ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 32,No. 2,Mar. 2017,pp. 408－416
DOI:10. 16337/j. 1004-9037. 2017. 02. 023

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

基于分层狄利克雷过程模型的文本分割

李天彩　王　波　席耀一　张佳明

(解放军信息工程大学信息系统工程学院,郑州,450002)

摘　要:文本分割在文本摘要、信息检索等诸多领域都有重要的应用。主题模型是该领域研究中的重要方法,但目前基于主题模型的方法普遍依赖于主题个数的人工设置。针对此问题,本文提出了一种基于分层狄利克雷过程(Hierarchical Dirichlet process,HDP)模型的文本分割方法。首先使用 HDP 模型获取文本在主题空间的向量表示,然后将主题向量用于 C99 分割算法实现文本分割,最后使用两种优化策略对结果进行优化。实验结果表明,基于 HDP 模型的方法能够摆脱对人工设置主题个数的依赖,有效提高了文本分割的性能。

关键词:主题模型;文本分割;分层狄利克雷过程;CRF 构造

中图分类号:TP391　　**文献标志码**:A

Text Segmentation Based on Hierarchical Dirichlet Processes

Li Tiancai, Wang Bo, Xi Yaoyi, Zhang Jiaming

(Institute of Information and System Engineering, PLA Information Engineering University, Zhengzhou, 450002, China)

Abstract: Text segmentation has important applications in many fields, including text summarization, information retrieval, and so on. Topic model is an important tool in text segmentation. However previous text segmentation methods based on topic model generally rely on manually setting of the number of topics influencing results significantly. To solve the problem, a novel text segmentation method based on hierarchical Dirichlet process(HDP) model is proposed. Firstly, texts are modeled with HDP model to get their expression with topic vectors. Then, the topic vectors are used in C99 segmentation algorithm for text segmentation. Finally, two optimization strategies are applied to result optimization. Experimental results show that the presented method can omit manually setting of the topics numbers and improve the performance of text segmentation.

Key words: topic model; text segmentation; hierarchical Dirichlet process; Chinese restaurant franchise (CRF) process

引　言

文本分割是指按照主题相关的原则将一篇较长的文本分割成语义段落序列,使得各个语义段落内

基金项目:国家高技术研究发展计划("八六三"计划)(2011AA7032030D)资助项目;全军军事研究生课题(2011JY002-158)资助项目。

收稿日期:2014-09-11;**修订日期**:2015-07-10

部具有最大的主题相关性,而语义段落之间具有最小的主题相关性[1]。文本分割在信息检索、文本自动摘要[2]等很多领域都有重要应用[3-5]。在信息检索中,文本分割可以缩小检索范围,提高检索的准确性;在文本自动摘要中,通过文本分割可以得到文本内描述主题不同侧面的语义段落,在此基础上得到的文本摘要不仅与问题紧密相关,而且涵盖了问题所涉及的多个侧面,具有较高的覆盖度。

目前,常用的文本分割方法有基于词汇聚集的方法、基于语言特征的方法和基于主题模型的方法。基于词汇聚集的方法认为描述同一主题的词汇倾向于出现在同一主题片段内。1997 年,Hearst[6] 提出了 TextTiling 方法,该方法先将文本分成固定长度的词汇序列,使用滑动窗口计算相邻序列之间的余弦相似度,通过相似度曲线的变化确定边界的位置。2000 年,Choi[7] 提出了 C99 算法,通过计算文本中句子之间的余弦相似度得到相似度矩阵,设置排序矩阵对相似度矩阵的表示进行优化,通过最大化分割单元内部密度实现分割。2008 年,Eisenstein 等[8] 提出了一种基于贝叶斯框架的文本分割方法。2011 年,Kazantseva 等[9] 将聚类算法(Affinity propagation,AP)用于文本分割,提出了基于相似性传播聚类的文本分割(Affinity propagation for segmentation,APS)方法。基于语言特征的方法认为文本中通常会包含标志主题转移的特征信息。2010 年,Yan 等[10] 在新闻事件识别中使用了多种特征信息用于分割,并指出由于 WordNet 的不完备性和更新速度较慢,将 WordNet 用于分割时对结果的改善非常有限。2012 年,邹箭等[11] 提出了一种针对中文的文本分割模型,根据词典和语料库计算词语的相关度进而计算句子间语义相关度并用于分割。基于词汇聚集的方法假定文本中的词都是孤立的,没有利用词与词之间的关系,因而分割结果的准确度有限;基于语言特征的方法虽然在特定领域的效果较好但是受到知识库的完备性、更新速度以及领域局限性的限制使得其移植性较差,无法适用于各种数据集。因此越来越多的学者探索将能反映文本语义信息的主题模型用于文本分割,并取得了较好的结果。2008 年,石晶等提出了基于概率潜在语义分析(Probabilistic latent semantic analysis,PLSA)模型[12] 和基于潜在狄利克雷分配(Latent Dirichlet allocation,LDA)模型[13] 的文本分割方法,通过实验发现基于 PLSA 模型的方法结果的随机性较大,而基于 LDA 模型的方法分割的准确度有明显提高。2012 年,Ridel 等对基于 LDA 模型的文本分割做了一系列研究[14-16],并将 TextTiling 和 LDA 模型相结合提出了 TopicTiling 方法,通过对 LDA 模型每一次采样得到的主题分布进行统计以确定最终的主题分布,提高了主题模型对文本表示的稳定性,使分割结果得到了显著的改善。2013 年,Du 等[17] 提出了主题分割模型(Topic segmentation model,TSM)方法,该方法通过结构化的主题模型获取文本的表示,并结合逐点边界采样算法进行分割,进一步提高了文本分割的准确度。

基于主题模型的方法虽然能够有效地提高文本分割的准确度,但是目前该类方法普遍依赖于主题个数的人工设置。Riedl 等[15] 指出 LDA 模型的主题个数对文本分割结果的影响很大,主题个数设置得过高会造成过拟合,过低则会造成对文本的描述不够充分,两者都会降低文本分割的准确度,而主题个数又与数据集有关,无法给出通用的经验参数。对于该不足,研究人员目前还没有给出很好的解决方案。在以往研究中通常是直接给出主题个数或者通过遍历参数来寻找最优主题个数,例如 Du 等[17] 为 3 个数据集分别设定了不同的主题个数;Riedl 等[15] 通过在 50～500 的范围内对主题个数进行遍历,确定 Choi 数据集最优主题个数为 100;石晶等[13] 在 20～160 之间对主题个数进行遍历,确定根据《人民日报》合成的文本分割数据集的最优主题个数为 80。由此可见,尽管主题模型能够有效提高文本分割的性能,但是针对不同的数据集,如何确定最优主题个数仍是一个有待解决的问题。

针对 LDA 模型主题个数需要人工设置的不足,Yee 等[18] 提出了分层狄利克雷过程(Hierarchical Dirichlet process,HDP)模型。周建英等[19] 提出,与 LDA 模型相比,HDP 模型能够自动生成主题个数,具有更好的鲁棒性。因此,本文将 HDP 模型用于文本分割,提出了一种基于 HDP 模型的文本分割方法。首先使用 HDP 模型获取文本在主题空间的向量表示;然后将主题向量用于 C99 分割算法实现文本分割;最后使用优化策略对结果进行优化。实验结果证明了本文方法的有效性和优越性。

1 HDP 模型

1.1 HDP 模型基本原理

HDP 模型是一种非参数贝叶斯模型，来源于狄利克雷过程(Dirichlet process, DP)，是狄利克雷过程混合模型的多层形式。该模型不仅能实现聚类和推断等功能，而且能够自动生成主题个数。其有向图模型如图 1 所示，对于一个包含 M 篇文本的集合，每篇文本的主题都来源于基分布 H，这保证了各篇文本之间可以实现对无限多个主题的共享。在实现过程中，首先从 H 中获取该文本集合的总体基分布 $G_0 \sim DP(\gamma, H)$，γ 为聚集参数；然后以 G_0 为基分布，以 α_0 为聚集参数，获取每一篇文本的主题分布 $G_j \sim DP(\alpha_0, G_0)$，$j=1,2,\cdots,M$；最后以该层狄利克雷过程为先验分布，构造狄利克雷过程混合模型(1)，即

$$\theta_{ji} \mid G_j \sim G_j, x_{ji} \mid \theta_{ji} \sim F(\theta_{ji}) \tag{1}$$

式中：函数 $F(\theta_{ji})$ 表示给定参数 θ_{ji} 时，观测变量 x_{ji} 的分布；参数 θ_{ji} 条件独立服从分布 G_j，而观测变量 x_{ji} 条件独立服从分布 $F(\theta_{ji})$，x_{ji} 代表第 j 篇文本的第 i 个词。

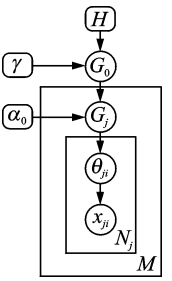

图 1 HDP 有向图模型

Fig. 1 Directed graph of HDP

1.2 CRF 构造

HDP 模型存在多种构造方式[18]，本文中使用的是基于中国餐馆连锁(Chinese restaurant franchise, CRF)构造的 HDP 模型。CRF 构造由中国餐馆过程(Chinese restaurant process, CRP)扩展而来。CRF 构造中假设一个中国餐馆代表一篇文本，餐馆里的一位顾客代表文本里的一个词，一道菜代表一个潜在主题，餐馆里的一张桌子代表共享同一道菜的顾客的群组，并且所有餐馆共享一个包含无穷多道菜的菜单，每一个中国餐馆里有无穷多张桌子，每张桌子能容纳下无穷多个顾客，进入餐馆 j 中的第 i 个顾客用 x_{ji} 表示。在 CRF 构造中，首先为每一位顾客分配餐桌，而后再为餐桌分配菜。顾客 x_{ji} 选择某一桌子就座的概率分布为

$$p(t_{ji}=t \mid \boldsymbol{t}^{-ji}, \boldsymbol{k}) \propto \begin{cases} n_{jt}^{-ji\cdot} f_{k_{jt}}^{-x_{ji}}(x_{ji}), t \text{ 被使用} \\ \alpha_0 p(x_{ji} \mid \boldsymbol{t}^{-ji}, t_{ji}=t^{\text{new}}, k), t=t^{\text{new}} \end{cases} \tag{2}$$

t^{new} 表示顾客 x_{ji} 选择一张新桌子坐下，此时有

$$p(x_{ji} \mid \boldsymbol{t}^{-ji}, t_{ji}=t^{\text{new}}, \boldsymbol{k}) = \sum_{k=1}^{K} \frac{m_{\cdot k}}{\gamma + m_{\cdot\cdot}} f_k^{-x_{ji}}(x_{ji}) + \frac{\gamma}{\gamma + m_{\cdot\cdot}} f_{k^{\text{new}}}^{-x_{ji}}(x_{ji}) \tag{3}$$

每一桌子所选择菜的概率分布为

$$p(k_{jt}=k \mid \boldsymbol{t}, \boldsymbol{k}^{-jt}) \propto \begin{cases} m_{\cdot k}^{-jt} f_k^{-x_{ji}}(\boldsymbol{x}_{jt}), k \text{ 被使用} \\ \gamma f_{k^{\text{new}}}^{-x_{ji}}(\boldsymbol{x}_{jt}), k=k^{\text{new}} \end{cases} \tag{4}$$

式中：观测数据 x_{ji} 服从分布 $F(\theta_{ji})$，$\boldsymbol{x}=(x_{ji}:\text{all } j, i)$，$\boldsymbol{x}_{jt}=(x_{ji}:\text{all } i \text{ with } t_{ji}=t)$，$\boldsymbol{t}=(t_{ji}:\text{all } j,i)$，$\boldsymbol{k}=(k_{jt}:\text{all } j,t)$，$n_{jtk}$ 表示餐馆 j 中坐在桌子 t 并选择菜 k 的顾客数，m_{jk} 表示餐馆 j 中选择菜 k 的桌子数，表示所有餐馆中不同菜的数目，$f_k^{-x_{ji}}(x_{ji})$ 表示观测数据的条件分布。上标表示该变量不被计算在内，例如 $x^{-ji}=\boldsymbol{x}/x_{ji}$，$\boldsymbol{k}^{-jt}=\boldsymbol{k}/k_{jt}$。$n_{jt}^{-ji\cdot}$ 表示第 j 个餐馆中的第 t 张桌子上的顾客数(不包括第 i 个顾客)，$m_{\cdot k}^{-jt}$ 表示所有餐馆中选择菜 k 的桌子数目(不包括第 t 张桌子)，其余符号所代表的物理意义依此类推。顾客 x_{ji} 进入餐馆，选择了桌子 t 坐下，并与同桌的其他顾客享用了菜 k 的过程即代表了待分割文本中的一个词 w_{ji} 被分配到主题 IDk 的过程[18]。

2 基于 HDP 模型的文本分割方法

2.1 利用主题向量表示文本

HDP 模型在每次迭代过程中都会通过采样为每个词分配主题 ID，因此可以使用主题 ID 代替词表

示文本。下面通过一个利用主题 ID 表示文本的例子对这一过程进行介绍。

244 1943 年 春节(34) 鲁艺(34) 闹(34) 秧歌(34) 运动(34) 出现(34) 高潮(34)

245 延安(34) 鲁艺(34) 文工团(34)演出(34) 小型(24) 秧歌剧(34) 兄妹(24) 开荒(34) 最为(24) 群众(34) 喜爱(34) 是 新(24) 秧歌(34) 运动(34) 初期(24) 代表(34) 作(24)

247 来自(0) 两(37) 方面(37)问题(0)说明(37)制度(0)还 不能(37) 保持(0) 国家(0) 企业(0) 分配(0) 关系(0) 稳定(0) 不 利于(37) 依法(0) 保护(0) 企业(0) 合理(0)收益(0) 难以(0) 让 企业(0) 形成(0) 真正(0) 压力(0) 动力(0) 更 不 利于(37) 国家(0) 有效(37) 集中(37) 资金(0) 保证(0) 重点(0) 建设(37)

上段文字包含三句话,除了停用词没有被标记主题 ID 之外,每个词都被分配了一个主题 ID。通过观察主题 ID 的分布,可以发现句子 244 和 245 是属于同一主题的,内容主要是关于延安时期的文化活动,这两句中词的主题 ID 大多被标记为 34;句子 247 主要是关于制度与经济建设,其中大部分词的主题 ID 被标记为 0。通过以上分析发现,使用主题 ID 代替词对文本进行表示并在此基础上进行文本分割是可行的。

传统的文本分割算法使用词汇作为特征,文本被看成由特征权重组成的向量,可表示为

$$\boldsymbol{s}=(w_1,w_2,\cdots,w_V) \tag{5}$$

式中:w_i 表示词汇 i 在文本中的权重,一般取词频(Term frequency, TF)或者词频-逆文档频率(Term frequency-inverse document frequency,TF-IDF),V 表示词汇个数。基于词向量的文本相似度计算方法无法度量特征之间的潜在语义关系,如句子 244 和 245 中提到的“秧歌”、“文工团”和“演出”之间存在着潜在的语义关系,但是在向量空间模型中这种相关性无法得到体现,而在主题模型下通过相同的主题 ID 可体现出。针对这个问题,本文在文本分割中使用主题向量代替词向量对文本进行表示。主题向量的形式可表示为

$$\boldsymbol{s}'=(t_1,t_2,\cdots,t_K) \tag{6}$$

式中:t_k 为主题 ID_k 在文本中出现的频率;K 为 HDP 模型自动聚类产生的主题个数。

2.2 C99HDP 实现原理

C99[7]是经典的基于词汇聚集的文本分割方法,该方法中以句子为基本单位,通过句子向量空间模型中表示计算向量的余弦相似度得到句子间的相似度。由 2.1 可知这种方法无法度量特征之间潜在的语义关系,因此本文提出使用 HDP 模型得到的主题向量代替词向量作为句子的表示,并将该方法记为 C99HDP。C99HDP 方法中相似度矩阵 $\boldsymbol{S}$ 中每一个元素 s_{ij} 可通过计算句子 i 和 j 的主题向量的余弦相似度得到,即

$$s_{ij}=\frac{\boldsymbol{s}_i'\cdot\boldsymbol{s}_j'}{|\boldsymbol{s}_i'|\times|\boldsymbol{s}_j'|} \tag{7}$$

式中:$\boldsymbol{s}_i'$表示句子 i 的主题向量。在得到 $\boldsymbol{S}$ 矩阵之后,以 $\boldsymbol{S}$ 矩阵为基础构造排序矩阵 $\boldsymbol{R}$。$\boldsymbol{R}$ 矩阵中的元素 r_{ij} 是通过对 $\boldsymbol{S}$ 矩阵中元素 s_{ij} 附近 $N\times N$ 的范围内值小于 s_{ij} 的元素的数量进行统计得到,文献[7]中 N 设定为 11。在得到排序矩阵 $\boldsymbol{R}$ 后,定义了分割单元的内部密度 D 的计算公式(8),在每次分割过程中通过最大化 D 来确定分割的位置 b_1 和 b_2,b_1 和 b_2 分别是分割单元 g 两端的句子编号($b_2>b_1$)。

$$D=\frac{\sum_{g=1}^{m}\mathrm{sum}(g)}{\sum_{g=1}^{m}a_g} \tag{8}$$

式中:m 表示分割单元总数,$\mathrm{sum}(g)$表示分割单元 g 内的 $\boldsymbol{R}$ 矩阵元素和,a_g 表示分割单元 g 的大小,$a_g=(b_2-b_1+1)$。在不断进行分割的过程中密度差 $\Delta(D^{(n)})=D^{(n)}-D^{(n-1)}$ 会逐渐减小,其中 $D^{(n)}$ 表示得到 n 个分割单元时的内部密度。当 $\Delta(D^{(n)})$ 达到分割停止的阈值 $\mu+c\times\sigma$ 时停止聚类,即文本分割完成,其中 μ 和 σ 分别为 $\Delta(D^{(n)})$ 的均值和方差,c 为常数,本文参考文献[7],将 c 设定为 1.2。

2.3 分割结果优化

Ridel 等[15]指出，仅使用 LDA 模型最后一次采样的结果作为文档的表示具有一定的不稳定性，而将 LDA 模型 Gibbs 每次采样的结果进行统计，对每个词使用所有结果中出现频率最高的主题 ID 作为其最终的主题 ID 可以有效提高文本表示的稳定性。与 LDA 模型不同，HDP 模型每次得到的主题个数是不固定的，因而无法通过对单个词汇的主题 ID 进行统计确定其最终的主题 ID。本文结合 HDP 模型的特点提出了两种对分割结果进行优化的策略：(1)在一次 HDP 模型建模的过程中，对不同迭代次数得到的采样结果进行文本分割，对得到的分割结果进行统计，选择出现概率大于 t_1 的分割位置作为最终的分割位置；(2)进行多次 HDP 模型建模，对每次建模得到的最终采样结果进行文本分割，对得到的分割结果进行统计，选择出现概率大于 t_2 的分割位置作为最终的分割位置。在不同数据集上进行的实验中发现，t_1 和 t_2 取 0.5 左右时效果最好。这是因为 t_1 和 t_2 取得过小会添加多余的分割，而过大则会遗漏一些正确的分割，两者都会造成最终分割结果的准确度下降。

3 实验设计与分析

3.1 实验数据集

由于针对中文的文本分割还没有公测的数据集，本文参考石晶等[13]和邹箭等[11]的实验设计，使用了两个数据集进行实验。

(1)PRF 数据集。在 1998 年 1 月份《人民日报》的基础上按照 Choi[7] 数据集格式构建的文本分割数据集，其中包括 4 个测试集 $T_{3\text{-}11}$，$T_{3\text{-}5}$，$T_{6\text{-}8}$，$T_{9\text{-}11}$，$T_{x\text{-}y}$ 表示所含主题片段的句数在 x 和 y 之间，具体构造方式是每次从 3 147 篇《人民日报》报道里随机选取 10 篇不同的文本，从每篇文本内提取 3～11 个句子，形成一个段落，将 10 个段落连接起来，组成一个新文本。由于每个片段来自不同的文本，讨论不同主题，因此其边界自然成为新文本中的主题边界。本文将该数据集记为 PRF 数据集，具体信息如表 1 所示。

表 1 PRF 数据集

Tab. 1 PRF dataset

测试集	$T_{3\text{-}11}$	$T_{3\text{-}5}$	$T_{6\text{-}8}$	$T_{9\text{-}11}$
段落长度	3-11	3-5	6-8	9-11
文本数量	400	100	100	100

(2)CRA 数据集。2012 年邹箭在文献[11]中选用 800 万字国家汉语语料库中的部分文本，通过人工标记主题边界构成的数据集。本文将该数据集记为 CRA，其中包括 3 个测试集，具体信息如表 2 所示。

表 2 CRA 数据集

Tab. 2 CRA dataset

编号	文本个数	主题个数	句子数	分割单元平均长度	文本平均主题个数
1	7	66	1 982	30	10
2	11	83	3 236	39	8
3	6	45	1 886	42	8

3.2 评价指标

为了便于同类算法进行对比，本文使用了常用的文本分割错误率 P_k[20] 和 WD(WindowDiff[21])作为评价指标。

P_k 可定义为

$$P_k = P_{seg} P_{miss} + (1 - P_{seg}) P_{false\ alarm} \tag{9}$$

式中:P_{seg}是距离为 k 的两个句子分属不同语义段落的概率;P_{miss}是算法分割结果缺少一个段落的概率;$P_{false\ alarm}$是算法分割结果添加一个段落的概率;k 取真实分割中段落平均长度的一半。本实验中按照石晶在文献[11,12]中的设置,取 $P_{seg}=0.5$。

WindowDiff 可定义为

$$\mathrm{WindowDiff}(ref, hyp) = \frac{1}{N-k} \sum_{i=1}^{N-k} (\mid b(ref_i, ref_{i+k}) - b(hyp_i, hyp_{i+k}) \mid > 0) \tag{10}$$

式中:ref 代表真实分割;hyp 代表算法分割;$b(i,j)$表示整句 si 和整句 sj 间的边界数量;N 表示文本中的整句数量;k 取真实分割中段落平均长度的一半。

3.3 实验设置与结果分析

两个数据集中的文本都已经进行过分词,在实验预处理过程中只进行了停用词的去除。为了与相关学者的研究进行比较,本文选取常用的 5 种方法进行对比实验,分别是 Hearst 提出的 TextTiling 方法[6]、Choi 提出的 C99 方法[7]、Eisenstein 提出的 BayesSeg 方法[8]、邹箭提出的 CRA 方法[11]和 Du 提出的 TSM 方法[17]。以上对比方法结果分别记为 TT,C99,BayesSeg,CRA 和 TSM。两个数据集上的 C99,BayesSeg 和 TSM 方法使用文献[16]中实验的参数设置,PRF 数据集上的 TT 方法使用了文献[7]中实验的参数设置。此外,为了验证 TSM 对主题个数的依赖性,在实验中除了使用文献[17]中的 10,25,50 之外,还增加了主题个数为 80 的实验。本文所提出方法的结果记为 C99HDP,实验中 HDP 模型的迭代次数为11 000,其他参数使用 Yee 等在文献[18]中的设置。为了验证两个优化策略的性能,分别设置 C99HDP-1 和 C99HDP-2 的两组实验,其中 C99HDP-1 对应优化策略 1,在 HDP 模型建模过程中每1 000次迭代保留一次采样结果并进行文本分割,对所有分割结果进行统计得到最终的分割结果并进行评分;C99HDP-2 对应优化策略 2,对每篇文本进行 11 次 HDP 模型建模,并对每次建模得到的最终采样结果进行文本分割,对所有分割结果进行统计得到最终的分割结果并进行评分。HDP 模型需要训练数据挖掘词汇之间的语义关系,为了避免从外部引入训练数据的影响,本文参考文献[15]中的实验设置,在 PRF 数据集上采用 10 折交叉验证的方法,每次使用 90%的数据用于训练,10%的数据用于测试,直到遍历所有测试数据。由于 CRA 数据集中文本较少,采用 10 折交叉验证的方法会存在训练数据过少的问题,所以每次只将 1 篇文本用于测试,其余用于训练,直到遍历所有测试文本。

3.3.1 在 PRF 数据集上的对比实验

在 PRF 数据集上进行实验,其结果如表 3 所示。由表 3 可以看出,C99HDP,C99HDP-1 和 C99HDP-2 的分割结果明显优于经典的 C99,TT 以及 BayesSeg。

对比 C99 与 C99HDP 的结果可以发现,使用主题 ID 代替词汇表示文本用于文本分割,可以提高文本分割的性能。从 C99HDP 与 C99HDP-1 和 C99HDP-2 的对比可以看出,使用了优化策略的结果要好于不使用时的结果。这主要是因为只使用最后一次采样结果对文本进行表示,会导致由于一些错误的主题 ID 标注进而造成错误的分割,而使用了多次采样结果进行分割并对结果进行统计可以过滤掉一些单次分割中出现的错误分割,进而提高分割的准确度。通过对 C99HDP-1 和 C99HDP-2 的结果进行对比发现,C99HDP-2 的结果要优于 C99HDP-1,这主要是因为 C99HDP-2 中使用每次建模的最终采样结果作为文本的表示,比迭代次数较少时的表示更加准确,因而在最终的文本分割准确度上会更高一些。从 TSM 与 C99HDP,C99HDP-1 和 C99HDP-2 的结果对比可以看出,TSM 在主题个数设置为 80 和 50 的时候性能表现较好,尤其是在主题个数设置为 80 时,TSM 在 $T_{3\text{-}5}$,$T_{6\text{-}8}$ 以及平均结果上都优于 C99HDP,C99HDP-1 和 C99HDP-2,因为 80 接近于 PRF 数据集的最优主题个数,在此情况下,TSM 能更好地发挥其模型的性能,而且 TSM 中整合的逐点边界采样算法也有助于降低其分割的错误率。从

另一方面来看，当 TSM 的主题个数 k 设置为 50，25 和 10 时其分割的错误率不断升高，这说明当 TSM 中主题个数的设置不断偏离最优主题个数的时候 TSM 的结果会不断恶化，错误率甚至会高于 C99HDP，C99HDP-1 和 C99HDP-2 的结果，这也说明了该类方法不适用于对主题个数未知的数据进行处理。从不同测试集的结果来看，各个方法在 $T_{3\text{-}5}$ 测试集上的错误率都较高，这说明对较短的分割单元进行分割的难度较大，有待进一步改进。

表 3 PRF 数据集上的实验结果

Tab. 3 Experimental results on PRF dataset

Experiment	$T_{3\text{-}11}$		$T_{3\text{-}5}$		$T_{6\text{-}8}$		$T_{9\text{-}11}$		Ave	
	P_k	WD	P_k	WD	P_k	WD	P_k	WD	P_k	WD
TT	0.268 1	0.281 1	0.283 1	0.292 7	0.151 6	0.172 8	0.191 2	0.233 3	0.242 6	0.2605
C99	0.271 1	0.283 5	0.300 4	0.308 9	0.227 2	0.258 6	0.208 4	0.269 9	0.260 1	0.281 6
BayesSeg	0.226 0	0.383 3	0.241 3	0.405 5	0.209 1	0.375 3	0.154 3	0.306 7	0.215 5	0.374 4
TSM(k=80)	0.047 0	0.067 0	**0.051 0**	**0.072 0**	**0.033 0**	**0.047 0**	0.052 0	0.060 0	**0.046 3**	**0.063 9**
TSM(k=50)	0.057 8	0.079 7	0.068 9	0.087 4	0.057 9	0.081 3	**0.038 4**	**0.048 0**	0.056 6	0.076 5
TSM(k=25)	0.083 5	0.121 7	0.100 7	0.143 0	0.063 2	0.090 7	0.061 6	0.077 0	0.079 9	0.113 9
TSM(k=10)	0.138 5	0.196 1	0.161 5	0.216 2	0.097 3	0.137 9	0.111 1	0.137 4	0.132 0	0.182 3
C99HDP	0.062 2	0.066 9	0.148 1	0.149 8	0.078 1	0.080 5	0.060 1	0.079	0.076 4	0.082 4
C99HDP-1	0.059 9	0.064 7	0.144 3	0.146 1	0.059 9	0.072 7	0.059 3	0.078 3	0.071 9	0.079 4
C99HDP-2	**0.043 5**	**0.052 4**	0.101 2	0.110 8	0.067 7	0.078	0.050 8	0.056 9	0.056 2	0.065 0

3.3.2 在 CRA 数据集上的对比实验

由于 PRF 数据集中的单篇文本长度较小，为了增加实验的全面性和可靠性，本文设置了 CRA 数据集上的实验，实验结果如表 4 所示。从表 4 中可以看出，在 CRA 数据集上进行的实验中，C99HDP-1 和 C99HDP-2 在平均错误率上都低于其他方法，这说明本文提出的方法在处理不同数据集时性能稳定，对新数据的鲁棒性较好。此外，TSM 在 CRA 数据集上的最优结果不是出现在 k=80 的实验组，而且 k=80 实验组的 P_k 错误率高于 BayesSeg，TT，CRA，C99HDP，C99HDP-1 和 C99HDP-2，这说明对于不同的数据集主题模型的最优主题数是不同的，依赖于主题个数设定的方法在处理新数据时会因为主题个数

表 4 CRA 数据集的实验结果

Tab. 4 Experimental results on CRA dataset

Experiment	1		2		3		Ave	
	P_k	WD	P_k	WD	P_k	WD	P_k	WD
CRA	0.116 0		0.146 0		0.102 0		0.126 3	
TT	0.142 0		0.153 0		0.124 0		0.142 5	
C99	0.159 1	0.189 3	0.134 6	0.194 4	0.200 2	0.307 1	0.158 1	0.221 1
BayesSeg	0.144 0	0.163 9	0.147 2	0.165 7	0.125 7	0.192 5	0.140 9	0.171 9
TSM(k=80)	0.117 7	0.159 5	0.151 4	0.187 3	0.167 6	0.204 2	0.145 6	0.183 4
TSM(k=50)	0.118 3	0.151 6	0.101 2	0.132 7	0.217 0	0.268 8	0.135 1	0.172 2
TSM(k=25)	0.124 5	0.196 9	0.113 4	0.130 8	0.124 0	0.136 9	0.119 3	0.151 6
TSM(k=10)	0.127 4	0.185 7	0.105 7	0.115 0	0.155 0	0.176 1	0.124 4	0.150 9
C99HDP	0.113 8	0.158 7	0.141 4	0.180 5	0.137 0	0.187 6	0.132 3	0.175 9
C99HDP-1	0.070 1	0.088 5	0.105 1	0.122 8	0.098 3	0.140 4	0.093 2	0.117 2
C99HDP-2	**0.043 1**	**0.078 5**	**0.048 4**	**0.081 3**	**0.060 7**	**0.099 4**	**0.049 9**	**0.085 0**

不合适而影响结果。对比表 3 和表 4 可以发现,基于主题模型的方法在 CRA 数据集上的表现性能都低于在 PRF 数据集上的表现,这是因为对于较长的分割单元,通常存在多个局部主题。例如,在一个关于天文学的段落中,其中靠前的部分主要是关于中微子理论的介绍,靠后的部分主要是关于天文观测设备发展的介绍,在 HDP 模型下对这两部分进行标记,靠前的部分出现较多的主题 ID 是 21,靠后的部分出现较多的主题 ID 是 19,在使用分割算法进行分割的过程中这个语义段落被错分成了两个语义段落。这种情况出现主要有两方面的原因:(1)训练数据集中关于天文知识的数据过少,对该领域相关词汇的关联挖掘不够充分;(2)对于这种因为颗粒度大小造成的分割不一致,即使是人工评判也有可能会出现,这与评判的标准和实际需求有关。

3.3.3 参数 N 的遍历

为了验证 C99 中生成 $\boldsymbol{R}$ 矩阵时使用参数 N 的大小对最终分割结果的影响,同时对 C99,C99HDP,C99HDP-1 和 C99HDP-2 的结果进行比较,在 CRA 数据集上进行实验,设置 N 在[3,101]范围内,以步长 2 变化。实验结果如图 2 所示,其中横轴为 N,纵轴为平均的 P_k 值。由图 2 可以看出,当 $N>9$ 时,C99HDP,C99HDP-1 和 C99HDP-2 的错误率均低于经典的 C99 算法,这说明使用 HDP 模型对文本进行表示可以有效降低文本分割的错误率并且总体性能稳定;C99HDP-1 和 C99HDP-2 的错误率均低于 C99HDP,这说明了两种优化策略的有效性;C99HDP-2 的错误率均低于 C99HDP-1,这说明 HDP 模型中的迭代次数较多时对文本进行表示的准确度要高于迭代次数较少时,进而对分割结果的优化更为显著。

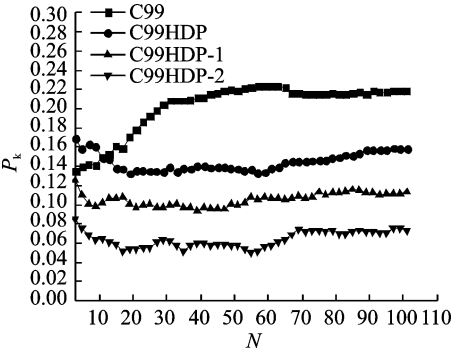

图 2 不同方法在不同 N 下的平均 P_k 对比结果

Fig. 2 Comparison of average P_k with different methods under different N

4 结束语

本文针对目前基于主题模型的文本分割方法无法自动确定主题个数的不足,提出了一种基于 HDP 模型的文本分割方法。首先利用 HDP 模型对待分割文本建模,获取文本在主题模型下的表示;然后将主题向量用于 C99 分割算法实现文本分割;最后提出了两种优化策略提高了分割结果的准确性。在两个数据集上实验的结果表明本文方法不仅能够降低文本分割的错误率,而且对新数据集有很好的鲁棒性。需要指出的是,本文提出的方法还有待进一步改进,如在较短段落的分割中该方法错误率较高,需要针对短段落的特点对其进行改进。除此之外,如何改进边界检测算法也是下一步工作中重要的研究内容。

参考文献:

[1] Wu J W, Tseng J C R, Tsai W N. A hybrid linear text segmentation algorithm using hierarchical agglomerative clustering and discrete particle swarm optimization[J]. Integrated Computer-Aided Engineering, 2014, 21(1): 35-46.

[2] 赵斌,吉根林,徐伟,等. 基于拓扑结构的微博话题摘要生成算法[J]. 数据采集与处理,2014,29(5):720-729.
Zhao Bin, Ji Genlin, Xu Wei, et al. Microblog topic summarization based on topology structures[J] Journal of Data Acquisition and Processing, 2014,29(5):720-729.

[3] Shafiq J, Giuseppe C, Gabriel M, et al. Exploiting conversation structure in unsupervised topic segmentation for emails [C]//Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. Massachusetts: Association for Computational Linguistics, 2010:388-398.

[4] 丁长林, 蔡东风, 王裴岩. 基于分类算法的专利摘要文本分割技术[J]. 山东大学学报,2012, 47(5):69-72.
Ding Changlin, Cai Dongfeng, Wang Peiyan. Text segmentation of patent summary based on a classification algorithm[J]. Journal of Shandong University, 2012, 47(5):69-72.

[5] Paula C F C, Maite T, Thiago A S P. Subtopics annotation in a corpus of news texts: Steps towards automatic subtopic segmentation[C]// Proceedings of the 7th Brazilian Symposium in Information and Human Language Technology. Fortaleza: Brazilian Computer Society, 2013:108-118.

[6] Hearst M A. TextTiling: Segmenting text into multi-paragraph subtopic passages[J]. Computational Linguistics, 1997, 23(1): 33-64.

[7] Choi F Y Y. Advances in domain independent linear text segmentation[C]// Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference. Denver: North American Chapter of the Association for Computational Linguistics, 2000: 26-33.

[8] Jacob E, Regina B. Bayesian unsupervised topic segmentation[C]//Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing. Honolulu: Association for Computational Linguistics, 2008: 334-343.

[9] Kazantseva A, Szpakowicz S. Linear text segmentation using affinity propagation[C]// Proceedings of the Conference on Empirical Methods in Natural Language Processing. Edinburgh: Association for Computational Linguistics, 2011: 284-293.

[10] Yan Rui, Li Yu, Zhang Yan, et al. Event recognition from news webpages through latent ingredients extraction[C]// The 6th Asia Information Retrieval Societies Conference. Tapei, China: Springer, 2010:490-501.

[11] 邹箭, 钟茂生, 孟荔. 中文文本分割模式获取及其优化方法[J]. 南昌大学学报, 2012, 35(6): 597-601.
Zou Jian, Zhong Maosheng, Meng Li. Method of Chinese text segmentation model acquisition and its optimization[J]. Journal of Nanchang University, 2012, 35(6): 597-601.

[12] 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.
Shi Jing, Dai Guozhong. Text segmentation based on PLSA model[J]. Journal of Computer Research and Development, 2007, 44(2): 242-248.

[13] 石晶, 胡明, 石鑫. 基于 LDA 模型的文本分割[J]. 计算机学报, 2008, 31(10): 1865-1873.
Shi Jing, Hu Ming, Shi Xin. Text segmentation based on model LDA[J]. Chinese Journal of Computer, 2008, 31(10): 1865-1873.

[14] Riedl M, Chris B. How text segmentation algorithms gain from topic models[C]// Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Denver: Association for Computational Linguistics, 2012: 553-557.

[15] Riedl M, Chris B. Text segmentation with topic models[J]. Journal for Language Technology and Computational Linguistics, 2012,27(1): 47-69.

[16] Riedl M, Biemann C. Topictiling: A text segmentation algorithm based on lda[C]// Proceedings of ACL 2012 Student Research Workshop. Jeju Island: Association for Computational Linguistics, 2012: 37-42.

[17] Du L, Buntine W, Johnson M. Topic segmentation with a structured topic model[J]. Stroudsburg Pa the Association for Computational Linguistics, 2013: 190-200.

[18] Yee W T, Michael I J, Matthew J B, et al. Hierarchical Dirichlet processes[J]. Journal of the American Statistical Association, 2006, 101(476): 1566-1581.

[19] 周建英, 王飞跃, 曾大军. 分层 Dirichlet 过程及其应用综述[J]. 自动化学报, 2011, 37(4): 389-407.
Zhou Jianying, Wang Feiyue, Zeng Dajun. Hierarchical dirichlet processes and their applications: A survey[J]. Acta Sutomatica Sinica, 2011, 37(4): 389-407.

[20] Beeferman D, Berger A, Lafferty J. Statistical models for text segmentation[J]. Machine Learning, 1999, 34(1): 177-210.

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

作者简介:

李天彩(1990-),男,硕士研究生,研究方向:文本分割与会话抽取,E-mail:litc1125@126.com。

王波(1970-),男,副教授,研究方向:网络协议分析与智能信息处理,E-mail:yyywb@163.com。

席耀一(1987-),男,博士研究生,研究方向:话题检测与追踪,E-mail:120479063@qq.com。

张佳明(1989-),男,硕士研究生,研究方向:情感分析,E-mail:550163421@qq.com。