Dotplotting 文本分割技术的分析与改进

罗海涛，叶娜，朱靖波

（东北大学自然语言处理实验室，沈阳 110004）

摘 要： 在线性文本分割领域，Dotplotting 是一个很著名的方法。本文对 Dotplotting 方法进行了详细的分析，发现了其中存在的两个问题。一个问题是，Dotplotting 方法的分割评价函数不是对称的，造成分割方式随阅读方向的不同而不同，这是违背事实的；另一个问题是，有时分割结果中会出现长度过短的语义段落，而长度过短的文本片段很难清楚地说明一个独立的主题。针对上述两个问题，本文通过在评价函数中同时考虑正向分割和反向分割，并加入惩罚因子，提出了改进的模型。实验结果表明改进后的模型比原始 Dotplotting 模型的性能有很大提高。

关键词： 点阵；正向分割；反向分割；长度惩罚因子

Analysis and Improvement of the Dotplotting Method for Text Segmentation

Luo Haitao, Ye Na, Zhu Jingbo

(Natural Language Processing Lab Northeastern University, Shenyang, 110004)

Abstract: Dotplotting is a well-known method in the domain of linear text segmentation. This paper gave a detailed analysis of the Dotplotting method, and discovered two faults in it. One is that the segmentation evaluation function is asymmetric, so the segmentation result varies with reading direction. This is in collision with fact. Another is that some semantic segments with too small lengths may appear in the segmentation result although a too short text can hardly explain a topic clearly. Focusing on these two faults, this paper presented improved models by considering forward and backward segmentation and applying length tradeoff in evaluation function. Experimental results show that the new models perform much better than the original Dotplotting model.

Keywords: Dotplotting; forward segmentation; backward segmentation; length tradeoff

1 引言

一般来讲，一篇文章会包含多个主题，即使是只有一个主题的文章，通常也会包含多个子主题，这些主题或子主题统称为语义段落。线性文本分割的目的是要找到这些语义段落的边界。在自然语言处理的很多应用领域中，如自动文摘和信息检索，文本分割都是一项重要的任务，它可以提高这些系统的性能，并避免把无用的信息提供给用户。

目前，已经有很多对文本进行线性分割的方法。它们主要可以分为两类：无监督方法[1,2,3,4,5,6]和有监督方法[7,8,9,10,11]。有监督方法具有较高的分割性能，但它们都需要大量的训练语料。无监督方法一般不需要训练语料，

本文工作部分得到国家自然科学基金（No. 60473140）和 国家教育部新世纪优秀人才计划项目资助。

作者简介：罗海涛（1974－），男，黑龙江省双鸭山市，在读硕士，E-mail:luohtao@gmail.com

且能够独立于领域。

在线性文本分割领域中，Dotplotting 是一个很著名的无监督方法，本文对它进行了比较详细的分析，发现了其中存在的两个问题。一个问题是，Dotplotting 的分割评价函数不是对称的，造成分割方式随阅读方向的不同而不同，这与事实相违背，叶娜[2]也对此进行了论述，并提出了改进的方案。另一个问题是，有时分割结果中会出现长度过短的语义段落，而长度过短的文本片段是很难清楚地说明一个独立的主题的。

为解决第一个问题，本文借鉴叶娜[2]提出的方法将分割评价函数对称化，从而使分割结果独立于阅读方向；对于第二个问题，本文在评价函数中加入了长度惩罚因子，对存在过短或过长语义段落的分割方式进行抑制。综合以上两点，本文提出了一个改进的模型。实验结果表明，改进的模型在测试语料上的性能显著高于原始的 Dotplotting 方法。

本文剩余部分组织安排如下：第二节对 Dotplotting 方法进行简要介绍和分析，第三节指出 Dotplotting 方法存在的问题，提出相应的解决方法。第四节给出对比实验结果和实验数据分析。最后，在第五节进行总结并介绍未来的工作。

2 Dotplotting 方法

在文本分割领域中，Dotplotting 是一种基于词汇聚合度和图像分析技术的方法，它通过一个反映整篇文本词汇重现情况的二维点阵图来寻找主题边界，具有全局性。

假设一篇文本的长度为 n 个词，如果某个词在文本中出现的位置集合为 $I=\{p_1,p_2,...,p_i\}(1\le i\le n)$，则在点阵图上由 $\{(p_i,p_j)\mid p_i\in I,p_j\in I\}$ 表示的所有坐标上用一个点标出该词。这样就把整篇文本表示为一个对称的二维点阵图。从这个图上，可以清楚地看到文本中词汇的密度分布情况。一般地说，语义段落内部的词汇重复程度会比较高，对应区域的点也会比较密集，以此来确定主题边界。

确定语义段落边界时，初始边界集合只有文本开始和结尾两个固定边界，每一次循环从候选的边界集合中根据评价函数选出一个最佳边界，并加入到已确定的语义段落边界集合中，直到边界数目达到指定的值为止。

对于一个候选的语义段落分割边界，其评价函数为

$$f_s=\sum_{j=2}^{|P|}\frac{V_{P_{j-1},P_j}\cdot V_{P_j,n}}{(P_j-P_{j-1})(n-P_j)} \qquad (1)$$

其中 P_j 表示第 j 个边界的位置，$V_{x,y}$ 是一个向量，它记录文本中位置 x 和位置 y 之间每个词条的出现次数。

从公式(1)可以看出，Dotplotting 沿主对角线计算由当前候选边界和已确定的边界集合分割出的各个语义段落区域上方各矩形区域密度之和，并将其作为当前候选段落边界的评价函数值。在选取最佳段落边界时，寻找评价函数值最小的边界作为最佳边界，从而使语义段落内部的总体词汇聚合度最大，语义段落之间的词汇聚合度最小。

3 Dotplotting 方法的分析与改进

3.1 正向分割结果与反向分割结果不一致的问题

从实际阅读经验来看，对于一篇文本来说，虽然人们习惯于从前向后阅读，但是文本的语义段落边界不应该因阅读方向的不同而改变。

以上假设可以从点阵图的角度解释如下，用图 1 表示对一篇文本的分割情况：图 1 中左下和右上顶点分别表示文本的开始和结尾，此时文本中共有 3 个语义段落，B1 和 B2 为两个语义段落分割点。Dotplotting 的原始评价函数为正向分割，根据公式(1)，I 和 II 是选取分割点时所考察的两个密度区域。

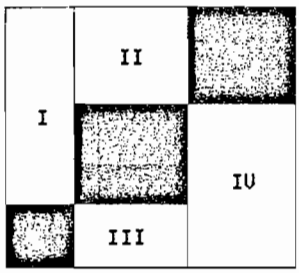

图 1 正向分割和反向分割示意图

Fig. 1 An illustration of forward segmentation and backward segmentation

从公式(1)可以看出，如果从后向前阅读文本，则分割点评价函数相应变为：

$$f_v = \sum_{j=2}^{|P|} \frac{V_{P_j,n} \cdot V_{0,P_j}}{(n-P_j)(P_j-0)} \tag{2}$$

根据公式(2)，反向分割在选取分割点时，所考察的两个密度区域是区域 III 和区域 IV。然而，由于在文本中词汇并不是均匀分布的，区域 I 和 II 的密度之和并不等于区域 III 和 IV 的密度之和，所以在计算分割点的评价函数值时，正向分割和反向分割方式的函数值不相同，导致正向阅读和反向阅读产生不同的分割结果。

为解决上述问题，本文对 Dotplotting 的分割点评价函数进行了修改，同时考虑正向阅读和反向阅读时的词汇聚合度，得到如下的评价函数：

$$f = \sum_{j=2}^{|P|} \frac{V_{P_{j-1},P_j} \cdot V_{P_j,n}}{(P_j-P_{j-1})(n-P_j)} + \sum_{j=2}^{|P|} \frac{V_{P_j,n} \cdot V_{0,P_j}}{(n-P_j)(P_j-0)} \tag{3}$$

公式(3)中，在评价分割点时，同时考察区域 I，II，III 和 IV 的密度，使分割评价函数对称化，从而使得正向分割和反向分割取得相同的分割结果，分割结果独立于阅读方向。

3.2 存在长度过短的语义段落的问题

一篇文章中，虽然各个语义段落的长度不等，但是不应该存在长度过短的语义段落，因为过短的文本片段很难清楚地表达一个独立的主题。如果某种候选分割方式含有过短的语义段落，那么该分割方式往往包含错误，需要对其进行抑制。

因此，本文采用长度惩罚因子来抑制过短或过长语义段落出现的可能性。本文定义惩罚因子如下：

$$LenTradeOff = \prod_{i=1}^{k} \frac{L_i}{L}$$

其中 L_i 为第 i 个语义段落的长度（词数），L 为文本的总长度，$\sum_{i=1}^{k} L_i = L$。对每一种分割方式，计算各语义段落在文本中出现的概率，即语义段落包含的词数与文本总词数的比值，并将它们相乘即为惩罚因子。由于惩罚因子在各语义段落长度相等时达到最大，所以，将分割评价函数值除以惩罚因子可以达到调节语义段落长度大小的效果。

结合考虑双向切割并应用长度惩罚因子后，得到如下的评价函数：

$$f_d = \left(\sum_{j=2}^{|P|} \frac{V_{P_{j-1},P_j} \cdot V_{P_j,n}}{(P_j-P_{j-1})(n-P_j)} + \sum_{j=2}^{|P|} \frac{V_{P_j,n} \cdot V_{0,P_j}}{(n-P_j)(P_j-0)}\right) \Big/ \prod_{i=1}^{k} \frac{L_i}{L} \tag{4}$$

公式(4)为改进后的模型的评价函数，该模型解决了 Dotplotting 原始评价函数中存在的两个问题。

3.3 总体分割算法

改进模型的文本分割算法总体流程如下图所示：

```
给定文本 W，K 为最佳的语义段落数目
初始化
    B={0, n}，Dmin=+∞，C={i | i 为 W 中可能的语义段落边界}，S=0
语义段落数从 1 递增到 K
    对 C 中的每一个元素 i
        P=B ∪ {i}
        计算分割方式 P 下的评价函数 D
        若 Dmin>D
            Dmin=D and S=i
    B=B ∪ {S}
    C=C-{S}
结束并输出分割结果 B
```

图 2　文本分割算法

Fig. 2 Text Segmentation Algorithm

算法把文本 W 的开始和结尾两个固定边界作为初始的已确定边界，每一次循环从候选边界集合中根据评价函数 D 选出一个最佳边界加入到已确定边界的集合中，直至语义段落数达到指定值 K 为止。

4　实验结果

4.1　测试语料及评价方法

本文中的测试语料来自 [3]，该语料在文本分割领域被广泛使用。语料集共有 700 篇英文文章，每篇文章由 10 个语义段落构成，而每个语义段落由 Brown 语料库中随机选择一篇文章中的前 n 个句子自动生成。根据 n 的不同，整个语料集被分为 4 个子集。每个子集所包含的文章数目见下表：

表 1 测试语料数目

Tab.1 Number of articles in the testing corpus

语料集编号	1	2	3	4
n 的范围	3-11	3-5	6-8	9-11
文章数目	400	100	100	100

在评价分割算法的性能时，本文使用 P_k 值作为评价标准[11]。P_k 值反映的是被错分的句子的比例，其值越低，说明分割的错误率越低，分割的结果越接近正确。它修正了以往用准确率和召回率评价时存在的问题，是目前评价文本分割算法的性能时最常用的一种评价方法。

4.2　实验结果

本文实验中所使用的 Baseline 系统是 Reynar 在 1993 年提出的 Dotplotting 系统[1]。

表 2 为 4 个测试语料集上的实验结果。表中的数值是由相应语料中所有分割结果的 P_k 值取平均值得到的。

表 2 对比实验结果

Tab. 2 Comparative experimental results

方法	测试集			
	3-11	3-5	6-8	9-11
Dotplotting	0.345	0.337	0.364	0.381
改进模型 1	0.259	0.274	0.289	0.296
改进模型 2	**0.203**	**0.169**	**0.157**	**0.133**

表 2 中，改进模型 1 是以公式(3)作为分割评价函数的改进模型；改进模型 2 是以公式(4)作为分割评价函数的改进模型。

从实验结果可以看出，改进模型 1 的性能比原始的 Dotplotting 方法在各个语料集上都有一定程度的提高，验证了本文前面对 Dotplotting 方法存在的问题的分析，同时说明在正向分割的基础上加入对反向分割的考察确

实有助于提高文本分割的性能；与改进模型 1 相比，改进模型 2 在各个语料集上的性能又都得到了较大幅度提高。说明通过加入长度惩罚因子来避免出现过短或过长的语义段落对提高性能有明显的效果。

4.3 错误分析

首先，通过分析表 2 的实验结果可以看出，随着测试集里语义段落长度的增加，Dotplotting 评价函数的性能变得越来越差，在将评价函数进行改进后也是如此，说明当语义段落的长度变大时，词的共现几率增加，Dotplotting 评价函数(1)区分语义段落的能力变弱。

而加入惩罚因子后，情况正好相反，在语义段落长度越大的语料集上性能越好，这是因为在一篇文本中边界分布得越均匀，加入长度惩罚因子对提高分割性能起到的作用就越大。而在语义段落之间长度差别较大的情况下，如语料集 1，加入长度惩罚因子对性能的提高就相对有限。由此看来，长度惩罚因子的计算方法还具有一定的局限性。

其次，根据 Dotplotting 的搜索策略，如果找到的第一个语义段落边界是不正确的，那么将影响到后面找出的所有边界的正确性，从这一点来看，边界搜索策略还有待改进。

5 结论以及未来的工作

本文分析了文本分割领域中的 Dotplotting 方法存在的两个问题，并针对这两个问题改进了 Dotplotting 的评价函数。改进后的模型使评价函数对称化，并应用长度惩罚因子来抑制过短或过长的语义段落。本文在文本分割研究中广泛使用的英文测试语料集上进行了实验和分析，实验结果表明所做的两个改进都对分割的性能有明显的提高。

本文方法的局限是只能在给定语义段落数目的情况下进行分割，无法自动确定语义段落的数目，也就是说，在把所有候选的分割边界加入到已经确定的边界集合之前，算法无法自动终止。下一步准备寻找能够自动确定语义段落数目的有效方法。

另外，如前文所述，边界搜索策略还存在着一定的不足，用来均衡语义段落长度的惩罚因子的定义也具有一定的局限性，寻找一种更加合理有效的搜索策略和惩罚因子也是下一步研究的重要内容之一。

参考文献：

[1] Jeffrey C. Reynar. An automatic method of finding topic boundaries. In proceedings of ACL'94(Student session). 1994.

[2] Na YE, Jingbo ZHU, Haitao LUO. Improvement of the Dotplotting Method for Linear Text Segmentation. In Proceedings of the Natural Language Processing and Knowledge Engineering. 2005. pp.636-641.

[3] Choi, F.Y.Y. Advances in domain independent linear text segmentation. In Proceedings of the 1st Meeting of the North American Charpter of the Association for Computational Linguistics. 2000. pp.26-33.

[4] P.Fragkou, V.Petridis, Ath.Kehagias. A Dynamic Programming Algorithm for Linear Text Segmentation. Journal of Intelligent Information Systems, 2004. 23:2, pp.179-197.

[5] Ji, X., Zha. Domain-independent Text Segmentation Using Anisotropic Diffusion and Dynamic Programming. In Proceedings of the 11th International Conference on Information and Knowledge Management, 2002. pp.211-218.

[6] M. A. Hearst. Multi-paragraph text segmentation of expository text. In Proceedings of the ACL'94, 1994. pp.9-16.

[7] H. Kozima. Text Segmentation based on similarity between words. In Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics. 1993, pp.286-288.

[8] Utiyama M, Isahara H. A Statistical Model for Domain-Independent Text Segmentation. In Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics. 2001. pp.491-498.

[9] D. M. Blei and P. J. Moreno. Topic segmentation with an aspect hidden markov model. In Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval. 2001. pp.343-348.

[10] T. Brants. F. Chen, I. Tsochantarides. Topic-based document segmentation with probabilistic latent semantic analysis. In Proceedings of the 11th International Conference on Information and knowledge Management. 2002. pp.211-218.

[11] D. Beeferman. A. Berger, J. Lafferty. Statistical model for text segmentation. Machine Learning 34(1-3), 1999. pp.177-210.