中 文 信 息 学 报

第 19 卷 第 6 期 JOURNAL OF CHINESE INFORMATION PROCESSING Vol.19 No.6

[综述]

文章编号:1003 - 0077(2005)06 - 0013 - 08

多文档自动文摘综述

秦 兵,刘 挺,李 生

(哈尔滨工业大学 计算机学院信息检索研究室,黑龙江 哈尔滨 150001)

摘要:多文档文摘是将同一主题下的多个文本描述的主要的信息按压缩比提炼为一个文本的自然语言处理技术。随着互联网上信息的日益丰富,多文档文摘技术成为新的研究热点。本文介绍了多文档文摘的产生和应用背景,阐述了多文档文摘和其他自然语言处理技术的关系,对多文档文摘国内外研究现状进行了分析,在此基础上汇总提出了多文档文摘研究的基本路线及关键技术,并总结了多文档文摘的未来及发展趋势。

关键词:人工智能;自然语言处理;多文档文摘;自然语言处理;文本压缩

中图分类号:TP391 **文献标识码**:A

Survey of Multi-document Summarization

QIN Bing, LIU Ting, LI Sheng

(Information Retrieval Laboratory, School of Computer Science and Technology,
Harbin Institute of Technology, Harbin, Heilongjiang 150001, China)

Abstract: multi-document summarization is a technology of natural languages processing, which extract important information from multiple texts about same topic according to ratio of compression. Multi-document summarization becomes new research spot with increasing of information in internet. In this paper, the background of multi-document summarization is introduced, the relationship with other technologies of natural language processing and the state of arts is analyzed, the key technologies and the methods of research of multi-document summarization are proposed. Finally, the feature of multi-document summarization is forecasted.

Key words: artificial intelligence; natural language processing; multi-document summarization; nature languages processing; compress of texts

1 引言

互联网的普及使人们的生活方式发生了巨大的变化,在网络带给人们大量信息的同时,人们的需求也随着网络信息的急剧增长不断地发生着变化,从而促进了许多新技术诞生和发展。人们面临的较多的问题是,面对成千上万的同一主题网页,它们多数具有相同的信息,而又包含着少量不同的信息,迫切需要一个帮助人们快速浏览信息的工具,该工具不仅提供的是直接的文档,而且是经过加工整理,包含这些文档的重要、全面的信息。这样会大大提高人们获取信息的效率,同时会使越来越多的人利用互联网来获取信息。单文档文摘技术和信息抽取技

收稿日期:2004 - 11 - 22 定稿日期:2005 - 05 - 08

基金项目:国家自然科学基金重点资助项目(60435020)

作者简介:秦兵(1968—),女,副教授,主要研究方向为中文信息处理,信息检索,多文档自动文摘.

术可以帮助人们快速,高效地获取主要信息。这些技术发展至今,已经具备了一定的理论基础,并且发展得比较成熟,得到了广泛的应用。随着互联网的普及,现有的这些方法已经不能满足人们新的需求,在解决目前互联网上多篇同一主题进行汇总和压缩的问题上仍然存在一些方法上的不足。

多文档文摘技术是信息时代发展到一定程度的必然趋势。多文档文摘可以将多篇同一主题的文档进行汇总,提供给人们简洁,全面的信息,将人们从繁琐、冗余的信息中解脱出来,在信息浏览中不仅可以单独作为一个系统应用,而且在其他自然语言处理系统中承担重要角色。例如,在新一代搜索引擎问答系统(Q&A)中,多文档文摘可以作为返回答案后处理模块。现有的搜索引擎只是将一系列与用户需求相关的文档提交给用户,问答系统则是将这些与用户需求相关的文档进行融合,直接提交给用户的是答案。多文档文摘技术也是话题的监测与跟踪技术 TDT(Topic Detection and Tacking)的组成部分,根据用户提供的信息,在互联网的文本流中不断发现与其相关的文本信息,并将新发现的文本与已有的文本进行汇总,生成线索报告提交给相应用户。在国家安全部门的非法信息监测,特殊信息的定制与融合的方面多文档文摘也能发挥重要作用。

多文档文摘的研究为用户提供了方便,提高了用户获取信息的速度和效率,为互联网的应用开辟了新的方向。

2 多文档自动文摘的定义及分类

多文档集合是指同一主题下不同文档的集合,特点是文档之间具有很多的共同信息,各个文档中包含与主题相关的不同的信息的文档集合。多文档文摘是将多文档集合中的多次重复信息以一次出现在文摘中,其他与主题相关的信息根据重要性及压缩比依次抽取的文本集合压缩技术。

多文档文摘的最终目的是为用户服务的,用户的不同需求决定了不同的文摘方法。根据用户的需求可将多文档文摘分为问题相关的多文档文摘和问题无关的多文档文摘。

问题相关的多文档文摘不仅汇总文档集合中的主要信息,去除冗余信息,在选择文摘单元时还需要考虑与问题相关程度。

问题无关的多文档文摘是对具有共同主题的多个文档的汇总,共同主题不是共同的标题,而是指内容中心相似的文档,对于该类文档进行文摘,重点是去除冗余信息,将多文档的内容以简洁、全面的信息呈现给用户。

3 多文档文摘与其他自然语言处理技术的关系

多文档文摘是随着互联网上的信息急剧膨胀而发展起来的新的文本信息处理技术,与其他的自然语言处理技术如信息检索、信息抽取、单文档文摘等有着千丝万缕的联系,同时又有它的独特之处。

信息检索只是找出满足一定检索条件(query)的整篇文档或段落,而人们仍然必须阅读所找到的每一个文档或段落才能获得所需要的信息,多文档自动文摘可以将这些相关的信息按用户需求或文本内容进行汇总生成全面简洁的文本直接提供给用户。

信息抽取具有预定的目标,根据固定的模板在文本中的提取信息。这种模板表示了某一领域中的结构信息,因此信息抽取被局限于特定域,信息抽取的主要任务是对需要的文本信息的进行识别,寻找与模板匹配的信息,不需要对文本进行全面地分析和理解。问题相关的自动

文摘同样含有预定的目标,但其目标是动态的,需从用户提出的问题中获取的,根据用户的需求将相关的答案进行整理汇总,以文摘的形式提交给用户。问题无关的自动文摘不具有预先规定目标,需要对文本的内容进行分析和处理,去除文本中的冗余信息,将其余信息进行有机的融合得到。目前的研究有将多文档文摘和信息抽取相结合的趋势,即从相似的文本中自动获取模板,将模板合并生成多文档文摘。该方法将信息提取与多文档文摘的优点结合起来应用于开放域信息检索,具有很大的应用前景。

单文档自动文摘及多文档自动文摘两者有相似之处,都需要对文本进行分析和理解,但是还有许多不同:多文档存在的冗余信息更多,对相同或相似的句子的关系的处理显得更为重要;在多文档集合中,不同媒体对同一事件的报道有时是不同角度、不同时间的信息,甚至是相反的信息,对于矛盾的句子的处理,将成为一个多文档文摘中特有的问题;单文档文摘生成按照文本单元在原来文档中的顺序就可生成文摘,多文档文摘生成过程中的排序将成为输出时的关键问题。尽管存在上述不同,但是仍然可以利用单文档文摘对多文档的集合进行初步的分类,将相似的文档聚集在一起,进一步生成最终文摘。另一方面,可以针对多文挡集合利用单文档文摘技术来生成多文档文摘,将多文档集当作一个文本,根据位置、词频、标题、段首等信息进行文本单元的抽取。

总之,多文档文摘技术可以看作为信息检索的后处理,单文档文摘技术的扩展,信息抽取技术的应用[1]。

4 国内外研究现状分析

从 20 世纪 50 年代末 Luhn 开创自动文摘领域,自动文摘技术逐渐地发展起来。多文档自动文摘的研究工作最早在 20 世纪 80 年代开始,当时的研究工作还不具有普遍性,主要在科技文章中通过多种关系描述对科技文章的多文档集合描述,科技文章的结构化统一些,比较好刻画,但这种方法是受限域的,不利于推广。网络的普及使跨文本的信息融合正在成为新的研究热点。真正的任意域的多文档文摘的研究是在 1997 年开始的。在国内,对中文的多文档文摘的研究目前还处于起步阶段,日本和中国台湾的学者在这方面做了一些工作[2],国内哈工大、中科院自动化所、中国科大等研究机构也开展了这方面的工作,总的来说,相关的文章发表得不多。

DUC(Document Understanding confereence)[3],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)赞助发起文本理解会议 DUC(Document Understanding Conference),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,每年的任务和评测都是针对单文档文摘和多文档文摘进行评测。随着人们的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。但是由于 DUC 没有针对中文的语料,因此还需要在中文多文档文摘的评价上作一些工作,从而客观的衡量系统生成的文摘的质量。另外,其他一些大型评测会议,例如信息检索最高级别的会议 TREC,以及话题识别与跟踪 TDT 会议也都涉及到多文档文摘的技术。

在多文档文摘近几年的研究中,诞生了许多多文档文摘系统,根据采用的方法的不同,大致可以分成以下几类:

1. 基于单文档文摘技术的方法

许多系统采用单文档文摘技术生成多文档文摘。例如南加州大学的 NeATS 系统[4]，该系统融合了单文档文摘技术，利用词频、句子位置、主题词等特征信息，利用 MMR(Maximal marginal relevance)的简化版本选择和过滤内容。该系统采用的并非是一些新技术，但是将这些技术应用于多文档文摘中并且在大规模公共任务上进行评价却是开创性的。由于是该系统是一个原型系统，特意采用了一些简单的技术：用统计的方法抽取重要的概念；利用它们的位置和主题词(stigma words)过滤句子；用 MMR 降低句子的冗余；根据时间表记按照年代信息进行排序。该系统在 2001 年的 DUC 评测中名列前茅。Newsblaster[5]是哥伦比亚大学在多文档文摘方面的一个比较成功的系统，它是一个新闻跟踪的工具，并可以为每天的主要新闻做出相关的文摘。Newsblaster 系统将新闻的浏览分为两部分：文摘的生成部分采用在 DUC - 2001 中参加评测的文摘生成器，将新闻进行划分归类，然后采用哥伦比亚大学开发 TDT 系统对相关信息进行检测与跟踪。该系统不同于其他的 TDT 系统，它采用确定文档相似度的几种特征的不同权值下合并对文档进行归类。Newsblaster 系统的独特之处在于它是将 DUC 和 TDT 结合起来的新闻浏览系统，目前已经推出了跨语言的新闻浏览系统。但是它仅仅将单文档文摘技术的方法应用于多文档文摘，忽略了多文档集合中文档之间的信息，在文摘质量的提高上必然存在一定的局限性。

2. 基于信息抽取的方法

信息抽取技术作为重要的文摘抽取工具也被应用到多文档自动文摘技术中。1998 年 Radev and McKeown 开始尝试将信息抽取技术应用到多文档自动文摘中来[6]，并成功地开发出一个应用于自然灾害领域文摘的原型系统 SUMMONS。该系统需要人工制定模板，但人工制定模板需要较大的人力，并且不易更新，仅适用于特定领域，不宜推广。SUMMONS 是第一个将自然语言处理技术与信息抽取相结合的多文档文摘系统，在当时的情况下是对一个新领域探索，必然有一些不成熟的地方，例如对于不同形式的数字表示不能很好的识别，并且也没有给出具体的评价。康奈尔大学的 michael white 等人开发的 RIPTIDES 系统也是一个基于信息抽取的系统[1]，和 SUMMONS 相比作了以下改进：抽取的句子使文摘更完整，更力求发现最相关的信息。SUMMONS 回避了数字表达形式问题，RIPTIDES 系统通过制定一些规则解决了这个问题。GISTEXTER 系统[7]也是基于信息抽取的多文档文摘系统，对于给定的领域利用信息抽取系统抽取主要的信息和公共模版，使文摘围绕着主要信息生成。当新的 topic 出现时，通过利用 WordNet 获得主题概念间的统计关系，生成 ad_hoc 模版。该系统融合了信息抽取和单文档文摘技术，并且提出了自动获取模板的方法，可以适用于非特定域的情况。不足之处在于自动获取模板需要较多的语料进行学习，并且该方法获得的文摘的语法不是很好。

3. 基于多文档集合特征的方法

目前多文档文摘的方法主要是将集中在利用多文档集合的信息，将多文档集合作为一个整体进行研究，通过对多文档集合中的句子按照其表达意思的相近程度重新组合聚类，然后从不同的类别中抽取文摘句。该方法可以在理解的角度上作文摘，相比较之下获得较好的文摘。美国密西根大学的 Radev 等人首先提出了质心的概念[8]，文摘的生成应从识别多文档集合的质心开始，在这里质心代表了文档集合的主题。2000 年，他们在此基础上开发了一个多文档自动文摘系统 MEAD。它采用统计的方法找出在多篇文档中出现频率最高的词和短语构成文档束的质心，构成伪句子，然后将文档集合中的其他句子与该伪句子计算相似度，进行排序。另一个具有代表性的研究是哥伦比亚大学的 McKeown 和 Radev 等人开发的基于片断聚类的多

文档文摘系统 MultiGen[9]。该系统利用重复信息(Repeated information)作为文摘内容的主要候选,从识别不同文章的异同点入手,将语义相似度高的段落融合到一起作为文档集的一个主题,并将主题中的短语或词组的交集作为关键词抽取出来并利用语言生成系统 FUF/SURGE 组成句子并生成一篇文档。一些学者提出了子事件概念[10~12],通过聚类方法将多文档集合原来的以文本为单元转化为以逻辑意义为单元的子集合看作子事件,通过对这些子事件抽取,生成文摘的主要内容。这种方法在理论上冗余性会更少、信息的覆盖率会更大,是目前比较流行的一种方法。

在中文处理方面,日本东京大学的 Minghui WANG 和 Hediheko TANAKA 开发了利用参考文献信息的多文档中文自动文摘系统[2]。其原理是通过抽取原文中作者讲述参考文献内容及其和原文异同关系的部分来组成文摘。这一方法应用领域狭窄,实际系统中对关于神经网络学习算法方面的科技论文实现了多文档文摘生成,而且仅是建立在文本浅层语法分析的基础上的,文摘的质量无法保证。中文多文档文摘的研究起步较晚,从技术上看,采用的主要技术手段大致是相同的,但是在这些技术使用过程中,需要利用的一些中文的资源和测试平台还不够成熟,例如,中文多文档文摘缺乏统一的评测,一些中文信息处理技术还不够成熟,在某种程度上制约了中文多文档自动文摘的发展。

评价是自然语言处理系统中一个关键的部分,也是最有争议的一个部分。近年来,由于多文档文摘和多语言文摘研究的重要性日益增长,对评价也提出了新的要求。评价作为自然语言理解技术的一部分,它提供了对结果进行比较和复述的环境,为产生更好的结果提供了竞争的自然环境。总之,建立一个好的评测系统将会对整个文摘质量的提高起促进作用。因此好的评价方法也就成为了一个亟待解决的问题。

传统的自动文摘评价方法主要由人根据以下几个指标:一致性,简洁性,文法合理性,可读性,及内容含量判断文摘的质量,但是人工评价在大规模文本进行评测时,需要消耗大量的人力,实现起来比较困难[24]。近几年,如何进行文摘的自动评价引起了人们极大的重视。Saggion 等[25]提出了通过计算余弦相似度,文本单元的重叠率,以及最长公共子串进行文摘评价方法,但是不足之处是该方法没有给出评价方法和人工评价方法的相关性。后来,由于 BLEU 方法在机器翻译评价中获得了成功,Lin and Hovy[23]提出了与其类似自动文摘系统评价系统 ROUGE (Recall-Oriented Understudy for Gisting Evaluation),该评价系统通过统计 n-gram 的共现对单文档文摘和多文档文摘进行评价。实验结果表明,该方法对单文档文摘的评价结果与人工评价的结果具有很好的相关性,对于多文档文摘的评价结果和人工评价的结果相关性不很理想,这也是多文档文摘评价方法需要进一步研究的目标。

总的来说,文摘的评价方法通常可以分为两类:第一类是内部的评价方法,即通过一系列的参数直接分析文摘质量的好坏。这可以借助于用户对文摘的连贯程度以及包含多少原文章关键信息的判断,也可以比较自动文摘与"标准"文摘的相似程度。第二类是外部的评价方法,即通过分析自动文摘对其它任务的完成质量的影响来判断。例如信息检索、自动问答、阅读理解等任务。

综上所述,多文档自动文摘技术,特别是中文多文档自动文摘技术的研究面临着社会需求特别巨大、学术研究急需大力开展的起步阶段。

5 多文档文摘研究的基本路线

目前的研究大多数都是基于句子抽取的多文档文摘。对于中文自然语言的处理系统,为

了更好的理解文本和句子的内容,常常需要进行分词处理。无论是中文多文档文摘系统还是其他语言多文档文摘系统,在系统的模块划分上是存在共性的,按照研究的技术路线划分为不可缺少的三部分:句子相似度的计算,文摘句的抽取和文摘句的排序,这些也是多文档文摘的关键技术。在此基础上对各个模块的不同方法的实现构成了不同的多文档文摘系统。

句子相似度的计算,目的是判断句子和句子之间的相似程度,确定表达意思的远近,为分析多文档集合奠定基础。

1. 句子相似度的计算

句子相似度计算是多文档文摘最关键也是最基础的一步。通过相似度计算可以判断多文档集合中冗余信息的多少,在句子的抽取时根据句子的相似度抽取冗余性最小的句子组成文摘句集合,可以看到句子相似度的值将在多文档文摘各项技术中发挥作用。句子相似度计算不仅在多文档文摘中充当重要角色,而且在问答系统、机器翻译等其他自然语言其他处理技术中也发挥着重要作用。国内外学者在句子相似度计算方面做了许多工作,总的来看大致归纳为以下几种方法:

基于 tf * idf 的句子相似度计算[17]:依靠句子之间词的匹配程度,确定句子之间的相似度。该方法简单直观,便于实现,但是由于没有考虑句子的深层信息,很难真正理解句子的意义,对于不同的词表达相似意思的句子不能很好的识别出来,应用上有一定的局限性。

基于隐含语义索引(LSI,latent semantic index)的句子相似度计算[18]:康奈尔大学的学者通过这种方法,根据上下文信息来确定词的语义。该方法不借助任何词典确定句子的语义关系,但是上下文的长度决定了计算的准确度,以句子为单位上下文有限,因此计算的准确度有一定的局限;另一方面,进行矩阵运算计算代价较大,对实时性会有一定影响。

基于语义辞典的句子相似度计算[19]:借助语义词典对句子中的词汇进行深层理解,通过词汇的语义相似度汇总后得到句子的相似度。该方法的特点是可以通过词汇的语义信息对句子进行深层理解,但是对于一词多义的现象首先进行词义消岐,保证每个词只有一个确切的语义表达,才能进行语义相似度计算。

基于句法分析的句子相似度计算[20,21]:句子构成不仅包括其中的词汇,而且还包括句子的结构。描述句子的结构通过词语之间的修饰关系表示,计算句子相似度同时从句子结构和词汇本身的信息来考虑,理论上会得到较准确的句子相似度的值。但是由于句法分析技术还不是很成熟,目前该方法只能停留在简单匹配上。

尽管句子相似度的计算在自然语言的许多领域得到应用,但侧重点也不同,在基于实例的机器翻译中更强调语法和语序的一致性,在信息检索领域更侧重于语义的相似。因此在不同的应用领域上句子相似度的计算方法会略有不同。

2. 文摘句的抽取

文摘句的抽取也就是对原始多文档集合主要信息的抽取,以句子为单元的信息抽取由于其含有较少的冗余信息并且具有一定的连贯性成为研究的主流。

在不同的多文档文摘系统中文摘句的抽取有两种方法:一是将文档集合中所有的句子按照某个特征或多个特征的组合统一进行排序,按照顺序进行文摘句抽取。二是将多文档集合划分为若干按意义相似文本单元组成的子集合,在不同的子集合中抽取句子,组成文摘。

在第一类多文档文摘系统中,得到文档集合中句子相似度的值之后,直接在多文档集合中进行句子抽取。例如,比较著名的方法是卡耐基梅隆大学的 Jade Goldstein 等人提出的基于 MMR(Maximal Marginal Relevance)的多文档自动文摘方法[13],通过 MMR 方法做文摘,将与主题

相关、而句子之间不相似的句子保存在文摘中，从而达到去除冗余信息的目的，这种方法主要适用于问题相关的多文档文摘。另一个比较典型的例子是密西根大学 Redev 提出基于质心的多文档自动文摘方法[14]，首先以词为研究单元，以该类中的高频词组成伪句子，以其为质心，按照句子与质心的相似程度，句子的位置，以及句子与首句的相关程度对句子打分，根据分数对句子排序，根据压缩比抽取句子，生成文摘。

在另一类多文档文摘系统中，对相似的句子进行聚类，形成多文档集合中的逻辑主题。然后在各个逻辑主题中抽取句子生成文摘[11]。这种方法生成的文摘可以降低冗余度，提高文摘的覆盖率。该方法由于对文档集合的文本单元的理解，不仅停留于浅层，而是从文档集合的逻辑结构理解文档集合，从而使文摘的质量更高，成为目前多文档文摘研究的主流。但不足是需要事先确定多文档集合逻辑主题类数，多文档集合的逻辑主题数是根据内容的紧凑程度确定，一般是未知的。一般的聚类方法的方法，都有一定的不足，基于划分聚类方法须要事先知道类别数，基于层次的方法可以不需要事先知道类别数，终止条件可以通过阈值判断，但该方法不足是聚类过程不能回溯，一旦确定对象的所属类别就不能更改。根据处理对象的特点选择合适的经过改进的聚类方法，该问题可以在一定程度上得到解决。

在文摘句的抽取上，多文档文摘的句子抽取方法不同于单文档文摘。单文档文摘需要抽取的信息的分布情况是一致的，即在原文中出现的信息的比例和在文摘中出现的比例是一致的。但是在多文档文摘中由于原始文档集合来自于不同的文本，重复信息较多，为了使用户获得全面简洁的信息，需要将多次重复的信息以一次出现在文摘中，并且将在不同文档出现的信息按照重要度和压缩比的要求抽取到文摘中。

3. 文摘句的排序

文摘句的排序也是一个很重要的过程，单文档文摘对文摘句的排序不太敏感，可以将抽取的句子按照原文的顺序输出。但对于多文档文摘，句子来自于不同的文档，句子的排序不仅解决流利度的问题，实际上还可以帮助人们正确理解原文的意思，因此是一个必不可少的过程。

不同的文本单元，生成方法不同，有的研究工作是基于段落单元的，通过找到的主要信息的段落，按与检索的相关程度，或者说按照信息量的多少进行排序输出；对于基于句子单元的文摘，生成复杂一些，要考虑句子的内容和时间信息，哥伦比亚大学 Regina Barzilay[14] 已经在这方面做了一些工作，Zhuli Xie[15] 通过进化算法 GEP(Gene Expression Programming)作为学习机制，通过对人工文摘和原文的对比找到排序的规律，对句子进行排序，不足之处在于目标文摘本身存在主观性，有待于找到更客观的特征刻画排序技术。

6 展望

多文档文摘的研究目前还停留在句子抽取阶段，实际上句子作为文摘的最小单位不是最理想的。这是由于有时在一个句子中还会包含冗余信息，有时一个句子表达的意思还不够完整，需要多个句子才能表达清楚。有的学者[16]提出了对句子进行压缩和融合，就是通过句法分析和统计的方法，对句子进行剪裁，利用文本生成技术重新组合，生成文摘，使文摘更加精炼。但是这种方法实现起来还有很多困难，需要涉及许多自然语言处理技术，例如句法分析技术，文本生成技术等，目前这些技术还不够成熟，因此仅停留于研究阶段，将来是多文档文摘发展的一个趋势。

随着国际间交流的增多，一些信息的描述不再局限于一种语言，在同一主题的多文档集合中会包含多种语言的文本，因此跨语言跨文档的研究必将成为多文档文摘的研究趋势[22]，随

着机器翻译技术的日益成熟,必将有广泛的应用前景。

对于中文多文档自动文摘由于缺乏大规模统一的测试集以及测试平台,从而制约了它的发展。同时其他自然语言资源的不成熟,对多文档文摘的实用化产生了一定的影响。随着更多的中文自然语言的资源库的健全和开放,中文句法分析等自然语言技术的成熟,以及中文多文档文摘统一的评测平台的建立和推广,相信中文多文档文摘的发展将会有一个质的飞跃。

参 考 文 献:

[1] Michael White, Tanya Korelsky, Claire Cardie, Vincent Ng, David Pierce and Kiri Wagstaff. Multidocument Summarization via Information Extraction[A]. In: Proceedings of the First International Conference on Human Language Technology Research[C]. 1998:36 - 44.

[2] Minghui Wang and Hediheko Tanaka. Summarization of Multiple Chinese Technical Articles[A]. In: The First International Conference on Information[C]. Fukuoka, Japan. 2002:16 - 19.

[3] http://www-nlpir.nist.gov/projects/duc/index.html.

[4] Chin-Yew Lin, Eduard Hovy. From Single to Multi-document Summarization: A Prototype System and its Evaluation [A]. In Proceeding of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL - 02) [C], Philadelphia, USA, 2002:25 - 34.

[5] Kathleen R. McKeown, Regina Barzilay, David Kirk Evans, etal. Tracking and summarizing news on a daily basis with columbia's newsblaster[A]. In Proceedings of the Human Language Technology Conference.2002[C].

[6] Dragomir R.Radev, Kathleen R.McKeovwn. Generating Natural Languages Summaries from Multiple On-Line Sources[J]. Computational Linguistics. 1998, 24(3):21 - 29.

[7] Sanda Harabagiu and Steven Maiorano. Multi-Document Summarization with GISTexter[A]. In: Proceedings of the Third LREC Conference 2002 (LREC 2002)[C]. June 2002, Canary Islands, Spain.

[8] R.Radev, Hongyan Jing and Malgorzata Budzikowska. 2000. Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies[A]. In: ANLP/NAACL 2000 Workshop[C]. 2000: 21 - 29.

[9] http://www1.cs.columbia.edu/~regina/demo4/.

[10] Naomi Daniel, Dragomir Radev, and Timothy Allison. Sub-event based multi-document summarization[A]. In: HLT NAACL Workshop on Text Summarization[C]. Edmonton Alberta.Canada.2003:9 - 16.

[11] Endre Boros, Paul B.Kantor, and David J.Neu. A Clustering Based Approach to Creating Multi-Document Summaries[A]. In Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval[C]. New Orleans, LA, 2001.

[12] Pascale Fung ,Grace Ngai Combining Optimal Clustering and Hidden Markov Model for Extractive[A]. In: Proceedings of the ACL 2003 workshop on multilingual summarization and question answering[C].2003:21 - 28.

[13] Carbonell, J. G., and Goldstein, J. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries[A]. In: Proceedings of SIGIR - 98. Info. Proc. and Management[C]. 31(5):675 - 685.

[14] Regina Barzilay, Noemie Elhadad, and Kathleen R. McKeown. Sentence Ordering in Multidocument Summarization [A]. In: Proceedings of the 1st Human Language Technology Conference[C]. San Diego, California, 2001:32 - 38.

[15] Zhuli Xie, Xin Li, Barbara Di Eugenio, Weimin Xiao, Thomas M. Tirpak and Peter C. Nelson Using Gene Expression Programming to Construct Sentence Ranking Functions for Text Summarization[A]. In:20th International Conference on Computational Linguistics[C]. 2004:1381 - 1384.

[16] Lin, C.Y. 2003. Improving summarization performance by sentence compression: A pilot study[A]. In: Proceedings of the Sixth International Workshop on Information Retrieval with Asian Languages[C]. 2003:1 - 9.

(下转第 56 页)

[6] Gale W., K. Church and D. Yarowsky. Using Bilingual Materials to Develop Word Sense Disambiguation Methods [A]. In Proceedings, Fourth International Conference on Theoretical and Methodological Issues in Machine Translation[C]. Montreal, 1992. 101 - 112.

[7] Li, C, and H. Li. Word Translation Disambiguation Using Bilingual Bootstrapping[A]. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL) [C], Philadelphia, 2002. 343 - 351.

[8] Diab, M. and P. Resnik. An Unsupervised Method for Word Sense Tagging using Parallel Corpora[A]. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL) [C], 2002. 255 - 262.

[9] Ng, H.T., Wang, B., Chan, Y.S. Exploiting Parallel Texts for Word Sense Disambiguation: An Empirical Study[A]. In Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL) [C], Sapporo, 2003. 455 - 462.

[10] 刘群,李素建.基于《知网》的词汇语义相似度计算[A].第三届汉语词汇语义学研讨会[C],台北,2002.

[11] 程葳,赵军,徐波,刘非凡.一种面向汉英口语翻译的双语语块处理方法[J].中文信息学报,2003,17(2):21 - 27.

(上接第 20 页)

[17] Yohei Seki: Sentence Extraction by tf/idf and Position Weighting from Newspaper Articles [A]. In: Proceedings of the Third NTCIR Workshop on Research in Information Retrieval[C]. Automatic Text Summarization and Question Answering, Tokyo, 2002: 55 - 59.

[18] Rie Kubota Ando, Branimir K. Boguraev, Roy J. Byrd and Mary S. Neff. Multi-documentSummarization by Visualizing Topical Content[A]. In: ANLP-NAACL 2000[C]. Advanced Summerization Workshop, Seattle, WA, 2000:12 - 19.

[19] 秦兵,刘挺,王洋,等.基于常问问题集的中文问答系统的研究[J].哈尔滨工业大学学报,2003,35(10):1179 - 1182.

[20] 穗志方,俞士汶. 基于骨架依存树的语句相似度计算模型[A]. 中文信息处理国际会议(ICCIP'98)[C]. 1998:23 - 27.

[21] Tsutomu Hirao, Jun Suzuki, Hideki Isozaki and Eisaku Maeda Dependency-based Sentence Alignment for Multiple Document Summarization[A]. In: 20th International Conference on Computational Linguistics[C]. 2004:446 - 452.

[22] David Kirk Evans, Judith L. Klavans and Kathleen R. McKeown. Columbia Newsblaster: Multilingual News Summarization on the Web. Demonstration[A]. In: HLT-NAACL 2004 [C]. 2004:1 - 4.

[23] Lin, C-Y, E. H. Hovy. 2003. Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics[A]. In: Proceeding of 2003 Language Techology Conference(HIT-NAACL 2003)[C], Edmonton, Canada.

[24] Over, P and J. Yen. 2003. An Introduction to DUC 2003 - Intrinstic Evaluation of Generic News Text Summarization Systems. http://www.nlpir.nist.gov/projects/duc/pubs/2003slides/duc2003intro.pdf.

[25] Saggion H., D. Radev, S. Teufel, and W. Lam. 2002. Meta-Evaluation of Summarization in a cross-Lingual Environment Using-Based Metrics. In: Proceedings of COLING - 2002, Taipei.