文章编号: 1003-0077(2008)03-0081-08

# 自动文摘评价方法综述

张 瑾[1,2],王小磊[1,2],许洪波[1]

(1. 中国科学院 计算技术研究所 信息智能与信息安全研究中心,北京 100190;
2. 中国科学院 研究生院,北京 100190)

**摘 要**:评价是自动文摘领域长期关注的焦点,对自动文摘技术的发展起着积极的促进作用。本文首先介绍了自动文摘评价方法的应用背景和面临的困难;然后对自动文摘评价方法进行了简单介绍和评价;接着在了解国内外研究现状的基础上详细分析了文摘评价方法的关键技术;最后对自动文摘评价方法未来的发展趋势进行了展望。

**关键词**: 计算机应用; 中文信息处理; 文本挖掘;自动文摘;自然语言处理;多文档文摘; 文摘评价方法
**中图分类号**: TP391 **文献标识码**: A

## Survey of Automatic Summarization Evaluation Methods

ZHANG Jin[1,2], WANG Xiao-lei[1,2], XU Hong-bo[1]

(1. Research Center of Information Intelligence and Information Security, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China;
2. Graduate University of Chinese Academy of Science, Beijing 100190, China)

**Abstract:** Evaluation has long been of interest to automatic summarization circle because of its effective promotion to the summarization progress. After a discussion of the background of summarization evaluation and the unsettled issues, this paper briefly introduces and comments on the current summarization evaluating methods. It further provides a detailed analysis of key technologies in the existing evaluation method. And finally, it presents some directions for future research.

**Keywords:** computer application; Chinese information processing; text mining; automatic summarization; nature language processing; multi-document summarization; summarization evaluation

## 1 引言

自动文摘技术用于自动从一篇或多篇文章中提取满足用户或应用需求的内容,加以组织后生成一篇内容完整、形式严谨的自动文摘[1]。它可以帮助人们在海量信息中准确、高效地寻找自己需要的信息,发展至今,已经得到了广泛的应用。自动文摘评价方法是自动文摘技术研究与发展的一个关键部分,规范合理的评价标准可以促进自动文摘技术的发展。但同时,自动文摘评价方法也是最具争议的,至今仍面临着许多挑战[2,3]:

· 人工评价成本高、耗时长,并且主观性强,一直以来很难对自动文摘系统开发提供切实有效的帮助,研究者更倾向于可以客观高效地进行自动文摘的自动评价方法。

· 有时文摘系统生成了一篇很好的文摘,却与作为评价标准的人工文摘相差甚远,给文摘评价带来了很大的困难,因此人工标准文摘的公平性也是自动文摘评价方法研究的一个难点。

**收稿日期**: 2007-06-06 **定稿日期**: 2007-09-24
**基金项目**: 国家 973 计划"大规模文本内容计算"课题资助项目(2004CB318109)

**作者简介**: 张瑾(1978—),男,博士生,主要研究方向为大规模文本处理,文本挖掘,多文档文摘技术; 王小磊(1984—),男,硕士生,主要研究方向为文本挖掘,特征选择;许洪波(1975—),男,博士,副研究员,主要研究方向包括文本挖掘,互联网搜索,信息过滤等。

· 文摘系统通常按照不同的压缩率来生成相应文摘。在不同压缩率下,文摘包含的原文档中的信息量是不同的。如何在评价中反映这一变化,也增加了自动评价的难度和复杂性。

· 文摘通常是面向需求的,应该根据用户或应用的需求包含特定的信息。如何在进行评价时也对用户或应用的需求加以考虑,这也使自动评价变得更加复杂。

为了解决这些问题,许多文摘自动评价方法都随着自动文摘技术的发展而被一一提了出来。

## 2 自动文摘评价方法的分类

1998 年 K. S. Jones[4] 提出从广义的角度将自动文摘的评价方法大致分为两类: 一种称为内部评价(Intrinsic)方法,它通过直接分析摘要的质量来评价文摘系统,内部评价主要评价文摘的连贯性和内容的完整性。另一种称为外部评价(Extrinsic)方法,它是一种间接评价方法,将自动文摘应用于某一个特殊的任务中,根据文摘完成这项任务的效果来评价自动文摘系统的性能,外部评价是测试文摘对自动问答、分类等任务的影响程度。

内部评价中文摘的连贯性主要是指文摘在文字上的流畅程度,包括是否出现主语悬挂,句子是否通顺,句子间语义是否连贯,句子间是否有关联词连接,逻辑结构是否合理等,主要采用主观性感觉进行评价。内部评价中文摘内容的完整性是指文摘中包含原文(或标准文摘)中的信息量的多少。常用的两种评价方法: 一种是以原文为参考,原文经过加工、标注,为评价提供判定依据;另一种是将专家根据原文生成的文摘作为标准文摘,来判断生成的自动文摘中所包含标准文摘中的信息程度。第二种评价方法是目前自动评价技术研究的热点。

为了克服内部评价方法的缺点,研究者陆续提出了一些外部评价方法,即通过一个具体的任务来评价系统的性能。很多不同的任务都可以作为外部评价的载体,例如 GMAT 测试、信息检索、自动问答等。1995 年,Brandow[5] 通过在 IR 任务中,利用检索的准确度来评价单文档文摘系统。1998 年,美国国防部高级研究计划署在 TIPSTER 文本计划项目下进行了一次自动文摘系统测试,具体方法是对 TREC 文本集的每篇文章生成文摘,然后根据文摘对原文章进行分类,将分类的准确度作为评价标准。

相对于内部评价方法来说,外部评价方法具有较少的主观性,易于对多个文摘系统进行对比,也有助于自动文摘在其他领域中的应用研究。但外部评价方法不足之处在于: 每次测评只是针对一个特定任务,有一定的局限性,不利于系统性能的全面改进。由于内部评价方法可以直接对文摘进行评价,并且独立于应用环境,所以现在仍是研究的重点。进一步考虑,人工评价有着很大的不确定性,不同的人对同一个文摘的质量好坏的判定不完全一样。而且人工评价代价太大,不利于大规模的对多个系统进行评价。自动评价技术才是现在研究的主流。但是有一些方面难以定量测试,如文摘是否连贯、表述是否合理等,只有让评价者将文摘与原文章(或标准文摘[6])做对比,根据主观感受做出评价。研究表明,这种方法可以作为自动评价方法的一个有益的补充[7]。

## 3 自动文摘评价相关理论

评价机器文摘包含标准文摘中信息的程度是内部评测中对文摘内容完整性的一种评测。机器文摘包含标准文摘中的信息内容越多,机器文摘与标准文摘越相似。因此,可以将判断机器文摘中包含标准文摘的多少转换成判断两个文本之间的相似程度,利用向量空间模型[8] 来对文摘内容的完整性进行自动评价。统计相关性是评估两变量的观测数据之间是否存在相关关系的一种手段。因此,可以采用相关性分析理论,利用 Spearman 等级相关系数和 Pearson 相关系数[9] 来验证机器文摘与标准文摘之间的一致性。

### 3.1 向量空间模型和余弦相似度

考虑一个文档空间由文档向量 $D_i$ 构成。每个文档向量 $D_i$ 由多个索引项 $T_j$ 构成。这些索引项可以根据它们的重要性赋以权重。假设共有 $t$ 个索引项,则每个文档向量 $D_i$ 可以表示为一个 $t$ 维向量 $(d_{i1}, d_{i2}, \cdots, d_{it})$。其中 $d_{ij}$ 表示第 $j$ 个索引项 $T_j$ 的权重。如果给定两个文档的对应向量 $D_1$ 和 $D_2$,就可以计算它们之间的相似度 $Sim(D_1, D_2)$。这个相似度可以用两个向量的内积来测量,就得到所谓的余弦相似度:

$$Sim(D_1, D_2) = \frac{\sum_{i}^{t} d_{1i} d_{2i}}{\sqrt{\sum_{i}^{t} d_{1i}^2 \sum_{i}^{t} d_{2i}^2}} \quad (1)$$

### 3.2 相关分析理论

一种评价方法产生的数值化分数,可以用于比较对相同文档产生的不同文摘。假设两个系统对一篇或一组文档产生了两个文摘,其中得分高的文摘预示着产生它的系统比产生另一个文摘的系统好,那么如果一个系统相对于另一个系统能连续地产生高分的机器文摘,则可以认为这个系统比另一个系统好。因此文摘评价的一个重要特征是在给定的测试文档集上系统的得分排序,而不是它们的具体分数。在此基础上,可以采用相关分析的方法统计这些评价结果之间的统计关系的强弱程度,检验不同评价方法的结果是否一致。

#### 3.2.1 线性相关分析

线性相关分析作为描述两变量间是否有直线关系以及直线关系的方向和密切程度的分析方法。通常认为,正相关程度越高,评测结果的一致性越好。线性相关分析要求两个变量服从正态分布。一般说来,两个变量都是随机变动的,不分主次,处于同等地位。两变量 $X$ 和 $Y$ 间的线性相关关系用 Pearson 相关系数描述。其定义如下:

$$r_p = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \qquad (2)$$

其中 $n$ 为样本数, $x_i$ 和 $y_i$ 分别为两变量的变量值; $\bar{x}$ 和 $\bar{y}$ 分别为两变量的均值。

两变量的线性相关程度与 $|r_p|$ 成正比。$r_p = 1$ 表示线性正相关; $r_p = -1$ 表示线性负相关; $r_p = 0$ 表示线性无关, Pearson 相关系数反映了两变量间线性关系的强度和方向。

#### 3.2.2 等级相关分析

若两变量不服从正态分布或总体分布未知,则可采用等级相关分析。它适用于非正态总体或总体分布未知;数据一端或两端有不确定值的变量或等级变量。与线性相关分析类似,等级相关分析也可以用于检验评测结果的一致性。等级相关分析的方法有多种,在此仅介绍 Spearman 等级相关分析。计算 Spearman 等级相关系数时不直接采用原始数据 $(x_i, y_i)$,而是用两变量的秩 $(R_i, Q_i)$ 来代替。两变量 $X$ 和 $Y$ 间的 Spearman 等级相关系数定义如下:

$$r_s = \frac{\sum_{i=1}^{n}(R_i - \bar{R})(Q_i - \bar{Q})}{\sqrt{\sum_{i=1}^{n}(R_i - \bar{R})^2 \sum_{i=1}^{n}(Q_i - \bar{Q})^2}}$$

$$= 1 - \frac{1}{n(n^2 - 1)} \sum_{i=1}^{n}(R_i - Q_i)^2 \qquad (3)$$

两变量的等级相关程度与 $|r_s|$ 成正比。$r_s > 0$ 表示等级正相关; $r_s < 0$ 表示等级负相关; $r_s = 0$ 表示等级无关, Spearman 等级相关系数是两变量 $X$ 和 $Y$ 等级间是否相关的一种测度。

### 3.3 Kappa 分析

相关分析在进行一致性检验时,有其局限性和不足。相关系数仅表示相关,并不表示真正一致;甚至在某些情况下应用不同的相关分析方法对同一批测定结果进行分析可能得出完全相反的结论。即它不能确切地综合反映评测结果之间的一致性。

Kappa 统计量是比较两个或多个观测者对同一事物,或观测者对同一事物的两次或多次观测结果是否一致,以由于随机因素造成的一致性和实际观测的一致性之间的差别大小作为评价基础的统计指标。Kappa 统计量和加权 Kappa 统计量不仅可用于无序和有序分类变量资料的一致性、重现性检验,且能给出一个反映一致性大小的"量"值。一般认为,若 $0.75 < k < 1$,说明一致性极好; $0.40 < k < 0.75$,一致性好; $0 \leq k \leq 0.40$,则一致性差。Kappa 统计量的定义为:

$$Kappa = \frac{P_o - P_e}{1 - P_e} \qquad (4)$$

其中 $P_o$ 为实际一致率, $P_e$ 为期望一致率。

## 4 国内外研究现状分析

近年来,关于自动文摘的专题研讨会纷纷出现在世界知名的会议——ACL[10]、COLING[11] 和 SIGIR[12]中。其中由美国国家标准与技术协会(The National Institute of Standards and Technology, NIST)[13] 支持的文档理解会议(Document Understanding Conference, DUC)[14],日本的搜索引擎评测型国际会议(NII Test Collection for IR Systems, NTCIR)[15] 主办的文摘挑战会议(Text Summarization Challenge, TSC)[16] 和 COLING-ACL 主持的多语言文摘专题讨论会(Multilingual Summarization Evaluation, MSE)[17] 在探讨自动文摘技术的同时,也为研究者提供了一个标准的文摘训练和评价平台,以便对参赛系统进行大规模的评测,从而推动自动文摘技术的发展。

在国内,由于缺乏大规模统一的测试集和测试

平台,同时由于相关自然语言资源的不成熟,中文自动文摘评价方法的研究还处于起步阶段,只有少数科研机构参与了相关工作。1995年,863专家组组织了一次国内单文档自动文摘系统评价,并发布了自动文摘测试大纲[18]。由于直接人工比较难度较大,1997年北京大学的俞士汶等人提出了一种机械式文摘质量的自动评价方法[19],基本采用了Edmundson的句子重合率的方法[20]。2001年上海交通大学的沈洲等人[21]提出了一种参照Turing测试的思想进行自动单文摘系统评价的方法。该方法是一种人工评价方法,具体方法是将人工文摘和自动文摘混合,交由专家组进行黑箱评价,根据自动文摘和人工文摘的排名结果评测自动文摘系统的性能。2004年华中师大的何婷婷等人[22]提出了一种内部评价方法,其中使用了主题覆盖度和内容冗余度两个指标。2005年哈尔滨工业大学的张姝等人[23]提出了基于向量空间模型文本相似度的评价方法,采用Spearman等级相关系数和Pearson等级相关系数对三种权重选择方法——基于词频tf的余弦相似度,去停用词的基于词频tf的余弦相似度和基于词tf-idf权重的余弦相似度——的一致性进行考察与验证。总的来说,中文领域相关的文章发表得不多,而且主要是针对单文档自动文摘评价方法,还有较大提高空间,值得进一步深入的研究。

DUC(文本理解会议)是目前在多文档文摘领域最有影响的评测会议,由NIST的系列会议之一TIDES (DARPA's Translingual Information Detection, Extraction, and Summarization program)赞助发起。DUC使研究者共同参与到大规模文本测试中来,促进了自动文摘包括多文档文摘的发展。DUC会议自2001年起每年举办一次,目的是针对单文档文摘和多文档文摘进行评价。随着人们需求的变化和各项技术的日益成熟,DUC从任务到测试文档以及评价方法都日益丰富和成熟。所有的参与者可以在大规模公共语料上进行评测,表明多文档文摘的研究正在向规范化、统一化方向发展[24]。但是由于DUC没有针对中文的语料,如何进行中文多文档自动文摘的评价,从而客观地衡量系统生成的文摘的质量,是中文文摘研究领域亟须解决的问题。

### 4.1 基于准确率和召回率的方法

主要思想是由人工生成一篇标准文摘(又称人工文摘),计算自动文摘中包含了标准文摘的多少句子,以此作为依据来评价自动文摘的质量[25,26]。如果标准文摘的长度为 $n$ 个句子,自动文摘的长度为 $k$ 个句子,并且有 $p$ 个标准文摘中的句子包含在自动文摘中,则准确率定义为: $Precision = p/k$;召回率定义为: $Recall = p/n$。

$F\text{-}Measure$[27]是一个对文摘的准确率和召回率综合考察的指标,定义为:

$$F\text{-}Measure = \frac{2 \times P \times R}{P + R}$$

其中 $P$ 为准确率, $R$ 为召回率。

考虑到标准文摘中的句子的重要程度不同,Sentence Rank方法通过对所有文档句按照其重要程度进行排序,然后可以应用相关度测量将自动文摘与标准文摘进行对比。

文摘的准确率和召回率是两个相互关联的指标。通常,系统的文摘召回率不会随着准确率提高而提高,反而可能会下降。因此,只用其中任一个指标来评价都未必理想。由于基于准确率和召回率方法只考察了句子是否相同,而忽视了句子内容本身的相似性。因此对于文摘句不同,而内容非常相似的两篇自动文摘会给出完全不同的评价结果。F-Measure方法综合考虑了文摘的准确率和召回率,但并没有本质的改进,不能完全克服原有的缺点。而Sentence Rank方法考虑了文摘句的重要程度,评测结果比单独考察一个指标要精确。

### 4.2 基于一致性评价的方法

基本方法是由几位专家根据原文章内容,直接对自动文摘中的文摘句评分,然后计算自动文摘的总得分。1996年,Carletta[28]首先把Kappa统计量[29]引入到分类任务的一致性评价中。Kappa方法可以对多位专家的意见进行一致性评价,缺点是只能对文摘句做二值判断,忽视了部分正确性。2000年,密西根大学的Radev和Tam提出了Relative Utility方法[30]。Relative Utility方法是一种改进的Kappa方法,可以对多个评价结果进行一致性判定。根据相对效用的思想,Relative Utility方法可以对两篇句子完全不同,但效果近似的自动文摘给出更一致的结果。基本思想是:多个专家同时给一篇自动文摘的文摘句打分,某个专家的最终评分是他与所有专家评分一致程度的平均。Relative Utility方法克服了Kappa方法只能做二值判断的缺点,可以评价文摘句的部分正确性。Relative Utility方法与Majority Vote方法[31]很相似,与基于

准确率和召回率的方法相比,它可以对不同压缩率的自动文摘进行评测。

### 4.3 基于内容相似度判别的方法

此前方法对文摘的评价都停留在句子的粒度上,这些方法过于粗糙,并不能正确反映自动文摘包含原文章信息的程度。因为句子中可能包含冗余信息,有时不同的句子中可能包含相同的信息。基于内容相似度的评价方法是对文摘内容完整性的一种评价,相比直接对文摘句进行打分的方法更准确。2002 年,英国谢菲尔德大学 Saggio 等人[32]提出了三种基于文摘内容相似度的自动评价方法,分别是基于余弦相似度(Cosine)、单元覆盖(Word Overlap)和最长公共子串(Longest Common Subsequence, LCS)的方法。

### 4.4 基于语义单元重合度的方法

为了更好地判断自动文摘包含标准文摘中信息的程度,一些研究者提出首先对标准文摘进行语义分析,取出其中的要点,即语义单元,根据自动文摘对这些语义单元的覆盖度来评价。2003 年,英国剑桥大学的 Teufel 和荷兰 Nijmegen 大学的 Van Halteren 提出了 Factoid 方法[33]。即用人工标注的方式将每个标准文摘句的内容表示为一组语义单元(Factoid),每个 factoid 包含一个词语或一个分句的信息,不同文摘句的相同内容对应同一组 factoid。自动文摘的质量则根据其包含 factoid 的数量来判断。

基于内容相似度判别的方法与基于语义单位重合度的判别方法的主要差别在评价时所选用的基本单元上,基于内容相似度主要选用单词和子串作为判别单元,而后者选用的是语义单元来作为评价的判别单元,然后再采用余弦相似度和单元覆盖等来进行评价。

## 5 主流评价方法

如前所述,停留在句子粒度的评价方法不够精确;而基于词的划分也不令人满意: 因为一个词在不同的上下文环境中的作用和重要性不尽相同,不具有可比性。如何划分文摘评价单元是当前面临的主要问题之一。而另一个困难是如何利用内容相似度和语义重合度等对这些单元进行比较,从而对自动文摘质量进行有效判定。因为不同单元的长度和重要性差别很大,不易评价。现在的主流评价方法在这两方面进行了有益的探索。

### 5.1 SEE

2001 年,美国南加州大学的 Chin-Yew Lin 开发了一个单文档文摘评价系统 SEE[34](Summary Evaluation Environment)。该系统首先根据评价的粒度将自动文摘和标准文摘打散成一系列单元(句子,分句等),通过计算自动文摘单元对标准文摘单元的覆盖程度,来评价自动文摘的质量。

### 5.2 ROUGE

2004 年, Chin-Yew Lin 等人参考了机器翻译的自动评价方法 BLEU[35],提出了 ROUGE(Recall-Oriented Understudy for Gisting Evaluation)评价方法[36]。该方法首先由多个专家分别生成人工文摘,构成标准文摘集。然后也是将系统生成的自动文摘与人工生成的标准文摘相对比,通过统计二者之间重叠的基本单元(*n* 元语法、词序列和词对)的数目,来评价文摘的质量。通过多专家人工文摘的对比,提高评价系统的稳定性和健壮性。该方法现已成为文摘评价技术的通用标准之一。

ROUGE 主要包括以下四种评价标准:

1) ROUGE-N 基于 n-gram 共现统计。

2) ROUGE-L 基于最长公共子串。

3) ROUGE-S 基于顺序词对统计。

4) ROUGE-W 在 ROUGE-L 的基础上,考虑串的连续匹配。

研究表明[37]: (1) ROUGE-2, ROUGE-L, ROUGE-W 和 ROUGE-S 用于单文档文摘评价效果很好, (2) ROUGE-1, ROUGE-L, ROUGE-W, ROUGE-SU4 和 ROUGE-SU9 在评价短文摘时结果令人满意, (3) 对于多文档文摘评价,各种方法难以达到很高的一致性。但是,如果在对自动文摘和标准文摘进行匹配时排除了停用词的干扰,那么 ROUGE-1, ROUGE-2, ROUGE-S4, ROUGE-SU4 和 ROUGE-SU9 的表现也很不错, (4) 通过使用标准文摘集而非单个标准文摘可以提高评价结果的一致性。

### 5.3 Pyramid

2003 年,哥伦比亚大学的 Mckeown、Nenkova、Passonneau 等人共同提出了 Pyramid 方法[38]。首先,将文摘句人工划分为若干个文摘内容单元(Summarization Content Unit, SCU),每个 SCU 表

示一个核心概念。一个 SCU 被越多的标准文摘包含就越重要。将所有 SCU 按照重要程度排序,同等重要的 SCU 排列在同一行,由上向下重要程度逐行递减,构成所谓的“Pyramid”。通过计算自动文摘包含的 SCU 的数量和重要程度来判断自动文摘的质量。初步研究表明,Pyramid 与人工评价有较好的一致性。但是,由于各个语义单元的大小不固定,且同一语义的表述方式多种多样,致使自动生成这些语义单元存在很大困难。而且人工标注成本高,不利于大规模地对多个系统进行评价。

### 5.4 BE

为了解决 Pyramid 方法的问题,Chin-Yew Lin 等人又在 2005 年提出了 BE(Basic Elements)方法[39]。首先由机器自动生成标准文摘的较小的 n 元语法单元,然后对它们进行合并,实现自底向上的构造语义单元。这样便可以实现单元的自动识别,而且在一定程度上降低了匹配表示相同概念的不同语义单元的难度,这些基本单元被称为 BE。具体方法是构造一个句法分析器,然后生成一棵分析树,并定义一系列剪枝规则从分析树中抽取有效的 BE。但是目前 BE 的定义、打分策略以及匹配方法等问题还没有得到很好的解决,有待通过研究得以解决。

## 6 评价方法研究的发展趋势

在 2000 年左右,研究界就开始了对自动文摘评价方法发展的规划,表 1 是 Baldwin 等人于 2000 年提出的自动文摘评价研究的 5 年发展规划[40]。在 DUC 等的支持下,英文领域文摘评价方法的研究近年来取得了长足的进步,其中尤其以南加州的 ROUGE 和 BE 最为著名,二者已成为 DUC 国际评价体系中的代表方法。但是需要看到,目前的更新文摘(Update Task)的评价研究还仅处于起步阶段,作为今年的 DUC2007 的先导任务(Pilot Task)被提出来。结合研究路线图和当前研究现状可见,短期的研究重点将是单语言多文档的文摘评价方法和刚刚开始的更新文摘的评价方法等的研究,随着评价方法研究水平的进一步提高,多类型、多语言的文摘评价方法将成为未来研究的重点。

表 1 自动文摘评价研究路线图

| | | | 2001 | 2002 | 2003 | 2004 | 2005 |
|---|---|---|---|---|---|---|---|
| 建立测试语料集 | | | √ | √ | √ | √ | √ |
| 开发自动评价软件 | | | √ | √ | √ | √ | √ |
| 抽取式文摘的内部评价方法 | 单文档 | 简单类型 | √ | √ | √ | | |
| | | 复杂类型 | | | √ | √ | √ |
| | 多文档 | 单类型 & 单语言 | √ | √ | √ | | |
| | | 更新文摘 | | | √ | √ | |
| | | 多类型 & 多语言 | | | | √ | √ |
| 理解式文摘的内部评价方法 | 单文档 | 简单类型 | √ | √ | √ | | |
| | | 复杂类型 | | | √ | √ | √ |
| | 多文档 | 单类型 & 单语言 | √ | √ | √ | | |
| | | 更新文摘 | | | √ | √ | |
| | | 多类型 & 多语言 | | | | √ | √ |

另一方面自动评价方法与人工评价方法间的偏差已经引起人们的关注[41],如何减小两者的差异将是下一步研究的一个热点。中科院计算所的 Jin Zhang 等人在文献[42]中分别给出了 DUC2007[43] 中文摘系统在 ROUGE, BE 和多项人工评价指标上的得分的对比分析,为了方便对比,这里我们只选用了 Linguistic Quality 来代表人工评价方法与用 ROUGE-2 代表的自动评价方法进行对比,因为两者得分在数量级有一定差异,图中所示的自动评价得分是经过平滑后的 ROUGE-2 得分。图 1 所示的是 DUC2007 Main Task 中 30 个参赛系统在人工评价与自动评价上的差异分析。通过图 1 的差异分析实验结果表明,当前自动评价方法与人工评价方法间仍存在一定的差距。如何缩小人工评价方法与自动评价方法间的差异是文摘自动评价面临的一个难题,值得进一步深入地研究,以期能够最终找到一种

可以完全独立于人工评价的自动评价方法。

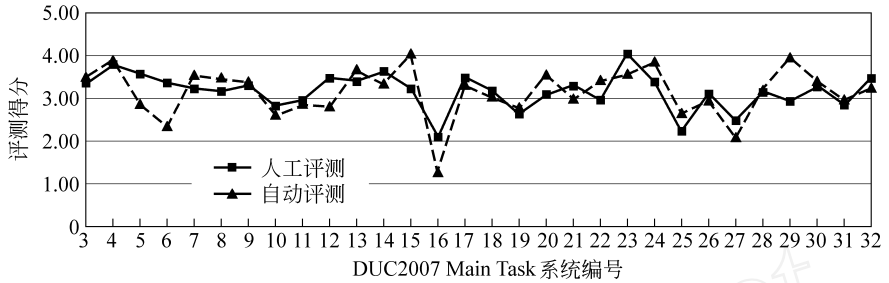


图 1　DUC2007 人工评价与自动评价差异分析

另外,如何去寻找一个更加合适的基本单元,并进行适当的划分也将是自动评价方法研究的一个重要方面。此外,随着中文文摘研究的深入,适合中文的自动评价方法的研究势必成为中文文摘领域的一个重要研究点,它将极大地推动中文文摘技术的迅速发展。

## 7　结语

在机器翻译和自动语音识别领域的经验表明:好的评测方法对系统的改进起到巨大的推动作用。在自动文摘领域,人们也在寻求一种与人工评测有很好一致性的自动评测方法。而目前的自动评测方法与人工评测方法之间还是存在很大差异,我们通过在 DUC2007 上对比人工评测方法和 ROUGE-2 方法,也证实了这一点。

尽管如此,人们在对自动文摘评价的研究过程中仍达成了一些共识,也取得了一些成果,尤其在英文文摘评价方法研究领域取得了较大进展。可以看到,自动评测方法的明显的发展趋势:考虑到人工文摘的不确定性,使用多篇人工文摘作参考;引入语义分析技术,解决同一内容的不同表示问题;给信息附加权重,更合理地反映信息的重要程度。然而自动文摘评价是一个相当复杂的问题,它所涉及的领域包括语言学、心理学、人工智能等多个学科,其实现还存在着很多困难,至今仍未能形成统一的标准,这也使自动文摘评价成为一个极具研究价值和挑战性的问题。尤其是中文多文档文摘评价领域,由于缺少统一的大规模测试集和评价平台,从而严重地制约了中文文摘技术的发展。如何有效地开展多文档文摘评价工作的研究,对推动中文多文档文摘技术的发展具有重要意义。

## 参考文献:

[1] Mani I. and Maybury M., eds. 1999. Advances in Automatic Text Summarization [M]. MIT Press.

[2] Mani I. Summarization Evaluation: An Overview [A]. In: Proc. of the NTCIR Workshop 2 Meeting on Evaluation of Chinese and Japanese Text Retrieval and Text Summarization [C]. Tokyo: National Institute of Informatics, 2001.

[3] Radev D., Teufel S., Saggion H., et al. Evaluation Challenges in Large-Scale Multi-Document Summarization [A]. ACL 2003 [C]. 7-12 July, Sapporo, Japan.

[4] Karen Sparck Jones, etc. Automatic Summarizing Factors and Directions Advance in Automatic Text Summarization [M], Cambridge MA: MIT Press: 1998.

[5] Brandow R., Mitze K. and Rau L. F. 1995. Automatic Condensation of Electronic Publications by Sentence Selection [J]. Information Processing and Management, 31(5): 675-685.

[6] Jing H., Barzilay R., McKeown K. and Elhadad M. Summarization Evaluation Methods: Experiments and Analysis [A]. In: Working Notes of the AAAI Spring Symposium on Intelligent Text Summarization [C]. 1998: 60-68.

[7] Mani I., House D., Klein G., et al. The TIPSTER SUMMAC Text Summarization Evaluation: Final Report [D]. MITRE Corp. Tech. Report, 1998.

[8] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing [J]. Commun. ACM, 1975: 613-620.

[9] 赵阳. 统计学原理 [M]. 北京: 中国财政经济出版社, 1993. 362-371.

[10] Hal Daume, III and Daniel. MarcuBayesian query-focused summarization[A]. In: Proceedings of ACL '06 [C]. 2006: 305-312.

[11] Hiroyuki Sakai and Shigeru Masuyama. A Multiple-document Summarization System with User Interaction [A]. In: Proceedings of COLING '04 [C].

2004: 1001.

[12] Sanda Harabagiu and Finley Lacatusu. Topic Themes for Multi-document Summarization [A]. In: Proceedings of SIGIR '05 [C]. 2005: 202-209.

[13] NIST, http://www.nist.gov/.

[14] DUC, http://duc.nist.gov/.

[15] Tsuneaki Kato, Mitsunori Matsushita and Noriko Kando. Expansion of Multimodal Summarization for Trend Information [A]. In: Proceedings of NTCIR-6 [C]. 2007.

[16] TSC, http://lr-www.pi.titech.ac.jp/tsc/index-en.html.

[17] Angelo Dalli, Roberta Catizone and Yorick Wilks. Clustering-Based Language Independent Multiple-Document Summarizer at MSE 2006 [A]. In: Proceedings of MSE 2006 [C]. 2006.

[18] 俞士汶,段慧明,田剪秋. 机械文摘自动评价的原理及实现 [A]. 智能计算机接口与应用进展 ——第三届中国计算机智能接口与智能应用学术会议论文集 [C],北京,清华大学出版社,1997: 230-233.

[19] 俞士汶,段慧明. 自动文摘评价报告 [J],计算机世界报, 1996 年 3 月 25 日:183.

[20] H. P. Edmundson. New Methods in Automatic Abstracting [J]. Journal of ACM, 1969, 16(2): 264-285.

[21] 沈洲,王永成,许一震,方澈. 自动文摘系统评价方法的研究与实践 [J]. 情报学报,2001,20(1):66-72.

[22] Po Hu, Tingting He, Donghong Ji, Meng Wang. A Study of Chinese Text Summarization Using Adaptive Clustering of Paragraphs [A]. In: Proceeding of Computer and Information Technology 2004 [C]. Wuhan, China, 2004: 1159-1164.

[23] 张姝,赵铁军,赵华,姚建民. 基于内容相似度的文摘自动评价方法及其有效性分析 [J]. 高技术通讯, 2006,16(3):241-245.

[24] 秦兵,刘挺,李生. 多文档自动文摘综述 [J]. 中文信息学报, 2005, 19(6): 13-20.

[25] Marcu, D.. The Automatic Construction of Large-Scale Corpora for Summarization Research [A]. In: Proceedings of ACM SIGIR 1999 [C], 137-144, University of California, Berkely, 1999.

[26] Jing H. and K. R. McKeown. The Decomposition of Human-Written Summary Sentences [A]. In: Proceedings of ACM SIGIR 1999 [C]. 129-136, University of California, Beekely, 1999.

[27] Van Rijsbergen, C.J. Information Retrieval, 2nd edition [M]. Dept. of Computer Science, University of Glasgow. 1979.

[28] Jean Carletta. Assessing Agreement on Classification Tasks: The Kappa Statistic [J]. CL, 1996, 22(2): 249-254.

[29] Sidney Siegel and N. John Jr. Castellan. 1988. Non-parametric Statistics for the Behavioral Sciences [M]. McGraw-Hill, Berkeley, CA, 2nd edition.

[30] Dragomir R. Radev and Daniel Tam. Summarization Evaluation using Relative Utility [A]. CIMK2003 [C], 508-511.

[31] Hassel, M. Exploitation of Named Entities in Automatic Text Summarization for Swedish [A]. In: Proceedings of NODALIDA '03 [C]. Iceland. 2003.

[32] Saggion H., Radev D., Teufel S., et al. Meta-evaluation of Summaries in A Cross-lingual Environment Using Content-based Metrics [A]. In: Proceedings of the 19th International Conference on Computational Linguistics [C]. Taipei, 2002: 849-855.

[33] Hans van Halteren, Simone Teufel, Examining The Consensus Between Human Summaries: Initial Experiments with Factoid Analysis [A]. Proceedings of the HLT-NAACL 03 on Text Summarization Workshop [C]. 57-64, May 31, 2003.

[34] Lin, C. Y. (2001). Summary Evaluation Environment (SEE). Available from hayden.isi.edu/SEE/.

[35] Kishore Papineni, Salim Roukos, Todd Ward, and WeiJing Zhu. Bleu: A Method for Automatic Evaluation of Machine Translation [D]. rc22176. Technical report, IBM T.J. Watson Research Center, 2001.

[36] Lin C. Y. ROUGE: A Package for Automatic Evaluation of Summaries [A]. In: Proceedings of the ACL 2004 Workshop on Text Summarization [C]. Spain, 2004.7: 4-8.

[37] Lin C. Y. Looking for a few good metrics: ROUGE and Its Evaluation [A]. Working Notes of NTCIR-4 [C]. 2004, Vol. Supl. 2: 1-8.

[38] Ani Nenkova, Rebecca Passonneau. Evaluating Content Selection in Summarization: The Pyramid Method [A]. HLT-NAACL 2004 [C]. 145-152.

[39] Eduard Hovy, Chin-Yew Lin, Liang Zhou, et al. Automated Summarization Evaluation with Basic Elements [A]. In: Proceedings of the Fifth Conference on Language Resources and Evaluation (LREC 2006) [C]. Genoa, Italy.

[40] Breck Baldwin, Robert Donaway, Eduard Hovy, et al. An evaluation road map for summarization research [D]. TIDES. July 2000.

[41] John M. Conroy, Judith D. Schlesinger, Dianne P. O'Leary. Bridging the ROUGE/Human Evaluation Gap in Multi-Document Summarization [A]. In: Proceedings of DUC2007 [C]. 2007.

[42] Jin Zhang, Hongbo Xu, Xiaolei Wang, et al, ICT CAS at DUC 2007 [A]. In: Proceedings of DUC2007 [C]. 2007.

[43] Hoa Trang Dang. Overview of DUC 2007 [A]. In: Proceedings of DUC2007 [C]. 2007.