基于维基百科的多文档自动摘要系统研究*

刘茂福，余 博，胡慧君
（武汉科技大学 计算机科学与技术学院，湖北 武汉 430065）

摘 要：设计并实现了一个基于维基百科的抽取式多文档自动摘要系统。使用 ROUGE 评测工具对使用维基百科前后的摘要进行对比实验。实验结果表明，维基百科能较大程度地提高多文档摘要的质量。

关键词：多文档自动摘要；维基百科；句子抽取
中图分类号：TP391.1 **文献标识码**：A **文章编号**：1674-7720(2011)16-0089-03

Research on Wikipedia-based multi-document summarization system

Liu Maofu, Yu Bo, Hu Huijun
(Computer Science and Technology Institute, Wuhan University of Science and Technology, Wuhan 430065, China)

Abstract: In this paper, the extractive multi-document summarization system based on Wikipedia is designed and implemented. We evaluate the final generated summaries with ROUGE, and the experiment results show that the Wikipedia can improve the quality of the final summary to a considerable extent.

Key words: multi-document summarization; Wikipedia; sentence extraction

在互联网搜索应用中，搜索引擎按照网页内容与用户查询主题的相关度线性排序并返回结果，往往会提供数量庞大、信息重复的多个页面集给用户，有时也会在新闻页面的末尾处提供多篇相关报道的链接。为了方便用户从海量信息中准确快速地获取用户想要的信息，针对返回的反映不同主题的页面集，可以利用多文档自动摘要技术，为每个包含多个相关文档的页面集自动生成一篇摘要并提供给用户，帮助用户进一步聚焦到真正需要的文档集合上。对生物学文献统计后发现，对文本进行人工标引时，42.7%的主题词从原文中产生，47%的主题词可以由原文中词语的同义词得到。根据这种分布规律，可以从文本中直接或间接抽取语句生成摘要[1]，因而出现了抽取式多文档自动摘要技术。

在对多文档生成摘要的过程中，相关联的文档因共同的关键词联系在一起，若能提供有关这些关键词的背景信息，对准确理解多文档的内容，提高摘要的质量将会提供很大帮助。例如，现在有很多关于中国政府在利比亚动乱冲突中撤侨行动的新闻页面，用户如果要为这些新闻页面所形成的文档集生成摘要，则可以利用利比亚冲突提供的背景信息对页面内容进行过滤，这样生成的摘要中将会只保留北非利比亚的撤侨行动，而将日本福岛核泄露事故中的撤侨行动内容过滤掉，这样生成的摘要内容就会更精确。

维基百科是目前世界上最大的面向互联网开放式的多语种百科全书，它的基本组成单元是“词条”，每一个词条都对应一个维基页面。根据 BBC 报道，通过测验证实，维基百科在科技方面与《大英百科全书》一样准确[2]。因此，本文利用维基百科作为提供背景信息的外部资源。对于给定的关键词搜索得到其对应的维基百科页面，选取与文档集主题关联度高的那部分内容，通过与文档集比对来缩小摘要句的选取范围，并用这部分内容对文档集里的句子进行过滤，提高为主题生成的最终摘要的精确度。

1 系统描述

本系统模型采用多文档自动抽取式摘要方法，将文本看作句子的线性组合，将句子看作词的线性组合[3]。计算句子 TF*IDF、句子位置、句子与主题相似度以及句子长度四个特征项的值，将各特征项权值按照组合优化

* 基金项目：湖北省自然科学基金(2009CDB311)，湖北省教育厅人文社会科学基金重点项目(2011jyte126)，国家自然科学基金重大研究计划(90820005)

后得到的结果对句子排序，用搜索出的维基百科内容对句子做相似度计算并降序排列，然后抽取摘要句。

系统包括 4 个模块，即文档预处理、独立特征计算和组合优化、基于维基百科的摘要句过滤和摘要生成。如图 1 所示。

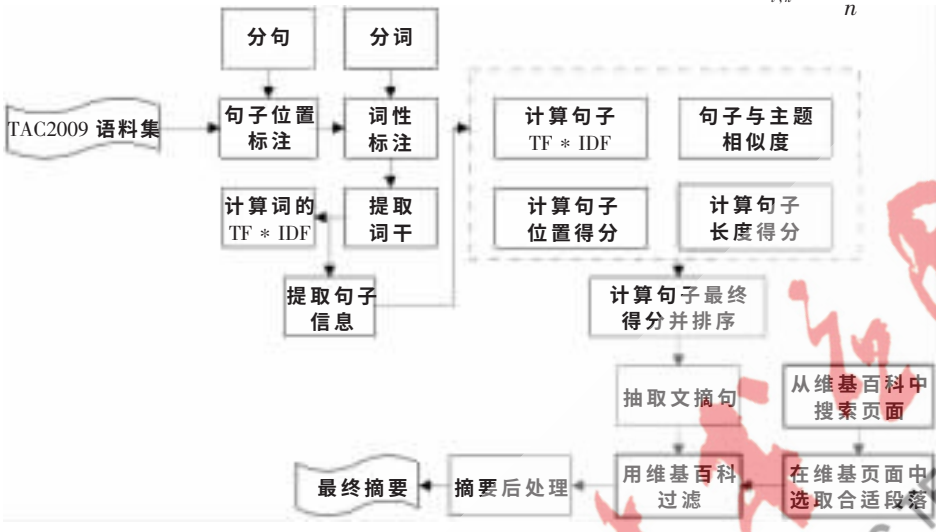

图 1 基于维基百科的多文档自动摘要系统模型

1.1 文档预处理

文档预处理的主要任务是基于分词词典与停用词表，切分文本中的词语并标注词性，进而统计词频并记录位置等词的基本信息。本系统将语料中每一个文档分割为句子单元，在预处理过程中利用 GATE 作为分词工具[4]。GATE 是一个应用广泛的信息抽取的开放型基础架构，为用户提供图形化的开发环境，被许多自然语言处理项目尤其是信息抽取研究项目所使用。

将语料文档用 GATE 分词、提取文档中每个单词的词干、标注每个词的词性以及停用词过滤，然后计算每个词的 TF*IDF 值，生成预处理文件。

1.2 特征选取和组合

根据预处理的结果，计算每个句子的四个特征项，即句子 TF*IDF、位置、与主题相似度以及句子长度。

(1)句子 TF*IDF

TF*IDF 是短语在文档中出现的频率和在全体语料中出现的文档频率的倒数之积[5]。也就是说，在本文档中出现比较频繁而在其他文档中不常出现的术语具有更高的信息量。该特征表示若句子包含的文档中重要的单词越多，该项得分越高。该项得分为句子中除去停顿词后所有单词的 TF*IDF 值的总和。

$$ST_{i,k}=\sum_{w\in Sen_{i,k}}(TDoc_{i,k}(W)+TTopic_{i,k}(W)) \tag{1}$$

其中，$Sen_{i,k}$ 表示第 k 个文档中第 i 个句子，w 表示 $Sen_{i,k}$ 中的非停顿词。$ST_{i,k}$ 表示 $Sen_{i,k}$ 的 TF*IDF 特征项的得分。$TDoc_{i,k}$、$TTopic_{i,k}$ 分别表示单词 w 在文档中的 TF*IDF 得分以及在主题信息中的 TF*IDF 得分[6]。

(2)句子位置

系统将文档内容看作是句子的线性组合，每篇文档中第一句话是最重要的，其他句子重要性按位置依次向后递减。

$$P_{i,k}=\frac{n-i+1}{n} \tag{2}$$

其中，$P_{i,k}$ 表示 $Sen_{i,k}$ 的位置特征项的得分；n 表示第 k 个文档的句子总数。

(3)句子与主题相似度

主题是文档集中讨论的中心内容，每个句子与主题相似度越大则表明该句包含的重要信息越多，句子就越重要。

$$S_{i,k}=SSenTopic_{i,k}+SDocTopic_{i,k} \tag{3}$$

其中，$SSenTopic_{i,k}$ 表示句子与主题标题句直接相似度得分，$SDocTopic_{i,k}$ 表示句子与主题标题句间接相似度[7]。

(4)句子长度

采用正态分布模型计算该特征项的得分。句子的长度越接近全部文档句子的平均长度，该特征的得分就越高。句子的平均长度是同一主题下的所有文档中的单词的总和除以句子总和的值。因此该特征项的得分为：

$$L_{i,k}=\frac{1}{\sqrt{2\pi\sigma}}e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \tag{4}$$

其中，μ 为同一文档集中所有句子的平均长度；x 为 $Sen_{i,k}$ 中包含单词的个数。

本文中给每个特征项设置一定的权重，每个特征项的最后得分为该特征项的得分乘以权重，句子的最终得分为每个特征项的最后得分的总和[8]。

1.3 基于维基百科的摘要句过滤

由于文档集 A 和文档集 B 都是围绕相同的主题展开，利用维基百科可以获得共同主题的维基页面内容。一般维基页面会在页面顶部对词条给出一个综述性介绍。本系统把维基百科的词条综述性介绍看成是维基百科词条页面自身的一个“特殊”的单文档摘要，利用这个“特殊摘要”对抽取得到的摘要句进行句子过滤计算所生成的多文档摘要中的句子和特殊摘要中的句子的相关度，将相关度低于预期阈值的句子删除掉；最后对剩余摘要句进行处理后生成最终摘要。

以 TAC2009 提供的文档集 D0901A 为例，其标题是“Indian Pakistan conflict”，其文档集中每篇文档的内容都紧紧围绕着“印巴克什米尔冲突”这一主题展开。因此，在维基百科中可以得到“Kashmir conflict”词条页面，其中，页面顶部是对克什米尔冲突以及双方各自立场的简

要介绍，下面分别按照冲突的时间线索、冲突的背后原因、当地人权状况和最新进展等四大方面做了详细描述。

在利用维基百科生成摘要时，首先从44个不同主题的文档集里找出合适的关键词或短语。本文选取的都是文档集中自带主题的关键词，输入到英文版的维基百科中搜索得到维基页面；在每个词条页面中选取页面顶部概述部分的前几小段内容作为词条的背景信息，修改成和语料相同的XML格式，保存为后续工作中将会使用到的集合 W。

1.4 生成摘要

生成摘要时，将语料集 A 输入到摘要系统计算出语料集 A 中每个句子4个特征值的得分，根据特征项组合优化的结果对句子降序排列得到摘要句的集合 A'，再将 A' 中每个句子与由维基百科生成的集合 W 中每一个句子做相似度计算，按照相似度值降序排列，根据摘要的长度限制按照一定的顺序选取句子组成摘要。

得到摘要后，还要对其进行后处理。简单的摘要后处理做法是遵循语法和习惯用语所生成的一些规则。本文共制定了11条规则来消除非限制性定语从句、时间短语从句等非重要信息，并使用正则表达式表示这些规则。最后从后处理得到的摘要句集合中根据摘要长度要求选取句子组合成正式摘要。

2 实验结果分析

2.1 实验准备

本文使用的语料来源于TAC2009，TAC2009共提供了44个不同主题的新闻文档集，每个主题的文档集都分为 A 和 B 两个集合，每个集合都包含10篇从全球四大新闻社选取的新闻材料。与此同时，TAC还组织了8位专家为44个语料集的每一个主题人工写了8篇摘要，并从中选取4篇摘要作为TAC人工评测对比的模板。

在44个主题的文档集中，并不是所有的主题都能在维基百科中定位到一个具体的页面，有一些由于指向模糊，在维基百科中无法提供有效参考页面。在所有的44个主题中，最终定位出40个主题的词条。正是由于维基百科极广泛的覆盖面，保证了90%的主题可以从维基百科中得到背景信息。对于与著名人物相关的主题，例如“美国前副总统切尼枪击误伤事件”和“迈克尔杰克逊亵童案”，或者生活类主题，例如塑料垃圾袋、太阳能、处方止痛药等，都很容易搜索出具体页面。而在事件类的新闻中，如果不是轰动一时的国际事件，例如“泰科(Tyco)前CEO科兹洛夫诈骗案”，则很难从维基百科中得到详细有效的信息。

而在能提供出具体页面的40个主题中，有少部分需要在维基百科中转义成另外的词条，例如，“世贸大厦纪念馆”应看做“911国家纪念和博物馆”的转义词条，“印航炸弹案嫌犯审判”在维基百科中则应该转义为“Air India Flight 182”词条。

2.2 实验结果

本文使用在自动摘要领域内广泛应用的ROUGE(Recall-Oriented Understudy for Gisting Evaluation)作为自动评测工具。ROUGE是一种基于要点召回率的评测方法，它通过考察专家摘要与机器摘要中相同评价单元(如n-gram、词序、词对等)的重叠数量来达到对文档质量进行自动评测的目的。TAC2009中采用的评价指标有ROUGE-2和ROUGE-SU4。表1给出了使用和没使用维基百科的两组自动文摘在ROUGE-2和ROUGE-SU4下的得分，以及两组自动文摘在人工评测下的得分。

从表1可以看出，不论是更新前的摘要还是更新后的摘要，在ROUGE-2和ROUGE-SU4两项评测中，使用了维基百科的自动摘要得分均高于没有使用维基百科的自动摘要，而且提高的幅度比较明显。相比ROUGE-SU4指标得分，ROUGE-2中的得分提高幅度更大一些。原因可能是44个主题的文档集中的绝大部分都能在维基百科中找到紧密关联的背景信息，这归功于维基百科越来越庞大的词汇量和中立客观的态度。

在人工评测部分，使用了维基百科的结果比没有使用维基百科的结果也有较大幅度提高。这说明引入维基百科的想法是正确的，引入这种外部资源确实能提高摘要内容和主题的关联度。但是，维基百科在提高摘要质量的同时，选取错误的维基页面内容也会对结果产生负面的影响。

仍然以“印巴冲突”为例，文档集中的内容围绕着在印巴冲突中双方对和平所作的努力，即“Efforts made toward peace in the India-Pakistan conflict”。而在维基页面中将之定位在页首第一部分，即“武装冲突”的叙述，却忽视“和平”这个最主要的关键词，因而适得其反。在维基百科的“Kashmir conflict”词条页面中，实际上可以在

表1 摘要在ROUGE评测和TAC人工评测中的对比结果

	没有使用维基百科的摘要	使用了维基百科的摘要
文档集 A 的ROUGE-2平均值	0.028 23(0.021 93-0.035 43)	0.051 86(0.044 78-0.059 26)
文档集 B 的ROUGE-2平均值	0.026 39(0.020 32-0.033 07)	0.042 66(0.036 26-0.050 60)
文档集 A 的ROUGE-SU4平均值	0.061 45(0.047 97-0.070 73)	0.090 82(0.083 35-0.098 69)
文档集 B 的ROUGE-SU4平均值	0.065 62(0.058 15-0.073 63)	0.084 57(0.077 58-0.092 35)
文档集 A 的人工评测得分	0.062 1.023 0.227 0.061 3.636 2.455	0.133 2.227 0.409 0.130 3.614 2.864
文档集 B 的人工评测得分	0.050 0.795 0.136 0.049 3.659 2.227	0.081 1.227 0.068 0.081 3.909 2.636

该页面中的“Recent developments”这一部分下找到“Efforts to end the crisis”这一符合语料要求的内容。由此可以看出，在维基百科中准确定位相关主题的内容是很重要的，这也是需要进一步研究的环节。

本文首先提出了一种通过引入维基百科作为外部资源来提高自动摘要质量的方法。描述了本自动文摘系统的4个模块，即文档预处理、独立特征计算和组合优化、基于维基百科的摘要句过滤以及摘要生成。作为对比，本文使用本系统对同一批语料在引入和不引入维基百科的条件下生成摘要，最后用ROUGE工具进行评测。实验结果证明，使用了维基百科的结果要明显优于没有使用维基百科的结果，这说明在自动摘要系统中适当地引入外部知识库(例如维基百科或者百度百科)可以有效提高生成摘要的质量。

参考文献

[1] LIN C Y.ROUGE：a package for automatic evaluation of summaries[C].In Proceedings of the Workshop on Text Summarization，Barcelona.ACL，2004.

[2] 周庆山，王京山.维基百科信息自组织模式探析[J].情报资料工作，2007(02)：29-32.

[3] SRAVANTHI M，CHOWDARY C R，KUMAR P S.QueSTS：a query specific text summarization system[C].In Proceedings of the 21st International FLAIRS Conference，Florida，USA.AAAI Press，2008.

[4] MIHALCEA R，TARAU P.TextRank：bringing order into texts[C].Proceedings of EMNLP.Barcelona，Spain：Association for Computational Linguistics，2004.

[5] 郭燕慧，钟义信，马志勇，等.自动文摘综述[J].情报学报，2005，21(5)：582-591.

[6] 徐超，王萌，何婷婷，等.基于局部主题关键句抽取的自动文摘方法[J].计算机工程，2008，34(22)：49-51.

(收稿日期：2011-04-25)

作者简介：

刘茂福，男，1977年生，研究生导师，副教授，主要研究方向：自然语言处理、分布并行处理、图像处理和图像挖掘。

余博，男，1985年生，硕士研究生，主要研究方向：多文档自动摘要。

胡慧君，女，1976年生，讲师，主要研究方向：图像处理。