文章编号：1003-0077(2009)06-0039-07

一种基于 LDA 的 CRF 自动文摘方法

吴晓锋，宗成庆

（中国科学院 自动化研究所 模式识别国家重点实验室，北京 100190）

摘 要：浅层狄利赫雷分配(Latent Dirichlet Allocation，LDA)方法近年来被广泛应用于文本聚类、分类、段落切分等等，并且也有人将其应用于基于提问的无监督的多文档自动摘要。该方法被认为能较好地对文本进行浅层语义建模。该文在前人工作基础上提出了基于 LDA 的条件随机场(Conditional Random Field，CRF)自动文摘(LCAS)方法，研究了 LDA 在有监督的单文档自动文摘中的作用，提出了将 LDA 提取的主题(Topic)作为特征加入 CRF 模型中进行训练的方法，并分析研究了在不同 Topic 下 LDA 对摘要结果的影响。实验结果表明，加入 LDA 特征后，能够有效地提高以传统特征为输入的 CRF 文摘系统的质量。

关键词：计算机应用；中文信息处理；自然语言处理；自动文摘；狄利赫雷分布；条件随机场

中图分类号：TP391 文献标识码：A

An Approach to Automatic Summarization by Integrating Latent Dirichlet Allocation in Conditional Random Field

WU Xiaofeng, ZONG Chengqing

(National Lab of Pattern Recognition, Institute of Automation,
Chinese Academy of Sciences, Beijing 100190, China)

Abstract: In recent years, Latent Dirichlet Allocation (LDA) has been widely applied in the document clustering, the text classification, the text segmentation, and even the query based multi-document summarization without supervision. LDA is recognized for its great power in modeling a document in a semantic way. In this paper we propose a new superivised method for the extraction-based single document summarization by adding LDA of the document as new features into a CRF summarization system. We study the power of LDA and analyze its different effects by changing the number of topics. Our experiments show that, by adding LDA features, the result of traditional CRF summarization system can be impressively increased.

Key words: computer application; Chinese information processing; natural language processing; automatic document summarization; latent Dirichlet allocation; conditional random field

1 引言

作为信息处理的一个重要分支，自动摘要自 Luhn[1] 提出以来，尤其近年来在许多国际评测，如文本理解会议(Document Understanding Conference，DUC)等，公布了可用于统一训练、测试的语料之后，发展较为迅速。很明显，在信息爆炸的今天，文本自动摘要技术的成熟必将为互联网用户提供极大的便利，而且对于语音等其他媒体文档，摘要技术也不乏其用武之地。

自动摘要方法的分类林林总总，大致上包括：1)按摘要的功用，可以分为指示型的、信息型的和评估型的；2)按输入文档的相关性，可以分为单文档的和多文档摘要；3)按应用，可以分为一般型的和基于查询的摘要[2]。本文所要采用的分类方法，也是目前来说使用最为广泛的分类方法，是基于摘要的产生方式来划分的，即抽取型摘要和生成型摘要。

收稿日期：2008-10-11 定稿日期：2009-03-24

作者简介：吴晓锋(1976—)，男，博士生，主要研究方向为自动文摘生成方法研究；宗成庆(1963—)，男，研究员，博导，主要研究方向为机器翻译。

一般来说,生成型摘要的特点是采用基于实体信息、信息融合及压缩的方法,文献[3]简明地指出,所有非抽取型的摘要都可以归类为生成型的摘要。生成型方法对自然语言处理的相关技术要求较高,而鉴于目前这些技术的发展还大都处于雏形阶段,所以该方法生成的摘要很难付诸实用。而抽取型摘要是提取文本中现成句子,虽然有些呆板,但因为简单易行,现在不但仍是理论研究的主要方向[4-6],而且也出现了一些实用型的文摘系统,例如,MEAD文摘系统(http://www.summarization.com/mead/)。本文将主要讨论有监督的抽取型摘要方法。

1.1 有监督的抽取型摘要发展

抽取型摘要的主要机制是给句子打分。早期的研究者一般根据词频及其分布特点、一些特殊的称谓[1],还有句子在段落中的位置[7]以及句子的相似度[8]等特征来确定一个句子的分值。这种方法的效果并不算差。而当前大多抽取方法是寻找合适的机器学习算法来更加有效地利用这些特征,或寻找更适合某些算法的新特征。

贝叶斯分类器和朴素贝叶斯模型是早期的应用[9-10],作者称其语料训练出的特征权重和文献[11]具有很好的一致性,而文献[11]中采用的是对所有的特征进行综合客观评分。这些分类器把每个句子单独对待,忽视了句子之间的有机联系。文献[5]采用了遗传算法来计算每个句子属于摘要与否的分值或置信度,但是这种方法也有上述缺点。

文献[12]首先用序列标注模型来尝试解决这个问题,作者使用了有较少的独立性假设的隐马模型(HMM)。然而,HMM对于表征描述句子间关系的特征能力有限。

近几年来,条件随机场(Conditional Random Field, CRF)[13]被证明是一种成功的序列标注模型。在词性标注、命名实体识别、语块分析中都得到了很好的应用。文献[14]尝试着用CRF来做文本摘要。CRF不但可以较好地使用上述所有以句子为单位的特征,而且还可以使用相关性较强的句子间特征。另外,还可以融合其他摘要系统的输出。文献[14]在文中除了引入了适当的特征,还对支持向量机(SVM)、HMM以非监督等其他系统的输出结果做了比较,其结果证明,CRF的结果存在较大的优势。

1.2 对文章的各种建模

无论是有监督还是无监督,词频特征及词的分布规律特征都是非常重要的。鉴于词汇集规模一般非常巨大,如何能有效的避免稀疏,如何能有效的对文本进行建模,并从模型中发掘潜在的词汇分布规律一直是学者们关心的问题。而进行降维一直是这个问题的主导解决思路。

通常使用的tf-idf方法[15],是单纯地把基本词汇的频度与该词的稀疏度相结合作为该词的分值,它的降维贡献在于把任意长度的整个语料中的数据规模减小到定长的词汇集的级别。这种缩减虽然非常可观,但即使是词汇集的级别仍是非常巨大的,并且文本内和文本间的词分布关系在这个降维中也没有反映出来。浅层语义索引(Latent Semantic Indexing, LSI)[16]是从事信息检索(IR)研究的学者们进行的一个有意义的尝试,LSI对tf-idf的矩阵进行奇异值分解,由此构筑的线性空间对比原空间的维度大大下降,并且被认为能捕捉到一些诸如同义词和多义词等基本的语言学特征。LSI方法认为,每一篇文本都是一个主题的产生物,反映了这个主题(topic)。显然这个假设有点过于硬性。概率浅层语义索引(pLSI)[17]通过引入概率放松了这个限制。它认为文本中可以有多个topic,每个词从topic中产生出来,而用topic的分布来表征这篇文本。这样表征文本的维数就从词汇集的量级降到了topic数的量级,这是一个重大的进步。pLSI的缺陷主要在于,它不是个完备定义的文本生成模型,它的参数只和训练文本有关,理论上不能将其直接用在未见过的文本上。

近年来,继tf-idf、浅层语义索引(LSI)、pLSI模型之后,浅层狄利赫雷分配(Latent Dirichlet Allocation, LDA)[18]引起了人们的重视。LDA是一种包含三层的层级贝叶斯生成概率模型,它把文本语料看作离散数据,数据中的每一个元素看作是由底层的有限个混杂在一起的话题(topic)产生出来的,而每一个topic又被看作是从一个更底层的topic的概率模型中产生出来的。LDA克服了pLSI的理论缺陷,并且继承了pLSI的降维优势。

1.3 提出本文方法的动因

文献[19]提出了用LDA分析而得到的主题作为特征,对基于提问的多文档文本抽取摘要的方法,取得了较为明显的效果。作者采用的是一种无监督的方法,而且对主题数目对于摘要生成的影响也没有加以分析。

有监督的摘要方法虽然有语料相对匮乏的缺

点，却也不可否认的是当前摘要生成方法中效果最好的一种。对有监督摘要生成方法的研究有不可替代的重要意义[14]。

本文提出的新摘要方法 LCAS 的创新点在于：采用一种简便易行的办法第一次将有监督摘要方法中效果较为优秀的 CRF 方法和采用 LDA 的文本模型揉合在一起，既突出了有监督摘要方法中 CRF 的传统优势，又结合了 LDA 文本模型细腻的主题的概念。本系统通过和原有的 CRF 方法的比对，得到了 4%的系统提高。并且，本文还首次深入分析了 LDA 在不同的主题数量下对摘要生成带来的影响。

本文其余部分包括：第二部分简要介绍 CRF 原理及传统的 CRF 摘要系统所采用的基本特征；第三部分介绍 LDA 模型及本文提出的 CRF 模型和 LDA 模型相融合的摘要生成方法的系统构成；第四部分介绍实验设计及结果分析。最后一部分是本文的结束语。

2　CRF 简介

CRF 最先由文献[13]提出，它是一种条件概率模型 $P(Y|X)$，这里的 X 和 Y 可以有较为复杂的结构。CRF 的显著优势是它不像 HMM 和最大熵马尔可夫模型（Maximum Entropy Markov Models，MEMM）有那么强的独立性限制，并且没有 MEMM 的所谓标注偏移（Label Bias）问题。在本文中我们假设状态和标注一一对应（摘要句对应的标注为 1；非摘要句对应的标注为 0）。

给定一个观测到的句子序列 $X=(x_1, x_2, \cdots, x_M)$，输出相应的标注序列为 $Y=(y_1, y_2, \cdots, y_M)$，这里的 y_i 从一个集合 Ψ 中取值，如前所述，对于文本摘要句 $\Psi=\{0, 1\}$。CRF 的目标是找到序列 Y，使得下式最大化：

$$P(Y|X, W)=\frac{1}{Z(X)}\exp(W\cdot F(X, Y)) \quad (1)$$

(1)式里的 $F(X, Y)=\sum_{i=1}^{M} f(i, X, Y)$ 是一个维数为 T 的垂直向量。其中的垂直向量 $f=(f_1, f_2, \cdots, f_T)'$ 代表的是 T 个特征函数，也就是特征模板，每一个可以写作 $f_t(i, X, Y)\in R$，$t\in(1, \cdots, T)$，$i\in(1, \cdots, M)$。举例来说，假设第 10 个特征模板是：[if the length of the third sentence is bigger than 25] &[if the third sentence is a summary]，再假设 *text*_1 的标注序列为 *label*_*sequence*_1，这样，该篇文章的第 3 句用上述模板抽得的特征可以表示为 $f_{10}(3, text\_1, label\_sequence\_1)$。其意义为：带有标注序列 *sequence*_1 的文章 *text*_1，其第 3 句话是否长度大于 25，并且该句子是否为摘要。W 为一个维数为 T 的水平向量，代表各个特征的权重。公式(2)给出了 $Z(X)$ 的定义，它是一个归一化因子：

$$Z(X)=\sum_{Y}\exp(W\cdot F(X, Y')) \quad (2)$$

2.1　CRF 的参数估计

对于 W 的估计一般采用最大似然法，即给出训练数据和标注序列 X、Y，最大化对数似然函数：

$$L_W=\sum_{i=1}^{M}\log(P_W(Y_i|X_i)) \quad (3)$$

为了避免过拟合，一般采用一些正规化方法，常用的一种是给参数加入高斯先验。用于优化 L_W 的方法很多，常用的有 GIS 和 IIS[13]。本文采用收敛较快的一种准牛顿方法：L-BFGS 方法（给出参考文献）。

2.2　推断

当给出了(1)式定义的状态序列的条件概率和参数 W，估计概率最大的标注序列由下式给出

$$Y^{*}=\arg\max_{Y} P_W(Y|X) \quad (4)$$

其值可用维特比（Viterbi）算法方便地求出。序列中每一点状态的边缘概率可以用动态规划推断过程求得，这个过程和 HMM 的前向后向过程类似[13]。

2.3　基本特征

传统文本摘要中用到的特征都可以在 CRF 里得到应用。这里我们采用有监督文本摘要中常用的一些特征。这里我们称之为基本特征：

基本特征

长度特征：去除停用词后的句子长度。

位置特征：句子处在文章中的位置。在文章的开始为 1，结尾为 2，其余位置为 3。

对数似然特征：句子 x_i 由其所在文章 D 生成的对数似然值 $\log P(x_i|D)$。其定义为：$\sum_{w_k} N(w_k, x_i)\log p(w_k|D)$。这里的 $N(w_{ki}, x_i)$ 是词 w_k 在句子 x_i 中出现的次数；$p(w_k|D)$ 可以用 $N(w_k, x_i)/\sum_{w_j} N(w_j, D)$ 来估计。

主题词特征：指去除停用词后的高频词。句子包含的这种词越多，被标定为摘要句的可能性也越大。我们用这个特征来记录句子中的主题词个数。

指示词特征：一些含有诸如“in summary”和“in conclusion”这种词的句子很可能是摘要句。用这个特征来标识一个句子是否含有这种词。

大写词特征：一些专有名词和作者要强调的词语一般会被大写，含有这种词语的句子是摘要的可能性也较大。用这个特征来标识一个句子是否含有被大写的词。

相邻句子相似度特征：我们用余弦相似度来度量两个句子的相似度。这个特征用来记录一句话与其前三个和后三个句子的相似度。

3 LDA 模型及 LCAS 系统实现方法

LDA 模型是建立在狄利赫雷分布以及多项式分布的基础上的模型。多项式分布可以简单地理解为二项式分布的推广，这里不再介绍。狄利赫雷分布是伽玛分布的推广，而伽玛分布是二项分布共轭分布，它是多项分布的共轭分布。在介绍 LDA 模型前，下面先介绍狄利赫雷分布。

3.1 狄利赫雷分布

狄利赫雷分布是多项式分布的共轭先验分布，简单来说，可以理解成多项式分布中概率参数的分布。其公式表示为

$$Dir(\theta \mid \alpha) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} \tag{5}$$

这里

$$\alpha_0 = \sum_{k=1}^{K} \alpha_k, \quad \theta_i \geqslant 0, \quad \sum_{i=1}^{k} \theta_i = 1 \tag{6}$$

3.2 LDA 模型

LDA 是一个生成概率模型。在该模型下文本看作是由随机混合的浅层 topic 组成的，每一个 topic 对应所有词汇的一种概率分布。LDA 假设语料 D 中的每一篇文本有如下的生成过程[18]：

算法 1

1）选择 N ~ Poisson(ξ)；

2）选择 θ ~ Dir(α)；

3）对每一个词 w_n：

a）选择一个 topic z_n ~ Multinomial(θ)；

b）从概率分布 $p(w_n \mid z_n, \beta)$中选择一个词 w_n，p 为在 topic z_n下的一个多项式概率分布。

这虽然是个完备的数学模型，但是过于理论化，为了使用方便，文献[18]使用的这个模型又做了如下假设：

1）认为 topic 的数量 k 为已知；

2）认为概率矩阵 β 是固定的；

3）Poisson 分布被更实际的文本长度取代；

4）忽略 N 的随机性。

算法中的第 2)步用到了狄利赫雷分布，一般采用对称狄利赫雷分布，即(5)式中的 α 不但满足(6)式中的 $\alpha_0 = \sum_{k=1}^{K} \alpha_k$，且满足 $\alpha_1 = \cdots = \alpha_K$。这样给定 α、β 后，topic z 和文本 w 的联合概率为

$$p(\theta, z, w \mid \alpha, \beta) = p(\theta \mid \alpha) \prod_{n=1}^{N} p(z_n \mid \theta) p(w_n \mid z_n, \beta) \tag{7}$$

这里的 θ 和 z 都为隐含变量，对其求边缘概率，则一篇文本的概率为

$$p(w \mid \alpha, \beta) = \int p(\theta \mid \alpha) \left(\prod_{n=1}^{N} \sum_{z_n} p(z \mid \theta) p(w \mid z, \beta) \right) d\theta \tag{8}$$

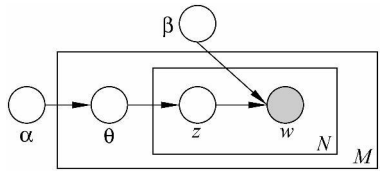

图 1 LDA 概率图模型

图 1 用概率图模型表示了 LDA 的生成过程。图中的方框代表了“重复”，也就是常说的词集(Bag of Words)的概念，即一种条件独立的假设。外层的方框代表了语料，其中有 M 个文本；内层的方框代表每篇文本的 N 个浅层 topic，以及由这 N 个 topic 生成出的 N 个词。图中隐含变量用空心园表示，实心变量为可见变量。从中可以清楚地看到 LDA 的三层概率模型：

1）α 和 β 是语料级的参数，一个语料库取一个值；

2）θ 是文本级的参数，每篇文本对应一个取值；

3）而 w 和 z 是词级的参数，每个词都要对应一次取值。

3.3 推断和参数估计

推断要解决的问题是，如果给出一篇文本，如何给出隐含变量的后验概率：

$$p(\theta, z \mid w, \alpha, \beta) = \frac{p(\theta, z, w \mid \alpha, \beta)}{p(w \mid \alpha, \beta)} \tag{9}$$

可是上式分母在求解时，因其 θ 与 β 的耦合(Coupling)关系，导致无法精确计算。不过，很多近似计算方法在近几年得到了广泛的研究，如马尔可夫蒙特卡罗马尔可夫方法(MCMC)、拉普拉斯近似以及本文要用到的变分近似(Variational Approximation)。

变分推断的基本思想是使用 Jesen 不等式来得到对数似然的一个可变下界。本质上说，就是用带索引的变分参数来得到一个下界函数族，然后通过最优化过程来确定最终的变分参数，最后得到一个最近的下界。通常逼近两个函数的方法是使两者的 KL 距离最小：

$$(\gamma^*, \varphi^*) = \arg\min_{(\gamma,\varphi)} D(q(\theta, z \mid \gamma, \varphi) \mid\mid p(\theta, z \mid w, \alpha, \beta)) \tag{11}$$

可以通过类似 EM 的迭代方法来搜索参数，所以总的运算约在 N^2K 的量级。

3.4　LCAS 系统设计

本文根据 LDA 模型的特点，提出了基于 LDA 的 CRF 自动摘要(LCAS)方法。我们在原有的 CRF 摘要抽取方法的框架下，引入 LDA 特征，LDA 特征描述的是文本通过 LDA 训练而得到的 topic 分布，这种 topic 的分布和词频特征一样可以作为 CRF 分类器的特征。系统框图见图 2。

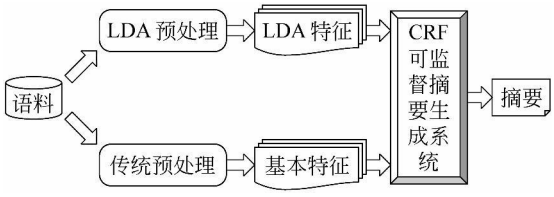

图 2　基于 LDA 的 CRF 摘要系统

图 2 中从语料出发向下的箭头为传统 CRF 摘要系统，通过预处理，提取基本特征送入 CRF 训练器训练，从而生成摘要；而从语料出发向上的箭头所指，则为本文系统特有的，依据 LDA 预处理而得到特征，将这些特征和基本特征合并，用 CRF 进行训练来提取摘要。

图中所用到的 LDA 特征定义如下：

LDA 特征

相邻句子的 topic 相似度特征(FA)：我们用余弦相似度来度量两个句子 topic 的相似度。这个特征用来记录一句话与其前 3 个和后 3 个句子的 topic 相似度。其计算方法和句子相似度的计算方法相同。

句子和所处文本的 topic 相似度(FB)：用余弦相似度度量每个句子和其所处文本的 LDA topic 相似度。

4　实验及结果分析

这部分首先介绍本文实验所采用的语料和评估标准。我们以文献[14] 为 baseline，为了验证 LDA 模型的有效性，实验设计和评估方法与文献[14] 所采用的严格一致。第二部分的实验中，我们对语料进行 LDA 分析，从实验结果中可以较清楚的看到 topic 的含义和作用。第三部分的实验是本文的核心，在这里系统的介绍了我们的 LCAS 系统的实验情况。我们给出了详细的实验结果，并和 baseline 进行比较。最终，4%的提高充分显示出了我们方法的优越性。同时还分析了 LDA 不同 topic 的数量对于实验结果的影响。

4.1　语料和评估标准

实验采用的语料是较为广泛使用的 DUC2001 年的摘要语料。DUC 是由 ARDA 赞助，NIST (http://www.nist.gov)举办的。DUC2001 的基本目标是引导研究人员在大规模的实验上推动摘要的发展。这批语料包括 147 篇新闻文本，文中每个句子是否是摘要句都被做了人工标注。这批语料本身就是为评测单文档抽取式摘要而设计的，并且做了很好的预处理。采用这个语料的另一个原因是，文献[14] 中的实验都是在这份语料上进行的，这样可以方便地将两个系统的性能进行比对。

摘要的评测共分为两种：一种是人工评测，也就是通过专家来对生成摘要的各个方面打分分出优劣；另一种是机器评测，现在使用的方法主要有 ROUGE 打分和通过 F1 值；文献[14] 采用了上述两种机器评测方法，从效果上说，这两种方法的评测结果基本一致。

为了方便比对，我们采用 F1 值来作为评估标准。

为了减小模型的不确定性，我们将语料分为 10 份并作交叉验证，最终的 F1 值是在这 10 个实验上的平均值。实验的设计和评测都是严格按照文献[14] 中所叙述的步骤进行，以保证比对的公正性。

4.2　实验 1：LDA 获取

我们在 DUC2001 年的语料上用算法 1 求得其 topic 分布，也就是多项式概率分布 $p(w \mid z)$，所得的

结果部分展示在表1。该表按概率从高到低的顺序给出了一些topic的分布情况。如topic1从高到低依次为eruption, ash, vocanic, vocanologist, bloom, information, activity, life-threatenning等等,可以认为这个topic应该主要是和火山活动的新闻有关的;topic2为kong, hong, patten, Chinese, people, politic, system, council, public, concession可以认为这是一个反映香港与大陆政治关系的一个主题;依次topic3和topic4分别为海上石油和北极问题。

表1 部分**topic**

Topic1	Topic2	Topic3	Topic4
eruption	kong	oil	ice
ash	hong	million	fish
vocanic	patten	valdez	antifreeze
vocanologist	chinese	ship	antarctica
bloom	people	tanker	sheet
information	politic	vessel	stream
acitivity	system	captain	scientist
life-threatenning	council	alaska	earth
scientist	public	mile	cold
warning	concession	set	water
dollar	government	crew	seal
torist	legislative	official	university

这个实验中我们采用的是基于变分近似的算法,采用EM迭代,因为语料规模较小,算法收敛的很快,一般在十几秒钟到一分钟以内就可以给出结果。

4.3 实验2:LCAS系统实验

我们对LCAS系统进行了实验,实验结果见表2。表2给出的数据是分别采用新特征FA和FB以及一起采用这两个特征FA+FB,并结合基本特征得到的F1结果。纵坐标为采用不同的topic数量,我们取从20个topic开始,每10个topic为单位逐渐递增,一直到100个topic。

图3中给出了实验结果的线条图形表示。

4.4 实验结果及分析

表2中的黑体数字,当topic取40,取特征FB,以及当topic取70,取FA特征和FB特征时,我们的系统达到最优值,0.405。它比文献[12]中的结果0.389提高了约4.1%。另外,从实验结果上看,FB特征比FA效果要好,它的曲线比较接近于FA+FB的曲线。本实验除了研究这两个特征的有效性,还对topic的数量对实验结果的影响作了测试。从结果上看topic取30到70之间的时候FB以及FA+FB的效果大都好于baseline,而且FA+FB对系统的提升对于topic的变化更不明显。所以应该说FA+FB的效果更具有鲁棒性。

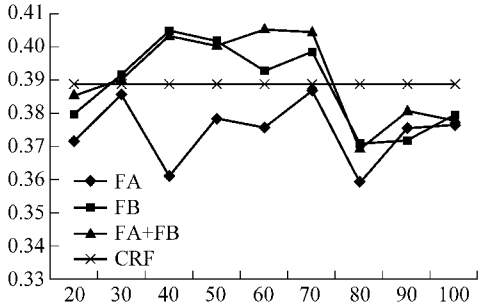

图3 LCAS系统实验结果

表2 **LCAS**系统实验结果

Topic	FA	FB	FA+FB
20	0.372	0.380	0.386
30	0.386	0.392	0.391
40	0.361	**0.405**	0.404
50	0.379	0.402	0.401
60	0.376	0.393	0.406
70	0.387	0.399	**0.405**
80	0.360	0.371	0.370
90	0.376	0.372	0.381
100	0.377	0.380	0.379

而从图3中我们可以清楚地看到,当topic数量由少到多增加时,实验结果无论是FA、FB还是FB+FA,都有明显的改善。而当topic数量过60~70时,结果又开始变差。

分析其原因,我们认为这是由于语料规模偏小,而topic数目设置过多从而导致了数据稀疏,使特征的作用下降。因此,在下一步研究中,我们将考虑语料的规模和topic的数量之间的关系。

另外,从表1中我们还可以看到,虽然LDA展现了一定的获得潜在topic的能力,但因词集(Bag of Words)的前提理论假设,使得同一个词可以被不

同的 topic 产生出来，如表中的“people”，“scientist”等词在不同的 topic 中都占有了很高的频率。文献[18]中也提到了这种情况，并提出了通过采用部分可交换性，或用马尔可夫性来描述词序列的方法来放松词集约束的理论假设。

5　结束语

本文提出了一种新的基于 LDA 的 CRF 自动摘要方法(LCAS)，该方法在传统的有监督的抽取型摘要的基础上，采用效果较好的 CRF 序列标注分类器，并结合 LDA 模型特点将新的特征加入到 CRF 分类器。实验证明本系统的性能比单纯采用传统的特征有较大的提高。并且本文还进一步分析了不同的 topic 对系统性能的影响并初步指出了原因。

我们的下一步工作是继续考察 LDA 模型对文本切割的作用，以及在文本初步切割的基础上提取摘要的效果。我们还希望研究在不同的语料规模下 topic 的数量对系统的影响。

参考文献：

[1] HP Luhn. The Automatic Creation of Literature Abstracts[J]. IBM Journal of Research and Development, 1958, 2(2): 159-165.

[2] 宗成庆. 统计自然语言处理[M]. 北京: 清华大学出版社, 2008.

[3] D. R. Radev, E. Hovy and K. McKeown. Introduction to the Special Issue on Summarization[J]. Computational Linguistics, 2002, 28(4): 399-408.

[4] Xiaofeng Wu, Chengqing Zong. A New Approach to Automatic Document Summarization[C]// International Joint Conference of Natural Language Processing, 2008: 126-132.

[5] J.Y. Yeh, H.R. Ke, W.P. Yang, and I.H. Meng. Text summarization using trainable summarizer and latent semantic analysis[J]. IPM, 2005, 41(1): 75-95.

[6] Hal Daum'e III, and D. Marcu. Bayesian Query- Focused Summarization[C]//ACL, 2006.

[7] P. B. Baxendale. Man-made Index for Technical Literature-An Experiment[J]. IBM Journal of Research and Development, 1958, 2(4): 354-361.

[8] Y. H. Gong and X. Liu. Generic text summarization using relevance measure and latent semantic analysis [C]//SIGIR, 2001: 19-25.

[9] J. Kupiec, J. Pedersen, and F. Chen. A Trainable Document Summarizer. Research and Development [C]//Information Retrieval, 1995: 68-73.

[10] C. Aone, N. Charocopos, J. Gorlinsky. An Intelligent Multilingual Information Browsing and Retrieval System Using Information Extraction[C]//ANLP, 1997: 332-339.

[11] H. P. Edmundson. New Methods in Automatic Extracting[J]. Journal of the Association for Computing Machinery, 1969, 16(2): 264-285.

[12] J. M. Conroy and D. P. O' leary. Text Summarization via Hidden Markov Models[C]//SIGIR, 2001: 406-407.

[13] J. D. Lafferty, A. McCallum, and F. C. N. Pereira. Conditional random fields: probabilistic models for segmenting and labeling sequence data[C]//ICML, 2001: 282-289.

[14] D. Shen, J. T. Sun, H. Li, Q. Yang, Z. Chen. Document Summarization using Conditional Random Fields[C]//IJCAI, 2007: 1805-1813.

[15] W. B. Frakes, R. Baeza-Yates. Information Retrieval Data Structures & Algorithms[M]. Prentice Hall PTR, New Jersey, 1992.

[16] S. Deerwester, S. Dumais, T. Landauer, G. Furnas, and R. Harshman. Indexing by latent semanticanalysis[J]. Journal of the American Society of Information Science, 1990, 41(6): 391-407.

[17] T. Hofmann. Probabilistic latent semantic indexing [C]//SIGIR, 1999.

[18] D. M. Blei, A. Y. Ng and M. L. Jordan. Latent Dirichlet Allocation[J]. Journal of Machine Learning Research 2003: 993-1022.

[19] Hal Daume III, Daniel Marcu. Bayesian Query-Focused Summarization[C]//ACL, 2006.

[20] C. M. Bishop, Linear Models for Classification, Pattern Recognition and Machine Learning[M]. chapter 4, 2006, Springer.

[21] 秦兵, 等. 多文档自动文摘综述[J]. 中文信息学报, 2005, 19(6): 14-20.