一种利用关键词提取的面向查询多文档文摘技术*

马亮[1,2] 何婷婷[1,2] 陈劲光[1,2] 李芳[1,2] 邵伟[1,2]

(1. 华中师范大学计算机科学系 武汉 430079

2. 国家语言资源监测与研究中心网络媒体分中心 武汉 430079)

Email: maliang1897@yahoo.com.cn

摘要： 针对面向查询的多文档自动文摘，本文提出了一种利用关键词提取技术的文摘句选择策略。通过计算多文档集中词语的查询相关性特征和语料相关性特征，并将词语的两个特征值进行特征融合得到每个词语的重要度，随后通过词语的重要度来给候选句打分，进一步利用改进的 MMR（Maximal Marginal Relevance）技术来调整候选句的得分，最后生成文摘。本文将特征融合引入到词语层面，在 DUC2005 的语料中测试效果很好。

关键词： 多文档文摘；关键词提取；文摘句选择

The Technique of Using Key Words Distillment for Query-focus Multi-document Summarization

Ma Liang[1,2] He Ting-ting[1,2] Chen Jin-guang[1,2] Li Fang[1,2] Shao Wei[1,2]

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

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

Abstract: This paper proposes a strategy of sentence selection for query-focus Multi-document Summarization through distilling key words from related multi-documents. It calculates the related feature with query and the related feature with other words in corpus for every word in related multi-documents, then obtains the word's importance by combining the two features. The score of candidate sentence is computed by the importance of words which the candidate sentence contains. The modified MMR technology is used to adjust the score of candidate sentence, and the summary is generated. Experiments showed that our method performed very well in DUC 2005 corpus.

Key words: Multi-document Summarization; key words distilling; sentence selection

1. 引言

为了从海量信息中快速、准确地获取有用信息，多文档自动文摘技术变得越来越重要。面向查询的多文档自动文摘技术的任务是基于特定的查询，将大量的查询结果文档中的相关内容浓缩为一个包含与查询相关的各个主题，并且内容简洁、组织良好、冗余低、满足个性化需求的摘要，它更具有针对性，更能适应当前 Internet 环境下对于信息获取的个性化需要。目前国内外学者关于面向查询的多文档自动文摘都进行了很多有意义的探索研究。

Prasad Pingali 等人[1]通过手工修剪句子，然后提出独立于查询的特征和依赖于查询的特征，并对两个特征分别进行打分，最后将每个句子的两个特征线性组合以得到最后的分数，再考

*基金项目：国家自然科学基金，编号：60773167；国家社会科学基金，编号： 06BYY029；湖北省自然科学基金计划项目，编号：2006ABC011； 973 国家重点基础研究发展计划，编号：2007CB310804；教育部/国家外国专家局高等学校学科创新引智计划，编号：B07042；

虑冗余性的情况下反复抽取出分数最高的句子作为最后的文摘句。该系统在 DUC2007 的评测比赛中取得了多项第一的成绩。

Ziheng Lin 等人[2]提出一种构建带时间戳的图模型来模拟人类写作和阅读的过程，从图中寻找关系对句子排序，并采用 MMR 技术抽取文摘句构成文摘，该系统在 DUC2007 年的评测中取得了较好的结果。

Chin-Yew Lin 和 Eduard Hovy[3]提出了一种在文本摘要中自动获取话题信号词的方法，该方法利用相关文档集和非相关文档集中词语的频率来计算话题信号词的统计量，并采用互信息量和最大似然估计量来近似计算，能够较好的提取出与话题相关的信号词。

John M. Conroy 等人[4]通过人工生成的多篇文摘来近似计算查询条件下信号词出现的概率，并将该值作为经验值用来挑选文摘句，还加入了语言知识的预处理（比如动名词短语和前导副词的消除）和冗余消除，最后生成的文摘性能在 ROUGE-2 的评测中超过了人工的文摘。

邵伟、何婷婷[5]采用一种多特征融合的文摘句选择策略，通过句子与查询的关联特征及句子的全局关联特征的融合来抽取文摘句以生成摘要，取得了较好的效果。

面向查询的多文档自动文摘系统中文摘句的选择相当重要，选择出的文摘句既要尽可能地满足用户的查询需要，又要尽可能地包含文档集的重要内容。词语是句子的基本组成单位，可以在更细的层面上刻画语义信息，本文提出一种利用关键词提取技术来选择文摘句的策略。首先根据查询相关性特征和语料相关性特征的线性组合来选择既与用户的查询相关性高、又包含文档集的重要内容的关键词语；然后根据这些关键词语的重要度给候选句打分，并利用改进的 MMR 技术来选择文摘句，最后生成文摘。

本文第二部分详细描述了面向查询的多文档自动文摘的流程，第三部分是本文的实验与结果分析部分，第四部分是结论与将来的工作部分。

2. 面向查询的多文档自动文摘的流程

从下面的流程图中可以看出，本系统分为语料获取、关键词打分和文摘生成等阶段，如图 1 。

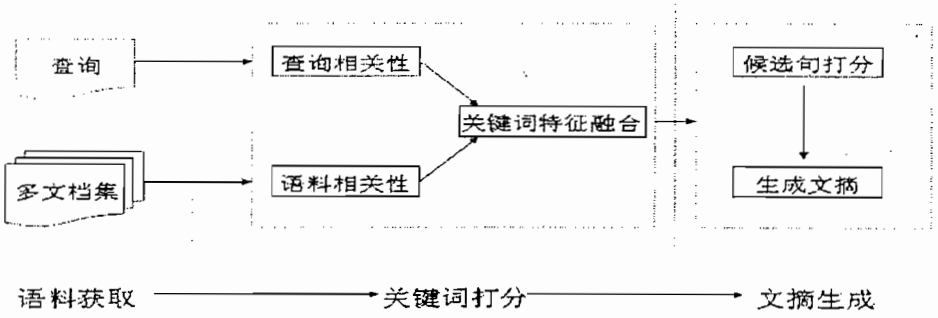

图 1　系统流程图

2.1 语料获取阶段

在面向查询的多文档自动文摘中，本文根据用户的查询条件得到相关的多文档集，并假设在该文档集中一定存在相应的文档或句子，它们能够尽可能的满足用户的查询需求，同时也较好

地概括了文档的主要内容。本文同时也选择与用户的查询条件不相关的文档集作为对比来给关键词语进行打分。

2.2 关键词打分阶段

面向查询的多文档自动文摘系统中为了让抽取出来的文摘句既能满足给定的查询话题，又能最大限度代表文档集的重要信息，采用两个重要特征来对关键词语进行打分：一个是查询相关性特征；另一个是语料相关性特征。最后将两个特征进行线性组合来对关键词语进行打分。

2.2.1 查询相关性特征

为了让提取出来的关键词语能尽可能的满足给定的查询条件，本文通过构建矩阵来度量相关多文档集中的词语与查询条件中词语的语义关系。

假设相关多文档集中的词语种数为TWT，查询话题中的词语种数为qwt，则可以构建一个$TWT*qwt$的矩阵，每行分别代表相关多文档集中的一个词语，每列分别代表查询话题中的一个词语，矩阵中的值代表该行与该列所对应的词语之间的语义关联度。

在计算两个词语w_1和w_2的关联度时，假设存在一个大小为K的窗口，两个词语w_1和w_2在大小为K的窗口中共现的次数为$n(w_1,k,w_2)$，其中k代表词语w_1和w_2在共现时的实际间隔词数，用$W(k)=K-k+1$表示词语w_1和w_2在共现时的共现强度，则两个词语w_1和w_2的语义关联度为：

$$A(w_1,w_2)=\sum_{k=0}^{K}W(k)*n(w_1,k,w_2) \quad (1)$$

那么，相关多文档集中的词语w_i的查询相关性特征可以用如下公式进行计算：

$$F_1(w_i)=\sum_{j=0}^{qwt-1}A(w_i,w_j) \quad (2)$$

2.2.2 语料相关性特征

为了让抽取出来的关键词语能最大限度代表相关文档集的重要信息，本文考查了词语的语料相关性特征，即利用与查询不相关的语料来凸显与查询相关的语料中词语的重要程度。

假设我们在语料获取阶段得到的与查询相关的文档集为 R，与查询不相关的文档集为 NR，分别统计出文档集 R 中的词语种数为TWT，文档集 R 中总共的词语个数为$TWNR$，文档集 NR 中总共的词语个数为$TWNNR$。假设语料中每个词语的出现概率相等，即语料中的每个词语w_i出现概率为：

$$p=1/TWT \quad (3)$$

记词语 w_i 在相关文档集 R 中的出现次数为 $fre(w_i)$，该词语 w_i 在文档集 R 中的出现概率为：

$$p_1 = fre(w_i)/TWNR \tag{4}$$

那么相关文档集 R 中其它词语的相对概率为：

$$p_2 = (TWNR - fre(w_i))/(TWNR*(TWT-1)) \tag{5}$$

同时统计出词语 w_i 在与查询不相关的文档集 NR 中的出现次数为 $freq(w_i)$，为了方便表述，令

$$M = TWNR*\log p + TWNNR*\log(1-p) \tag{6}$$

$$N = fre(w_i)*\log p_1 + freq(w_i)*\log(1-p_1) + (TWNR - fre(w_i))*\log p_2 + (TWNNR - freq(w_i))*\log(1-p_2) \tag{7}$$

此时可以根据下面的公式来计算词语 w_i 的语料相关性特征分数为：

$$F_2(w_i) = -2*(M-N) \tag{8}$$

2.2.3 关键词特征融合

关键词语要尽可能地满足上面的两个特征，所以本文将上述的两个特征值进行了线性组合，以便词语 w_i 的最后得分 $Weight(w_i)$ 能够代表词语 w_i 的重要度：

$$Weight(w_i) = \theta * F_1(w_i) + (1-\theta) * F_2(w_i) \tag{9}$$

其中，θ 是参数，需要通过大量的实验来确定。

2.3 文摘生成阶段

关键词语的重要度得到后，如何由这些关键词语来选择文摘句就成了重要的问题，这需要充分挖掘和利用关键词语的重要度信息。在利用关键词语重要度信息对候选句进行打分后，再利用改进的 MMR 技术动态调整候选句的分数，并每次选择分数最高的候选句作为文摘句，直到生成的文摘长度达到要求。

2.3.1 候选句打分

候选句的得分会影响到该候选句能否成为文摘句，在由关键词语的重要度对候选句进行打分的时候有两个方法。第一个方法是：将关键词语的重要度从高到低排序，选择前面的 T 个关键词作为核心词，假设这些核心词是平等的，包含这些核心词越多的候选句越有可能成为文摘句。第二个方法是：根据关键词语的重要度按照打分公式对每个候选句进行打分，分数越高越有可能成为文摘句。

方法一：候选句的分数是由包含的核心词语的个数来确定的，这样做是为了保证文摘句尽可能覆盖核心词语。候选句的分数可以由下面的公式来计算：

$$F(S_i)=\frac{1}{|n|}\sum_{j<=T}S_i(w_j) \tag{10}$$

其中，n 是句子 S_i 包含的所有词语数目，$S_i(w_j)=\begin{cases}1\\0\end{cases}$ 表示句子 S_i 是否包含核心词语 w_j。

方法二：候选句的分数是由其所包含的关键词语的重要度之和来确定的，这样做可以保证在文摘句尽可能覆盖关键词语的同时也考虑到每个关键词语之间重要度的差异。候选句的分数可以由下面的公式来计算：

$$F(S_i)=\frac{1}{|n|}*\sum_{w_j\in S_i}Weight(w_j) \tag{11}$$

其中，n 是句子 S_i 包含的所有词语数目，$Weight(w_j)$ 表示句子 S_i 包含的关键词语 w_j 的重要度。

2.3.2 生成文摘

按照每个候选句 s_i 的分数从高到低将所有的候选句进行排序，抽取出分数最高的候选句作为文摘句。

经典的MMR技术[6]可以使得候选句尽可能的与查询话题相关并且与已经选择出来的文摘句不重复，其定义为：

$$MMR=\arg\max_{D_i\in R-S}[\lambda sim(D_i,Q)-(1-\lambda)\max_{D_j\in S}sim(D_i,D_j)] \tag{12}$$

其中，S 表示从文档集合中选择出来的句子集，Q 表示查询话题，R 表示已经排好序的文档集。

本文在关键词的重要度计算时已经考虑了关键词的查询相关性，所以在生成文摘的阶段重点是考查文摘句的冗余性，即候选句与已经被选的文摘句之间的相似性。利用下面改进的公式来动态地调整候选句的分数：

$$F(S_i)=\lambda*F(S_i)-(1-\lambda)*\max sim(S_i,S_j) \tag{13}$$

其中，S_i 表示候选句，S_j 表示已经选择出来的文摘句，λ 是参数，由实验确定。

S_i 与 S_j 之间的相似性仍然采用向量夹角余弦公式来计算。

3. 实验结果与分析

本文采用 DUC2005 提供的语料[7]来做实验。DUC2005 的官方语料给定了 50 个话题（topic）

和 50 个相应的查询问题，每个话题选择 25—50 篇来自新闻周刊和金融时报的相关文档，其任务是生成能够回答话题中问题的简洁、组织良好、流畅的多文档文摘。

在英文的文章中，常常涉及到词形的变化，比如："keep"，"keeps"，"keeping" 等表示的是相同的意思，只是词形或时态不同而已，在计算词频的时候应该引起注意，因此本文在实验的过程中采用了词干化技术[8]对所有的词语进行词干化处理，这样可以使得提取出来的关键词语的重要度不会因为词形的差异而产生误差。

DUC2005 的官方语料给定了 50 个话题的相关文档集，本文在关键词语打分阶段计算词语的语料相关性特征时每选定一个话题下的文档集作为相关文档集，那么剩余的 49 个文档集就会作为不相关的文档集参与词语的语料相关性特征计算。

为了使实验数据具有可比性，本文选择了 2005 年的参赛队伍 SystemID15 的评分结果进行对比。参赛队伍 SystemID15 在 2005 年的 Rouge2 和 RougeSU4 官方评测中获得了第 1 名的好成绩。本文针对部分话题得到的文摘的 Rouge 评测结果数据如表 1 所示。其中，表中的 d307b、d313e 、d321f 、d331f 、d343c、d357i、d438g、d446j、d694j、d699a 分别代表话题名称。

本文对候选句打分阶段的方法一和方法二都进行了实验并将 Rouge 评测结果归纳于表 1 中，方法一是指本文在候选句打分阶段选用的公式（10），方法二是指本文在候选句打分阶段选用的公式（11）。

表 1　本文对部分话题得到的文摘 Rouge 评测结果

topic	方法一		方法二		SystemID15	
	Rouge2	RougeSU4	Rouge2	RougeSU4	Rouge2	RougeSU4
d307b	0.05894	0.12436	0.08397	0.12176	0.06063	0.11587
d313e	0.05162	0.13774	0.05533	0.14242	0.08911	0.1446
d321f	0.07943	0.12254	0.09632	0.15946	0.09574	0.16345
d331f	0.0826	0.12823	0.0943	0.14923	0.09501	0.15637
d343c	0.08115	0.11972	0.08341	0.13214	0.07932	0.12148
d357i	0.11489	0.16957	0.12631	0.17946	0.12898	0.18263
d438g	0.0845	0.13928	0.09364	0.1253	0.09349	0.14573
d446j	0.05251	0.0937	0.06798	0.10615	0.05529	0.10403
d694j	0.02418	0.0864	0.02695	0.09347	0.03226	0.09923
d699a	0.07952	0.10856	0.08146	0.12489	0.07369	0.12216

从表 1 可以看出，本文提供的方法二在话题 d307b、d321f 、d343c、d438g、d446j、d699a 下的评测结果要高于 SystemID15 的评测结果，在其它话题下的评测结果略低于 SystemID15 的评测结果，我们对照 Rouge 评测分数将源文档集中的源文档内容进行分析发现，采用本文方法二得分较高的话题下的源文档中各标点符号有利于候选句的切分与识别，并且文档中的句子长短适中，有利于利用关键词语的重要度来给候选句打分，最后生成的文摘中每句话的长度也适中；得分较低的话题下的源文档中部分句子语法结构复杂，句子太长或太短，造成利用关键词语的重要度来给候选句打分时产生误差。比如在话题 d313e 下生成的文摘中有几句话只有 3 个或 4 个词语，影响了文摘的质量。

本文提供的方法一产生的文摘质量比方法二产生的文摘质量要差，原因在于方法一产生的文摘质量受源文档中句子长短的影响更大，利用关键词语的重要度来给候选句打分时，是否一定应该除以候选句中的总词数或者是对总词数这个参数进行一定地处理（比如开平方）还需要根据具体的语料来定，大量实验表明，如果文档中的句子长短适中，本文所提出的文摘句选择策略生成的文摘质量会提高。

在实验过程中还发现，采用方法一时如果将关键词语划为核心词语的数目扩大，比如将前300个关键词语划为核心词语改为将前400个关键词语划为核心词语，文摘的质量有了一定的提高。

4. 结论与将来的工作

词语作为句子的基本组成单位，可以在更细的层面上刻画语义信息，本文提出一种利用关键词提取技术来选择文摘句的策略。首先根据查询相关性特征和语料相关性特征的线性组合来选择既与用户的查询相关性高、又包含文档集的重要内容的关键词语；然后根据这些关键词语的重要度给候选句打分，并利用改进的MMR技术来选择文摘句，最后生成文摘。

本文利用词语在给定窗口中的共现强度来度量相关多文档集中的词语与查询条件中词语的语义关系，并且利用与查询不相关的语料来凸显与查询相关的语料中词语的重要程度，这两个特征很好地刻画了关键词语的重要度，既体现了查询条件，又体现了文档集的重要内容。

在下一步的工作中，如何确定关键词语打分阶段和文摘生成阶段的公式中的最佳参数是我们需要研究的问题，在本文的实验中，我们还发现参数与语料的多少以及语料的合理性有很大的关系。我们还将研究文摘句的排序问题和文摘句修剪策略使生成的文摘在可读性方面进一步地提高；同时也将对中文语料进行实验，开发中文语料的面向查询的多文档自动文摘系统。

参考文献

[1] Prasad Pingali, Rahul K and Vasudeva Varma.2007.IIIT Hyderabad at DUC 2007. In Proceedings of DUC2007.

[2] Ziheng Lin, Tat-Seng Chua, Min-Yen Kan. 2007. NUS at DUC 2007: Using Evolutionary Models of Text . In Proceedings of DUC 2007.

[3] Chin-Yew Lin and Eduard Hovy. 2000. The Automated Acquisition of Topic Signatures for Text Summarization. In Proceedings of the 18th conference on Computational linguistics, Morristown, NJ, USA. Association for Computational Linguistics.

[4] John M. Conroy, Judith D. Schlesinger, Dianne P. O'Leary. 2006. Topic-Focused Multi-document Summarization Using an Approximate Oracle Score. Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions. Association for Computational Linguistics.

[5] 邵伟，何婷婷等. 一种面向查询的多文档文摘句选择策略. 2007. 第九届全国计算语言学学术会议. 2007. 8.

[6] J. Carbonell and J. Goldstein.1998. The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries. In Proceedings of SIGIR'98. Melbourne, Australia, 1998.

[7] http://duc.nist.gov/

[8] The Porter Stemming Algorithm http://tartarus.org/~martin/PorterStemmer/