶 娜, 罗海涛

(东北大学 计算机软件研究所,辽宁 沈阳 110004)

Text Segmentation Model Based on Multiple Discriminant Analysis

ZHU Jing-Bo+, YE Na, LUO Hai-Tao

(Institute of Computer Software and Theory, Northeastern University, Shenyang 110004, China)

+ Corresponding author: Phn: +86-24-83672481, E-mail: zhujingbo@mail.neu.edu.cn, http://www.nlplab.com

Zhu JB, Ye N, Luo HT. Text segmentation model based on multiple discriminant analysis. *Journal of Software*, 2007,18(3):555−564. http://www.jos.org.cn/1000-9825/18/555.htm

Abstract: This paper proposes a new domain-independent statistical model. In this model, four multiple discriminant analysis (MDA) criterion functions are defined and used to achieve global optimization in finding the best segmentation by means of the smallest within-segment distance, the largest between-segment distance and segment length. To alleviate the high computational complexity problem introduced by the new model, genetic algorithms (GAs) are used. Comparative experimental results show that the methods based on MDA criterion functions have achieved higher P_μ than that of TextTiling and Dotplotting algorithms.

Key words: text segmentation; multiple discriminant analysis; within-segment distance; between-segment distance; segment unit length; genetic algorithm

摘 要: 提出了一种独立于具体领域的文本线性分割统计模型,其中采用多元判别分析方法定义了 4 种全局评价函数,实现对文本分割模式的全局评价,寻找满足分割单元内距离最小化和分割单元间距离最大化条件的最好分割方式.该模型采用遗传算法来解决新模型的高计算复杂度问题.比较性实验结果显示,新模型比 TextTiling 和 Dotplotting 算法取得了更高的 P_μ评价性能.

关键词: 文本分割;多元判别分析;分割单元内距离;分割单元间距离;分割单元长度;遗传算法
中图法分类号: TP391 文献标识码: A

文本智能处理系统中,如信息检索一般是以整篇文档为基本处理单位,隐性假设一篇文档主要讨论一个主题.实际上,一篇文档往往涉及到一个或多个子主题,因而,基于整篇文档的处理颗粒度在很多实际应用中难以满足用户更高和更准确的要求.很明显,从用户的角度来看,基于段落的检索技术能够比全文检索技术提供更加准确、更少冗余信息的答案[1].另外,如果能够自动识别和划分文本的子主题结构,将语义段落作为文本处理单

* Supported by the National Natural Science Foundation of China under Grant No.60473140 (国家自然科学基金); the National High-Tech Research and Development Plan of China under Grant No.2006AA01Z154 (国家高技术研究发展计划(863)); the Program for New Century Excellent Talents in University under Grant No.NCET-05-0287 (新世纪优秀人才支持计划); the National 985 Project of China under Grant No.985-2-DB-C03 (国家 985 工程)

Received 2005-11-29; Accepted 2006-01-24

位,将有益于改善问答系统和文档摘要技术的性能.

文本分割是指将文本中属于同一个子主题的相邻段落合并为一个语义段落*,简称为一个分割单元(segment unit,简称 segment),这样可以把一篇文本线性分割成为若干个语义段落,形成语义段落序列.但是在不同的应用中,分割点的位置可能有所不同,主要有 3 种方式:词与词之间、句子与句子之间以及段落与段落之间.例如,在语音识别数据中,由于缺乏句子和段落标记,则采用词与词之间的分割点方式;在文本数据流中(如 TDT 任务),由于缺乏段落标记,可以采用句子与句子之间的分割点方式;如果对一篇自然文本进行分割,可以采用句子与句子之间,或者段落与段落之间的分割点方式.实际上,针对不同的分割点方式,适用的文本分割技术本质上没有太大区别,主要不同点在于候选分割点位置的预测和选择.本文主要研究基于段落层次的文本线性分割技术,采用段落与段落之间的分割点方式.

研究人员提出了一些文本分割技术,其中利用了一些语言学特征信息[2−5],包括线索短语特征(cue phrases)、新词出现(new words occurrence)、重现特性(包括词汇重现、*n* 元词语重现、*n* 元字符重现)、专名和代名词使用(named entities and pronoun usage)、同义词重复(synonymy)、语速停顿等.还有采用统计方法[6−17],包括词共现技术、词汇链、文本片段的相似性计算技术、动态规划算法、聚类算法、HMM 模型等.

在文本分割过程中,需要解决两个关键问题:主题边界的自动识别和分割单元(语义段落)数目的确定.一些研究方法主要基于文本中相邻片断(可能是段落或者文本片断)的相似性,然后根据相似性的变化程度解决主题边界的识别[2]:若相似度大,则两者合并为一个分割单元;否则识别为主题边界.实际上,在识别主题边界的过程中,预先确定相似度阈值是非常困难的,这就造成难以准确识别主题边界.这种技术属于局部优化方法,实现过程比较简单.另外一些方法主要考虑文本所有片断之间的相似性,不仅仅限于相邻文本片断,属于全局优化方法.Reynar[6]和 Choi[10,11]采用 Dotplotting 技术实现了文本线性分割.严格上来说,该方法考虑了全局优化和局部优化的一些特性,介于两者之间.Yaari[13]采用了一种集聚聚类技术实现文本层次分割过程,然后通过一定规则将层次分割转换成为线性分割,属于无指导聚类.但是,该方法主要通过合并最相似的相邻两个文本片断来实现聚类过程,也无法真正克服局部优化技术的弱点.Fragkou 等人[17]采用动态规划方法(dynamic programming,简称 DP)来实现文本分割,主要考虑两个因素:分割单元内的词汇相似性和分割单元的长度分布,属于全局优化技术.但是,这些方法都需要一些训练语料来实现模型先验参数的估计[15−17].在实际应用中,预先构造一个合适的训练语料用于模型先验参数的估计也不是很容易的.

本文提出了一种基于多元判别分析(multiple discriminant analysis,简称 MDA)的文本分割技术,其中考虑了 3 个因素:分割单元内距离(within-segment distance,简称 WSD)、分割单元间距离(between-segment distance,简称 BSD)和分割单元的长度信息(segment unit length,简称 SUL),采用多元判别分析方法定义 4 种分割全局评价函数(MDA criterion function),实现对文本分割的全局评价.其中,假设分割单元内距离越小(强凝聚性)、分割单元间距离越大(强发散性)的分割模式是全局最佳的.最后根据全局分割评价结果,选择具有最高评价值的分割模式作为正确分割,从而自动判定主题边界和确定语义段落的最佳数目.本文的文本分割技术属于全局优化方法,并且无须使用任何训练语料,因此属于与具体领域无关的方法.

本文第 1 节给出文本分割的统计模型.第 2 节详细讨论基本思想,给出全局评价函数的定义.在第 3 节中给出文本分割算法.第 4 节详细给出实验设计和性能分析.最后讨论将来的主要工作.

1 统计模型

首先定义一个文本为词序列 $W=w_1w_2\ldots w_t$,其中,t 表示文本 W 包含词的个数;定义文本 W 的分割模式为 $S=s_1s_2\ldots s_c$,其中,c 表示文本分割 S 包含的分割单元个数.给定一个文本 W,文本分割统计模型的关键问题在于寻求具有最大概率的分割模式,计算方法是

* 在本文的论述中,分割单元(segment unit or segment)与语义段落(semantic paragraph)属于同一个概念,主要是区别于传统意义的文本段落.

$$\hat{S} = \arg\max_{S} P(S \mid W) \tag{1}$$

由于大多数文本中包含的句子或段落的长度差别很大,在文本分割过程中会造成不平衡比较现象[2].例如,假设模型采用句子为最小比较单元,很明显,两个较长句子的相似度评价值可能高于两个较短句子或者其中一个是短句子的相似度评价值,这种不平衡比较现象将导致分割错误偏移.同样,如果采用段落为最小比较单元,也会出现类似现象[2].Hearst 在 TextTiling 算法[2]中采用块(block)方法来解决这个问题,采用块作为最小比较单元.在 TextTiling 算法中,块定义为包含 *blocksize***个词的文本片断.采用具有相同长度的块参与分割评价过程,能够有效解决不平衡比较现象.

在本文的统计模型中也引入块方式来重新定义文本 W 为块序列 $B=b_1b_2\ldots b_k$,其中,k 表示文本 B 包含块的个数,则模型(1)可以修正为给定一个文本 B,最大概率的文本分割模式计算方法是

$$\hat{S} = \arg\max_{S} P(S \mid B) \tag{2}$$

实际上,直接利用上述统计模型进行求解具有最大概率的分割方式的任务是非常困难的.Utiyama 和 Isahara[16]引入分割单元的描述长度(description length)来计算先验概率 $P(S)$,采用 Laplace 法则来计算条件概率 $P(W|S)$,最后采用动态规划方法(dynamic programming,简称 DP)实现计算过程.Fragkou 等人[17]也采用动态规划方法来实现文本分割过程,其中考虑分割单元内词相似性和分割单元的长度分布,定义了一个分割成本函数(segmentation cost function),该函数包含一个独立变量和 3 个需要从训练语料中训练得到的参数.

本文提出了一个基于 MDA 的全局评价函数 J 来评价具体分割方式,评价值的大小表示分割方式的好坏.评价函数主要考虑 3 个因素:分割单元内距离 WSD、分割单元间距离 BSD 和分割单元的长度 SUL,具体定义和计算方法将在第 2 节进行详细论述.因此,模型(2)求解最大概率的文本分割方式的过程可以转换成求解具有最大评价值的文本分割模式的过程,称为模型(3),计算公式是

$$\hat{S} = \arg\max_{S} P(S \mid B) \overset{\text{def}}{=} \arg\max_{S} J(B,S) \tag{3}$$

2 评价函数

2.1 基本思想

在统计模式分类领域,多元判别分析[18](multiple discriminant analysis,简称 MDA)方法用于类别可分离性判定,是一种有效线性转换方法(linear transformations).该方法能够实现满足最小方差条件下对数据空间进行最佳分割.其基本思想是:数据空间中各类样本可以分开是因为它们位于数据空间的不同区域内,这些区域之间距离越大,类别的可分离性就越大.如图 1 所示.

图 1 中,点表示样本向量表示,根据数据空间的样本分布,可以将数据空间分割成 3 个类别:类别 A、类别 B 和类别 C.当类内平均距离达到最小时,表示类内样本分布凝聚性最强;类间平均距离达到最大,表示不同类别的样本分布发散性最强,可以获得全局最优的数据空间分割效果.

基于这种思想,在文本分割任务中,假设给定文本为数据空间,分割单元为类别,块向量为样本向量,则文本分割的过程就相当于图 1 所示的数据空间分割过程.同理,可以认为当分割单元内距离达到最小(强凝聚性),分割单元间距离达到最大(强发散性),就可以获得全局最优的文本分割模式.在实际情况下,两者可能存在一定的相互制约性,难以同时达到最理想状态.在本文提出的文本分割模型中,采用模型(3)进行文本分割.

在模型(3)中,基于上述思想,本文提出了基于多元判别分析的分割评价函数,称为 MDA 评价函数 J(MDA criterion function).该函数主要考虑了 3 个因素:分割单元内距离 WSD、分割单元间距离 BSD 和分割单元的长度 SUL,可以用于分割模式的全局评价,评价值越大,表示该分割模式越好.下文采用分割单元内散布矩阵 S_W

** 实验显示:当 *blocksize*=100 时,中文文本分割性能最好,因此,在本文的实验中,块的长度为 100 个词.如果文本分割过程不去掉禁用词的话,块的长度计算也不考虑标点符号(,.?!“”等).

(within-segment scatter matrix)和分割单元间散布矩阵 S_B(between-segment scatter matrix)来计算分割单元内距离 WSD 和分割单元间距离 BSD[***].

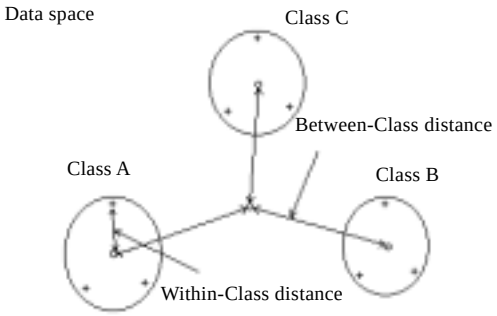

Fig.1 The greatest separation over data space is shown when average within-class distance is the smallest and average between-class distance is the largest

图 1 数据空间的最佳分割:类内距离最小,类间距离最大

2.2 分割单元内散布矩阵S_W

采用第 1 节中的定义,定义给定文本为 $B=b_1b_2\ldots b_k$,其中,下标 k 表示文本 B 包含块的个数,每个块采用 d 维词向量表示,则 b_i 表示 i^{th} 块的向量表示;$S=s_1s_2\ldots s_c$ 表示文本 B 的一个分割模式,其中,下标 c 表示当前文本分割模式 S 包含的分割单元个数.每个分割单元至少包含一个块,则定义分割单元内散布矩阵 S_W 为

$$S_W=\sum_{i=1}^{c}P_i\frac{1}{n_i}\sum_{b\in s_i}(b-m_i)(b-m_i)^t \tag{4}$$

其中:P_i 为分割单元 s_i 的先验概率,等于分割单元 s_i 的块个数与当前文本 B 的所有块个数的比值;n_i 表示分割单元 s_i 中块的个数;m_i 为分割单元 s_i 的 d 维中心向量

$$m_i=\frac{1}{n_i}\sum_{b\in s_i}b \tag{5}$$

2.3 分割单元间散布矩阵S_B

由于分割单元间散布矩阵 S_B 不能直接算出,需要进行简单推导.定义 m 为当前分割模式 S 的总体平均向量

$$m=\frac{1}{n}\sum_{b\in B}b=\frac{1}{n}\sum_{i=1}^{c}n_im_i \tag{6}$$

其中,n 表示当前文本 B 中块的个数,则分割模式 S 的总体散布矩阵 S_T 为

$$S_T=\sum_{i=1}^{c}P_i\frac{1}{n_i}\sum_{b\in s_i}(b-m)(b-m)^t=S_W+\sum_{i=1}^{c}P_i(m_i-m)(m_i-m)^t \tag{7}$$

满足

$$S_T=S_W+S_B \tag{8}$$

根据式(7)和式(8),可以得到分割单元间散布矩阵 S_B:

$$S_B=\sum_{i=1}^{c}P_i(m_i-m)(m_i-m)^t \tag{9}$$

*** 分割单元内距离 WSD 相当于分割单元内散布矩阵 S_W 的迹 $tr(S_W)$;同样,分割单元间距离 BSD 相当于分割单元间散布矩阵 S_B 的迹 $tr(S_B)$.矩阵的迹等于对角线数字之和.

2.4 长度因子S_L

文本分割的实验结果显示:一些位置非常接近的分割点产生一些不正确的较小分割单元.为此,本文假设包含一个或很少数目句子的文本片断难以表达一个独立的主话题.例如,文本中存在与前后文本片断相关性不强这样一个句子,则可以猜测该句子可能是一个插入语,或者是为了论述连贯性的目的而增加的一个句子,实现承上启下的目的,不能独立表达一个独立子主题.

基于上述考虑,本文提出的评价函数考虑了分割单元的长度分布,定义了一个长度因子 S_L 来解决这个问题.如果当前分割模式 S 中存在较小的分割单元,则长度因子 S_L 被赋予较小值,起到一个惩罚的作用.

首先定义 L 为文本 B 的长度,等于该文本包含词的个数;L_i 表示分割单元 s_i 的长度,等于分割单元 s_i 包含词的个数;很明显,$L=L_1+L_2+\ldots+L_c$.长度因子 S_L 定义为

$$S_L=\prod_{i=1}^{c}\frac{L_i}{L} \tag{10}$$

2.5 MDA评价函数J

在模型(3)中,为了评价给定文本 B 的分割模式 S,本文定义了 4 种 MDA 评价函数:

1) 考虑分割单元内距离 WSD 和分割单元间距离 BSD,定义 MDA 评价函数 J_1:

$$J_1(B,S)=\frac{tr(S_B)}{tr(S_W)} \tag{11}$$

2) 考虑分割单元内距离 WSD 和分割单元间距离 BSD,定义 MDA 评价函数 J_2

$$J_2(B,S)=tr(S_B)\times tr(S_W) \tag{12}$$

3) 考虑分割单元内距离 WSD、分割单元间距离 BSD 和分割单元的长度 SUL,定义 MDA 评价函数 J_3:

$$J_3(B,S)=S_L\times\frac{tr(S_B)}{tr(S_W)} \tag{13}$$

4) 考虑分割单元内距离 WSD、分割单元间距离 BSD 和分割单元的长度 SUL,定义 MDA 评价函数 J_4:

$$J_4(B,S)=S_L\times tr(S_B)\times tr(S_W) \tag{14}$$

其中,$tr(.)$表示矩阵的迹,等于矩阵对角线元素之和.

在本文的实验中,将分别使用上述 4 种 MDA 评价函数来实现文本分割过程.实验结果显示:J_3 和 J_4 的组合性能最好.另外,J_1 和 J_3 用于文本分割过程中的主题边界自动识别;J_2 和 J_4 用于文本分割过程中的分割单元数目的自动确定.

3 实现算法

本文采用模型(3)来实现文本分割过程,其中分别采用 4 种 MDA 评价函数 J_1,J_2,J_3 和 J_4 对文本分割模式 S 进行全局评价.文章最开始部分提到,文本分割技术需要解决两个关键问题:主题边界的自动识别和分割单元数目确定.

在本文的文本分割模型中,实现算法分为两步:

第 1 步:在给定当前文本 B 的分割单元数目已知条件下,利用 MDA 评价函数 J_1 或 J_3 自动识别主题边界.实现算法见算法 1.

算法 **1**. 自动识别主题边界基本算法.

Given a text $B=b_1b_2\ldots b_k$, where k is the number of blocks in B; c is the given number of segments
Initialization: $S_{best}=\{\}$, $J(B,S_{best})=0.0$
Text segmentation:
Begin
1) Construct possible segmentation set $SSet=\{S_1,S_2,\ldots,S_n\}$
Loop
2) Get a segmentation S from $SSet$, and delete S from $SSet$;
3) If $J(B,S_{best})<J(B,S)$ Then

Begin
$S_{best}=S$ and $J(B,S_{best})=J(B,S)$.
Endif
Until $SSet=\{\}$.
End
Output best segmentation S_{best}.

在算法 1 中的第 3)步,采用 MDA 评价函数 J_1 或 J_3 来计算 $J(B,S)$评价值.实验分别给出了比较结果.本文主要研究基于段落层次的文本分割技术[****],也就是说,分割点只能出现在每个段落的结束位置.从该算法可以看出计算复杂度为 $O(C_m^c)$,其中:m 表示当前文本 B 包含的段落数目;c 表示给定分割单元的数目.

可想而知,这个计算的复杂度还是很高的.为了解决这个问题,本文采用遗传算法[19](genetic algorithms,简称 GAs)来实现算法 1 的实现过程,达到优化算法的计算复杂度的目的.实验采用 MATLAB 中的遗传算法模块来实现上述算法.

第 2 步:在算法 1 分析结果的基础上,自动确定当前文本的最佳分割单元数目.即在未知文本分割数目的基础上,利用 MDA 评价函数 J_2 或 J_4 来确定最佳的文本分割模式.实现算法见算法 2.

算法 **2**. 自动确定分割单元数目基本算法.

Given a text $B=b_1b_2\ldots b_k$, where k is the number of blocks in B; m is the number of paragraphs in B
Initialization: $S_{best}=\{\}$, $J(B,S_{best})=0.0$
Text segmentation:
Begin
For K =2 to m
Begin
1) Suppose K is the desired number of segments, determine the best segmentation S with algorithm 1;
2) If $J(B,S_{best})<J^*(B,S)$ Then
Begin
$S_{best}=S$ and $J(B,S_{best})=J^*(B,S)$.
Endif
End For
End
Output best segmentation S_{best}.

在算法 2 中的第 2)步,采用 MDA 评价函数 J_2 或 J_4 来计算 $J^*(B,S)$评价值.本文实验分别给出了比较分析结果.从算法中可以看出计算复杂度为 $O(m)$,其中,m 表示当前文本 B 包含的段落数目,因为文本分割单元的数目肯定少于文本段落数目.

4 实 验

4.1 评测语料

目前还没有一个公开的通用的中文文本分割评测语料.为此,本文构造了一个规模为 106 篇的中文语料库作为评测数据集.语料来源于电子版的人民日报,体裁和内容较为广泛,涵盖了科技说明文、人物传记、时事评论等领域.考虑到通常情况下的中文文本长度,语料库中的每篇文本选取上主要集中于 5~8 个语义段落,平均每篇的自然段落数目为 25.8 个.每篇文本的标准分割模式通过人工方式标注给定[*****].本实验中,测试语料分为两部分:测试语料 1(5 个语义段落)和测试语料 2(6~8 个语义段落).

**** 本算法也可以直接用于基于句子层次的文本分割过程,假设分割点可以出现在任何两个句子之间;同理,该算法也可以直接用于基于词层次的文本分割过程,假设分割点可以出现在任意两个词之间;很明显,计算复杂度将比本算法的计算复杂度更高,这也是本文下一步研究工作的重点.

***** 实验室的 3 位研究生通过相互讨论方式,人工构建了该评测语料及其参考标准分割答案.

4.2 评价方法

本文首先采用传统的正确率(precision)、召回率(recall)、$F1$ 值来评价文本分割算法的性能.正确率是指分割结果中正确分割点个数占所有分割点的比重;召回率是指算法正确判断的分割点个数占标准答案中分割点的比重;$F1$ 值按下式计算:

$$F1-Measure=\frac{2\times Precision\times Recall}{Precision+Recall} \tag{15}$$

对于文本分割评价来说,用传统的正确率和召回率并不能全面和公正地评价分割性能.主要原因在于:用正确率和召回率主要考虑绝对匹配的结果,实际上,离正确分割点较近的错误分割点比较远的错误分割点性能更好,但是正确率和召回率一视同仁,无法体现出这种差别[20].

为了克服上述缺点,本文同时采用了 Beeferman 等人[20]提出的 P_μ 评测方法来评价文本分割系统的性能.P_μ 评测方法******的具体计算公式是

$$P_\mu(ref,hyp)=\sum_{1\le i\le j\le n}D_\mu(i,j)(\delta_{ref}(i,j)\overline{\oplus}\delta_{hyp}(i,j)) \tag{16}$$

其中,ref 指标准答案中参考分割模式;hyp 是指系统给出的分割模式;n 为文本里的句子数;$\delta_{ref}(i,j)$为指示函数,当句子 i 和 j 在 ref 中同属一个分割单元时,其值为 1,否则为 0;同理,当句子 i 和 j 在 hyp 中同属一个分割单元时,$\delta_{hyp}(i,j)$值为 1,否则为 0.上式对$\delta_{ref}(i,j)$和$\delta_{hyp}(i,j)$值进行异或运算.D_μ为随机选取的句子对在文本中的距离概率分布,其值依赖于参数μ.其中,D_μ为参数为μ的指数分布[20]:

$$D_\mu(i,j)=r_\mu e^{-\mu|i-j|} \tag{17}$$

其中,r_μ为归一化因子.

在本文的实验中,参数 $1/\mu$设为 11,是测试文本集中分割单元包含的平均句子数.文本预处理使用的分词系统采用由东北大学自然语言处理实验室开发的 CipSegSDK 系统[21].

4.3 实验结果

实验中,采用 Hearst 提出的 TextTiling 算法[2]和 Reynar 提出的 Dotplotting 算法[6]作为 Baseline 系统,其中 Dotplotting 算法只能支持给定分割单元数目条件下的文本分割,因此,在本实验中,只在分割单元数目已知的条件下分析 Dotplotting 算法的性能.为了正确评测各种文本分割模型的性能,本文设计两个实验进行比较分析.

在实验 1 中,预先给定文本分割单元数目,该实验的目的在于评测本文提出的 MDA 方法与 TextTiling 和 Dotplotting 算法对主题边界的自动识别能力.表 1 中带有下划线和加黑的数字表示单项总体平均最高指标.从表 1 可以看出:两个 MDA 模型(J_1 和 J_3)比 TextTiling 和 Dotplotting 算法具有更好的 P_μ评价值、正确率、召回率和 $F1$ 值.

另外,从表 1 中可以看出:虽然使用评价函数 J_1 的 MDA 模型取得最好的正确率、召回率和 $F1$ 值,但是,P_μ 评价值指标不如使用评价函数 J_3 的 MDA 模型.由此可以看出:第一,正如前面论述的,正确率和召回率属于绝对匹配方法,由于对较小偏差的错误分割与较大偏差的错误分割,无法给出公正评价;而 P_μ评价值指标就能够较好地反映出上述两个错误分割的不同,前者的性能好于后者.所以,具有较高正确率和召回率的模型不一定具有较好的 P_μ评价值;第二,评价函数 J_3 使用了分割单元长度因子,一定程度上考虑了分割单元的长度分布,对较小的错误分割单元能够起到一定的纠正作用.即考虑分割单元长度分布有助于改善文本分割的性能.这一点与 Fragkou 等人[17]的实验结果和分析是一致的.

在实验 2 中,由于没有给定文本分割单元的数目,MDA 模型和 TextTiling 算法需要在识别主题边界的基础

****** 实际上,Beeferman 等人又提出了 P_k 评价方法,有些研究人员采用该方法进行评价文本分割的性能.但是,P_k 评价方法存在一些问题,包括评价方法不直观、对分割单元数目的惩罚力度不够等,详细分析参见文献[22].本文主要研究基于段落层次的文本分割技术,我们认为,P_k 评价方法比较适合于基于句子层次和基于词层次的文本分割技术的性能.

上自动确定文本分割单元的最佳数目.由于 Dotplotting 算法难以实现自动确定文本分割单元的数目,所以本实验中没有采用 Dotplotting 算法作为 Baseline 系统.

Table 1 Comparison experimental results with known number of document segments

表 1 分割单元数目已知条件下的比较实验结果

Algorithm	Evaluation etric	Testing corpus 1 (5 segments)	Testing corpus 2 (6~8 segments)	Average
MDA model (J_1)	*Precision*	0.459	0.519	**0.486**
	Recall	0.459	0.519	**0.486**
	*F*1	0.459	0.519	**0.486**
	P_μ	0.854	0.889	0.869
MDA model (J_3)	*Precision*	0.448	0.445	0.447
	Recall	0.448	0.445	0.447
	*F*1	0.448	0.445	0.447
	P_μ	0.899	0.913	**0.906**
TextTiling (Hearst)	*Precision*	0.424	0.389	0.408
	Recall	0.424	0.389	0.408
	*F*1	0.424	0.389	0.408
	P_μ	0.814	0.839	0.825
Dotplotting (Reynar)	*Precision*	0.371	0.385	0.389
	Recall	0.371	0.385	0.389
	*F*1	0.371	0.385	0.389
	P_μ	0.719	0.700	0.709

从表 2 的实验指标可以看出:MDA 模型比 TextTiling 算法具有更好的 P_μ评价指标.实际上,在文本分割过程中,MDA 模型采用了全局评价函数进行主题边界识别和文本分割单元数目的确定.从算法 2 可以看出:在 MDA 模型中,最佳分割模式的主题边界识别和文本分割单元数目的确定是同时实现的,属于全局优化过程.

从表 2 可以看出:使用评价函数 J_3和 J_4的 MDA 模型比使用评价函数 J_1和 J_2的 MDA 模型取得了更好的召回率、*F*1 值和 P_μ评价指标.虽然后者比前者获得更好的正确率,但只提高 0.3%,而前者比后者的召回率高出 2.1%,*F*1 值高出 0.9%,P_μ评价指标高出 3.8%.从这点可以看出:考虑文本分割单元长度分布,有助于改善分割单元数目确定性能.

Table 2 Comparison experimental results with unknown number of document segments

表 2 分割单元数目未知条件下的比较实验结果

Algorithm	Evaluation etric	Testing corpus 1 (5 segments)	Testing corpus 2 (6~8 segments)	Average
MDA model (J_1 and J_2)	*Precision*	0.485	0.457	**0.473**
	Recall	0.440	0.514	0.474
	*F*1	0.462	0.483	0.473
	P_μ	0.799	0.870	0.832
MDA model (J_3 and J_4)	*Precision*	0.485	0.452	0.470
	Recall	0.455	0.544	0.495
	*F*1	0.470	0.494	**0.482**
	P_μ	0.840	0.906	**0.87**
TextTiling (Hearst)	*Precision*	0.425	0.350	0.391
	Recall	0.452	0.568	**0.504**
	*F*1	0.438	0.433	0.441
	P_μ	0.781	0.842	0.808

4.4 相关研究对比

本文探讨领域无关的线性文本分割方法,定义 4 种全局评价函数,实现对文本分割的全局评价,同时自动确定分割数目.与本文的研究最相关的工作包括 TextTiling[2],Dotplotting[6]等.

对比实验结果表明:MDA 模型具有更好的性能,体现出基于全局评价的方法比基于局部评价的方法更有效.主要原因在于 MDA 评价函数 J_1和 J_3是全局评价函数,属于全局最优方法.TextTiling 算法主要通过确定相邻文本片断的相似性变化程度来猜测主题边界,属于局部最优方法.Dotplotting 技术严格上来说,考虑了全局优化和局部优化一些特性,介于两者之间,主要考虑类内密度和类外密度,实现主题边界识别.

在分割单元数目的确定方面,Dotplotting 算法难以实现自动确定文本分割单元的数目,TextTiling 算法主要

采用一个简单的 cutoff 函数,根据相邻文本片断的相似度改变程度来确定文本分割单元数目.寻找基于相邻文本片断的相似度变化曲线的“波峰”和“波谷”,选择最大落差相似度改变的位置作为分割点.该方法属于局部最优方法,难以实现全局优化过程.由此可见,全局评价方法在分割单元数目的确定上具有一定的优势.

TextTiling 算法比 MDA 模型具有更好的召回率,表明简单利用相邻片断相似性变化程度可以找出更多的正确分割点,但也引入较多的错误分割点,这一点从较低的正确率中可以看出.过多的错误分割点的引入,造成 $F1$ 和 P_μ评价指标下降.

与其他一些基于统计模型的文本分割算法相比[15−17],本文的算法无须训练语料,是一种领域无关的方法,具有更好的通用性.

5 结束语

本文深入研究了文本分割的两个关键问题:主题边界的自动识别和分割单元数目的确定,并提出了一个独立于具体领域的文本分割统计模型——MDA 模型,其中,采用多元判别分析方法定义 4 种分割全局评价函数,实现对文本分割的全局评价.该评价函数主要考虑了分割单元内距离、分割单元间距离和分割单元长度分布信息.实验结果显示:MDA 模型总体性能优于 TextTiling 和 Dotplotting 算法,反映了全局评价方法比局部评价方法具有更好的文本分割性能.通过对 4 种 MDA 评价函数的比较分析,证明了考虑文本分割单元的长度分布信息将有助于主题边界识别和分割单元数目的确定.但是,本文提出的 MDA 模型具有较高的计算复杂度,为此,本文采用了遗传算法来解决计算复杂度问题.该高复杂度的问题主要是由于本文提出的 MDA 模型是一个无序模型(disordered model),无法采用动态规划算法来实现优化.为解决计算复杂度过高的问题,下一步将深入研究属于有序模型(ordered model)的 MDA 文本分割模型,对分割单元内距离和分割单元间距离的计算方式进行改进,使之适应动态规划或其他复杂度较低的搜索策略的要求.也将深入研究其他有效的优化算法或近似优化算法来解决模型的高复杂度问题.在未来的工作中,将考虑如何将基于 MDA 的文本分割技术用于文档摘要、信息检索、问答系统等应用中.

致谢 在本文的研究工作中,非常感谢 Keh-Yih Su,Matthew Ma 和 Benjamin K Tsou 教授提出意见,也非常感谢常兴治同学、陈文亮博士所做的一些实验工作.

References:

[1] Salton G, Singhal A, Buckley C, Mitra M. Automatic text decomposition using text segments and text themes. In: Bernstein M, Carr L, Osterbye K, eds. Proc. of the 7th ACM Conf. on Hypertext. New York: ACM Press, 1996. 53−65.

[2] Hearst MA. TextTiling: Segmenting text into multi-paragraph subtopic passages. Computational Linguistics, 1997,23(1):33−64.

[3] Morris J, Hirst G. Lexical cohesion computed by thesauri relations as an indicator of the structure of text. Computational Linguistics, 1991,17(1):21−42.

[4] Kozima H. Text segmentation based on similarity between words. In: Proc. of the 31st Annual Meeting of the Association for Computational Linguistics. 1993. 286−288. http://acl.ldc.upenn.edu/P/P93/P931041.pdf

[5] Passoneau RJ, Litman DJ. Intention-Based segmentation: Human reliability and correlation with linguistic cues. In: Proc. of the 31st Meeting of the Association for Computational Linguistics. 1993. 148−155. http://acl.ldc.upenn.edu/P/P93/P931020.pdf

[6] Reynar JC. Topic segmentation: Algorithms and application [Ph.D. Thesis]. Pennsylvania: University of Pennsylvania, 1998.

[7] Ponte JM, Croft WB. Text segmentation by topic. In: Peters C, Thanos C, eds. Proc. of the 1st European Conf. on Research and Advanced Technology for Digital Libraries. Berlin, Heidelberg: Springer-Verlag, 1997. 120−129.

[8] Reynar JC. Statistical models for topic segmentation. In: Proc. of the 37th Annual Meeting of the Association for Computational Linguistics. 1999. 357−364. http://acl.ldc.upenn.edu/P/P99/P991046.pdf

[9] Kauchak D, Chen F. Feature-Based segmentation of narrative documents. In: Proc. of the 43rd Annual Meeting of the Association for Computational Linguistics. 2005. 32−39. http://acl.ldc.upenn.edu/W/W05/W05-04.pdf

[10] Choi FYY. Advances in domain independent linear text segmentation. In: Proc. of the North American Chapter of the Association for Computational Linguistics Annual Meeting. Seattle: Association for Computational Linguistics. 2000. http://acl.ldc.upenn.edu/A/A00/A002004.pdf

[11] Choi FYY, Wiemer HP, Moore J. Latent semantic analysis for text segmentation. In: Lee L, Harman D, eds. Proc. of the 6th Conf. on Empirical Methods in Natural Language Processing. Somerset: Association for Computational Linguistics. 2001. 109−117.

[12] Blei DM, Moreno PJ. Topic segmentation with an aspect hidden Markov model. In: Croft BW, Harper DJ, Kraft DH, Zobel J, eds. Proc. of the 24th Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. New York: ACM Press, 2001. 343−348.

[13] Yaari Y. Segmentation of expository texts by hierarchical agglomerative clustering. In: Mitkov R, Nicolov N, Nikolov N, eds. Proc. of the Conf. on Recent Advances in Natural Language Processing. Series: Current Issues in Linguistic Theory. 1997. 59−65.

[14] Ji X, Zha H. Domain-Independent text segmentation using anisotropic diffusion and dynamic programming. In: Callan J, Cormack G, Clarke C, Hawking D, Smeaton A, eds. Proc. of the 26th Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. New York: ACM Press, 2003. 322−329.

[15] Heinonen O. Optimal multi-paragraph text segmentation by dynamic programming. In: Proc. of 17th Int'l Conf. on Computational Linguistics. San Francisco: Morgan Kaufmann Publishers, 1998. 1484−1486. http://acl.ldc.upenn.edu/P/P98/P982244.pdf

[16] Utiyama M, Isahara H. A statistical model for domain-independent text segmentation. In: Proc. of the 9th Conf. of the European Chapter of the Association for Computational Linguistics. 2001. 491−498. http://acl.ldc.upenn.edu/P/P01/P011064.pdf

[17] Fragkou P, Petridis V, Kehagias A. A dynamic programming algorithm for linear text segmentation. Journal of Intelligent Information Systems, 2004,23(2):179−197.

[18] Duda R, Hart P, Stork D. Pattern Classification. 2nd ed., John Wiley & Sons, 2001.

[19] Mitchell TM. Machine Learning. McGraw-Hill, 1997.

[20] Beeferman D, Berger A, Lafferty J. Text segmentation using exponential models. In: Cardie C, Weischedel R, eds. Proc. of the 2nd Conf. on Empirical Methods in Natural Language Processing. Somerset: Association for Computational Linguistics. 1997. 35−46.

[21] Yao TS, Zhu JB, Zhang L, Yang Y. Natural Language Processing-Research on Making Computers Understand Human Languages. 2nd ed., Beijing: Tsinghua University Press, 2002 (in Chinese).

[22] Pevzner L, Hearst M. A critique and improvement of an evaluation metric for text segmentation. Computational Linguistics, 2002, 28(1):19−36.

附中文参考文献:

[21] 姚天顺,朱靖波,张俐,杨莹.自然语言理解——一种让机器懂得人类语言的研究.第 2 版.北京:清华大学出版社,2002.

朱靖波(1973－),男,浙江金华人,博士,教授,CCF 高级会员,主要研究领域为计算语言学理论.

叶娜(1981－),女,博士生,主要研究领域为文本分割,信息抽取.

罗海涛(1974－),男,硕士生,主要研究领域为文本分割.