一种面向查询的多文档自动文摘系统实现方法*

桂卓民[1 2] 何婷婷[1 2] 陈劲光[1 2] 李芳[1 2]

1. 华中师范大学 计算机科学与技术系，湖北 武汉 430079

2. 国家语言资源监测与研究中心网络媒体分中心，湖北 武汉 430079

E-mail: Wenzheng38@sina.com tthe@mail.ccnu.edu.cn

cjg2003@hutc.zj.cn fang_lf@163.com

摘 要： 针对面向查询的多文档自动文摘，本文提出了一种系统实现方法。首先通过对句子结构的分析发现，句子中某些成分并不能反映该句子的重要信息，提出在一定句子的修剪基础上，基于倒几率比的词权计算方法与改进的HAL语言模型方法，并应用于文本的自动摘要。实验证明该方法对自动文摘的质量有一定提高。

关键词： 自动文摘， 倒几率比， HAL， 面向查询

A System of Approach to Achieve for Query-Focused Multi-Document Automatic Summarization

Zhuomin Gui[1 2] Tingting He[1 2] Jinguang Chen[1 2] Fang Li[1 2]

[1]Department of Computer Science, HuaZhong Normal University, 430079, Wuhan

[2]Monitor and Research Center of National Language Resource Network Multimedia Sub-branch Center, 430079, Wuhan

Wenzheng38@sina.com tthe@mail.ccnu.edu.cn

cjg2003@hutc.zj.cn fang_lf@163.com

Abstract: In this paper, we propose an approach to achieve a system for query-focused multi-document summarization. At first, based on the analysis of sentence structure, some of certain components of the sentence do not reflect the important information. This paper put forward the inverse odds ratio of the calculation weight of the words and the improved relevance-based language modeling (HAL) to create the auto-summary based on some pruning of the sentence. Experiments show that our method achieves a certain improvement of the quality of the automatic summarization.

Keyword: automatic summarization, inverse odds ratio, HAL, query-focused.

0 前言

随着互联网和信息高速公路的快速发展，网上的文本数据信息急剧增长，迫切需要对其进行有效的组织、总结和分析，以便快速、高效地掌握相关信息。因此如何快速、准确地获取信息成为信息处理研究的一个重要课题。自动文摘[1]正是致力于实现这种目标，为了满足特定任务的要求，利用计算机自动地从信息源文本中提炼出最重要的内容以生成一个简练的版本。

*项目资助：国家自然科学基金(60773167)；国家十一五科技支撑计划课题“网络文化安全预警技术研究”(2006BAK11B03)；973国家重点基础研究发展计划(2007CB310804)；教育部/国家外国专家局高等学校学科创新引智计划(B07042)。

1 简介

面向查询的多文档自动文摘就是基于特定的查询，将查询结果中多个文档的相关内容浓缩为一个覆盖主要相关主题、简洁、组织良好的摘要[2]。它除了能为用户提供所需信息的一个较为全面的摘要外，还能帮助用户判断和浏览感兴趣的具体信息内容。我们实验室一直致力于多文档自动文摘方面的研究，并参加了TAC 2008 的比赛。

1.1 TAC简介

TAC（Text Analysis Conference）是由国家标准技术局（NIST：National Institute of Standards and Technology）的信息技术实验室的信息存取分部 (IAD：Information Access Division)的检索小组组织的。于2008年，TAC从NIST的文件理解会议（DUC：Document Understanding Conference）的文章摘要和文本检索会议（TREC：Text Retrieval Conference）的问答追踪中脱颖而出。TAC由NIST和美国先进的智力研究计划活动（IARPA：Intelligence Advanced Research Projects Activity）主办。

TAC的目的是支持自然语言处理（NLP：Natural Language Processing）的研究，为NLP方法学的大规模评估提供必要的基础设施。它的目的不是竞争；而是强调通过对结果的评估来推进技术发展水平。特别是，TAC将致力于以下目标：

- 基于大范围共同的测试集来促进NLP的研究；
- 改进NLP的评估测试方法；
- 建立一系列的测试集，为NLP系统模型的评估需要；
- 为学术的交流研究提供一个开放性的论坛；
- 改进NLP在现实生活中研究的方法，加速从实验室研究到商品化生产的技术转化；

TAC 2008 年的主要任务有：

- 问答系统：从大型文档集合中检索确切的答案；
- 多文档文摘系统；

1.2 我们的工作

在2008年的第一届TAC比赛中，我们的系统，CCNUS (Centre China Normal University for Summarization)，采用的是基于词权计算和改进的HAL语言模型的文摘生成方法。首先，从给定的文本中抽取内容并进行句子的切分；然后，基于句法分析我们建立句子的句法树，对句子进行指代消解和句子修剪；最后，采用倒几率比的思想对词语进行权重计算，并结合改进的HAL相关语言模型计算查询相关，来获取句子的得分，基于MMR算法来抽取句子生成文摘。

本文的第二部分将介绍我们在TAC 2008中使用的方法。第三部分列出了我们的部分实验结果。第四部分是总结和将来的工作。

2 系统实现方法

系统的实现过程分为预处理、句子压缩、动态权重计算和后期生成处理四个部分，可以通过图1 来表示。下面将详细介绍系统的每个过程。

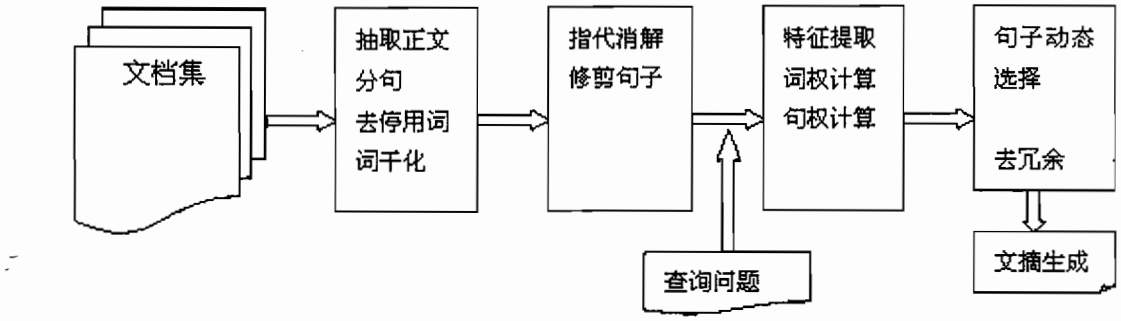

图1　系统框架

2.1 预处理过程

对比赛提供的数据进行预处理，以确定哪些是重要的内容，哪些词最能表达文档的内容，具体的步骤包含有：

①去掉标签，提取正文；②对句子进行切分、标号，去掉疑问句和感叹句；③去除停用词；④对词语进行词干化。

2.2 句子压缩过程

①指代消解

系统中，我们使用了Long Qiu，Min-Yen Kan 等人提出的基于语法的指代消解模型[3]。通过句法分析，它不是找最近的名词（通常的做法），而是找出符合匹配规则的短语或者语言单位，用来处理第三人称代词、省略的主语、宾语。

在应用于文摘时，我们根据需要将句子的指代消解过程分为句子之间的指代消解和句子之内的指代消解，为了句子的清晰性和简略性，本系统只考虑句子之间的指代消解过程。

②句子修剪

这一阶段我们使用了基于句法的句子压缩方法。Charniak提出基于最大熵的句法解析树模型，它将一个句子解析成一颗句法树[4][5]。在基于句法树的基础上，我们提出了一些规则来修剪句法树。

我们简单列出五条修剪规则，在保证合乎语法的基础上消除尽可能多的冗余信息。这些规则随着强度的增加，可信性在下降，即规则n（$1 \le n \le 4$）造成的句子非合法性肯定比规则n+1少。规则1消除的是介词短语，特点是以逗号结尾或者父亲结点是从句；规则2消除的是关系从句；规则3消除的是同位格；规则4消除的是动名词短语；规则5消除一些特殊短语，例如“, John said”。

修剪过程可能会丢掉一些有用的信息。本系统动态性地对句子进行一定程度的修剪。即句子的修剪是在不丢掉重要信息的前提下。

2.3 动态权重计算过程

对句子的动态权重计算，以确定选择哪些句子作为文摘句，通过句子自身的得分与句子和查询问句的得分来反映。在这一过程中主要步骤有：

①特征向量提取

在整个过程中，本文使用特征向量 {<t1,w1>,<t2, w2> ,…,<tn, wn >} 表示每一个句子,其中wi表示第i个特征ti的权重。

本系统采用的是基于几率比的方法来提取特征的。几率比的思想是：一个词语如果在正样本中出现的频率较大，而在负样本中出现的频率较小，说明这个词语越能反映正文本的主题，所以就给它较高的权重。

$$OR(t) = \log \frac{p(t \mid Cpos)(1 - p(t \mid Cneg))}{p(t \mid Cneg)(1 - p(t \mid Cpos))} \tag{1}$$

其中Cpos表示正样本的情况，Cneg表示负样本的情况。p(t|Cpos)表示词项t在正样本中出现的概率，p(t|Cneg)表示词项t在负样本中出现的概率。

在实验中把话题下的文本分为正样本和负样本，把话题集的文本看成是正样本，把当前的文本看成是负样本，统计词语在正样本和负样本中出现的频率，以及在正样本和负样本中没有出现该词语的频率。

②句子权重计算

句子单元的得分是基于多特征融合的打分机制。第一种特征是句子单元本身含有的信息量，第二种特征就是句子单元与查询话题的共现概率，两者采用线性加权和就构成句子单元的得分。经过去掉停用词和词干化预处理后，句子留下来的都是有效词，及其对应的词权。通过这些词权的平均大小来衡量句子单元含有的信息量。

$$\text{Score1}(S) = \frac{1}{n}\sum_{i} w(ti) \tag{2}$$

n表示一个句子中有效词的个数，w(ti)表示词项ti的权重。

本系统中，在计算句子单元和查询话题的共现概率时，在Jagadeesh基于相关语言模型的基础之上，我们进行了改进。

Jagadeesh证明了，在高维中用HAL空间（Lund and Burgess, 1996）可以更好地计算相关语言模型（Lavrenko and Croft, 2001）与语义代表性的词之间的相关性。模型假设查询话题和文档是来自相关模型R中的例子，计算P(w|R)的概率，即计算文档中的词w与特定信息模型R发生的概率，可用P(w|Q)替代， Q=(q_1,q_2,……,q_k)。通过定义，条件概率可以通过联合概率分布来实现：

$$p(w \mid R) \approx p(w \mid Q) = p(w \mid q_1, q_2, ..., q_k) = \frac{p(w, q_1, q_2, ..., q_k)}{p(q_1, q_2, ..., q_k)} \tag{3}$$

假设查询词之间是相互独立的，根据独立性假设有：

$$p(w, q_1, q_2, ..., q_k) = p(w)\prod_{i} p(q_i \mid w) \tag{4}$$

$p(q_i|w)$，可用HAL（Hyperspace Analogue to Language）模型来解释，HAL模型是基于词w和查询词qi在一定窗口K之下共现的概率。

在2003 DUC比赛中，International Institute of Information Technology (IIIT)使用该模型取得了较好的成绩。但在该方法上有个问题，事实上，给定一个查询Q=(q_1,q_2,……,q_k)，假设情况一：

一个词和查询的每个词都弱相关；情况二：一个词与查询中的某个词强相关而与其他的词不相关，人们往往接收后者。

假设查询词q_1,q_2和保留词w_1,w_2，它们的概率如下：$p(w_1)=p(w_2)=0.1$，$p(q_1 \mid w_1)=p(q_2 \mid w_1)=0.1$，$p(q_1 \mid w_2)=0.01$，$p(q_2 \mid w_2)=0.9$。在Jagadeesh模型中则有：

$$p(w_1,q_1,q_2)=p(w_1)p(q_1 \mid w_1)p(q_2 \mid w_1)=0.1\times 0.1\times 0.1=0.001$$

$$p(w_2,q_1,q_2)=p(w_2)p(q_1 \mid w_2)p(q_2 \mid w_2)=0.1\times 0.01\times 0.9=0.0009$$

实际上，词w_2更能体现与查询Q之间的相关性。为此，我们将公式(4)改进如下：

$$p(w,q_1,q_2,\ldots,q_k)=p(w)\sum_i p(q_i \mid w) \tag{5}$$

根据上面假设，基于改进后的联合概率有：

$$p(w_1,q_1,q_2)=p(w_1)(p(q_1 \mid w_1)+p(q_2 \mid w_1))=0.1\times(0.1+0.1)=0.02$$

$$p(w_2,q_1,q_2)=p(w_2)(p(q_1 \mid w_2)+p(q_2 \mid w_2))=0.1\times(0.01\times 0.9)=0.091$$

正好词w_2体现了与查询主题的相关性。

对于句子S，查询Q，假设句子中的词语是相互独立的，结合公式(3)和(5)，则句子S和查询Q相关性描述如公式(6)：

$$Score2(S,Q)=\frac{\sum_{t_i\in S} p(t_i \mid Q)}{L}=\frac{\sum_{t_i\in S} p(t_i)\sum_{q_j\in Q} p(q_j \mid t_i)}{L}=\frac{\sum_{t_i\in S} p(t_i)\sum_{q_j\in Q} pHAL(q_j \mid t_i)}{L} \tag{6}$$

其中L是句子的长度，pHAL是HAL的概率，即查询词q_j与给定词t_i在给定窗口K大小的共现概率。

综合考虑句子单元本身的信息量和与查询句的相关度可得出句子的得分为：

$$\mathrm{Score}(S)=\alpha Score1(S)+(1-\alpha)Score2(S,Q) \tag{7}$$

α是实验参数，可以通过测试语料得到。

2.4 文摘生成过程

在文摘生成的过程中，根据候选句的权重，首先我们动态的挑选文摘句；在保证冗余度最小的前提下，然后再选取权重高的候选句。具体细节如下：

①句子的动态选择

通过上面步骤知道了原句子与修剪后句子的得分，在选择原句子还是修剪后的句子，我们采用动态的方法。即是，不在丢掉重要信息的前提下来选择句子。假设原句子与修剪后的句子的得分分

别为：$\mathrm{Score}_{original}$、$\mathrm{Score}_{cut}$，原句子与修剪后的句子的长度分别为：$\mathrm{Length}_{original}$、$\mathrm{Length}_{cut}$。如果满足下面公式(8)就采用修剪句子，否则的话说明修剪的过程剪除了原句子中认为是重要的信息，就依然选择原句子。

$$\frac{\mathrm{Score}_{original}-\mathrm{Score}_{cut}}{\mathrm{Length}_{original}-\mathrm{Length}_{cut}} \le \delta \frac{\mathrm{Score}_{original}}{\mathrm{Length}_{original}} \tag{8}$$

δ 是实验参数，可以根据测试集来调整。

②去冗余

通过上面步骤抽取的文摘句冗余度较大，要通过句子相似度计算减少冗余句子。具体过程如下：

i 首先将句子按其得分从高到底的排序；

ii 抽取得分高的句子Si；

iii 选取候选句Si后，对于待选句Sj，其得分将按公式(9)进行调整；

$$\mathrm{Score}\ (S_j) = Score(S_j) - sim(S_i, S_j) * Score(S_i) \tag{9}$$

其中 $sim(S_i, S_j)$ 是两个句子的相似度。

iv 再对剩下的句子按其得分进行从高到底的排序，选取得分高的句子；

v 如此重复，直至满足合乎词数为止。

3 实验分析

在TAC 2008年的比赛中，我们提交了两个系统。1 号系统中，我们没有做句子的指代消解和修剪工作，2 号系统中，我们做了句子的指代消解和修剪工作。系统的评价统计数据见表1：

表1 系统的比较

系统	mod Pyramid score	numScus	linguistic quality	overall responsiveness	Rouge2 R	RougeSU4 R
1	0.300	4.521	2.396	2.379	0.081	0.1225
2	0.300	4.667	2.458	2.547	0.083	0.1343
	-0.0%	+3.2%	+2.6%	+7.1%	+2.5%	+9.6%

从上面的表中可以看出，系统在加入了句子的指代消解和修剪工作之后，各项的评价得分都有明显的提高，尤其是在总体召回率和语言质量上。

我们也与哈尔滨工业大学信息检索实验室提交的#11号系统得分进行了比较，评价统计数据见表2：

表2 我们#2与哈工大#11的比较

系统	mod Pyramid score	numScus	linguistic quality	overall responsiveness	Rouge2 R	RougeSU4 R
11	0.336	4.781	2.406	2.542	0.088	0.125
2	0.300	4.667	2.458	2.547	0.083	0.134
	-12.0%	-2.4%	+5.2%	+0.3%	-5.8%	+6.7%

哈工大的11号系统采用了子主题聚类，并且是基于改进的多样化句子评分。因此，在金字塔评测和Rouge2 R 上，好过我们；但由于我们采用了指代消解和句子修剪，并且是动态给句子评分，所以2在语言质量和RougeSU4 R 好过11.

4 结论与展望

本文介绍了一种基于句法统计规则的面向查询的多文档自动文摘系统。我们使用DUC 2007年的语料作测试集，通过实验来调整参数，改善系统。并参加了TAC 2008的比赛，在自动文摘的质量上也有一定的提高。

在多文档自动文摘方法中，基于语义计算的方法一直是我们实验室关心的重点，也是我们下一步的工作。在实验中，我们遇到了一些困难，如何将语义计算方法进一步扩展，以及如何运用机器学习的策略来完善语义计算等，是我们今后主要努力的任务。

参 考 文 献

[1] 郑义, 黄萱菁, 吴立德. 2003. 文本自动综述系统的研究与实现. 计算机研究与发展.

[2] Wauter Bosma. Query-Based Summarization using Rhetorical Structure Theory. In Proceedings of CLIN04.2005

[3] Long Qiu, Min-Yen Kan and Tat-Seng Chua. (2004). A Public Reference Implementation of the RAP Anaphora Resolution Algorithm. In proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004). Vol. I, pp. 291-294.

[4] E. Charniak. A maximum-entropy-inspired parser. In Proceedings of the Conference on North American Chapter of the Association for Computational Linguistics (ANLP-NAACL), pages 132-139, 2000.

[5] E. Charniak and M. Johnson. Coarse-to-fine n-best parsing and MaxEnt discriminative reranking. In Proceedings of the Annual Meetings of the Association for Computational Linguistics (ACL), pages 173-180, 2005.

[6] 邵伟，何婷婷，胡珀. 一种面向查询的多文档文摘句选择策略. 2007年第九届全国计算语言学学术会议.

[7] Lappin, S. Leass, H. 1994. An algorithm for pronominal anaphora resolution, Computational Linguistics, 20(4), 535-561.

[8] Jaime Carbonell, Jade Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In Proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval.

[9] Wauter Bosma. 2005. Query-Based Summarization using Rhetorical Structure Theory. In Proceedings of CLIN04.

[10] http://www.nist.gov/tac, 2008.