

自动文摘技术及应用*

金 博^{1,2}, 史彦军², 滕弘飞^{1,2}, 艾景波²

(1. 大连理工大学 机械工程学院; 2. 大连理工大学 计算机科学与工程系, 辽宁 大连 116024)

摘 要: 综述了自动文摘技术的研究进展。面向自动文摘系统用户, 介绍了自动文摘系统及其应用, 分析了机械文摘和理解文摘各自的特点, 展望了自动文摘技术今后的发展方向和趋势。

关键词: 自动文摘; 机械文摘; 理解文摘

中图法分类号: TP391. 1 文献标识码: A 文章编号: 1001-3695(2004)12-0013-03

Automatic Abstracting Technology and Its Application

JIN Bo^{1, 2}, SHI Yan-jun², TENG Hong-fei^{1,2}, AI Jing-bo²

(1. School of Mechanical Engineering; 2. Dept. of Computer Science & Engineering, Dalian University of Technology, Dalian Liaoning 116024, China)

Abstract: The research development of automatic abstracting technology is reviewed. Oriented automatic abstracting system users, this paper introduces the automatic abstracting system and its application, and analyzes the characteristic of statistical abstract and understanding abstract. The developing direction and trend of automatic abstracting technology is prospected as well.

Key words: Automatic Abstracting; Statistical Abstracting; Understanding Abstracting

1 引言

概括介绍一篇文章的内容可以有多种方式, 其中最主要的方法就是作文摘。中华人民共和国国家标准《文摘编写规则》(GB6447-86)对文摘的定义是:“(文摘是)以迅速掌握报道内容概略为目的而编写的文章, 不加主观的解释和评论, 简洁而准确地记述报道的重要内容^[1]。”自动文摘技术的作用是生成给定原文的中心内容, 或把所需要的内容从文章中自动抽取出来, 并用同于或不同于原文的句子表示出来^[2]。其软件系统称之为自动文摘系统。

自动文摘技术的研究开始于 20 世纪 50 年代末, IBM 公司的 Luhn 首次设计了一个自动文摘系统^[3]。进入 90 年代以来, 随着 Internet 的开通, 自动文摘的价值充分显露出来, 引起了世人的极大关注, 越来越多的学者纷纷开始从认知心理学、情报科学、计算语言学等各个方面展开研究, 提出了实现自动文摘的新的思路和方法, 自动文摘的研究进入了前所未有的繁荣期。

迄今为止的自动文摘系统主要经历了以下两个阶段: 基于统计的机械文摘和基于意义的理解文摘^[4]。这两个阶段的特点可以从表 1 的几个方面的比较中看出来。各自都有其优缺点: 机械文摘方法简单, 容易实现, 但结果不尽如人意; 理解文摘是在对全文理解的基础上进行的, 比较接近于人提取摘要的过程, 文摘的结果一般较好, 但难度较大, 不易实现。

表 1 机械文摘与理解文摘的比较

类型	原文内容	原文输入形式	采用的处理方法
机械文摘	相对非受限域	自然的文本形式	模式匹配, 启发函数, 词频统计
理解文摘	受限域	知识表达形式	知识工程方法, 语法语义, 知识库

需要指出的是, 自动文摘的上述两个阶段是彼此重叠的, 即使在目前, 机械文摘仍然是自动文摘技术领域最重要的文摘方法之一。其根本原因是由于目前自动文摘技术还远没有被完美解决, 众多的研究人员还都在摸索。

2 机械文摘

机械文摘是指根据文章的外在特征抽取原文中的部分句子作为摘要。机械文摘的基本原理是: 文章中有一些能够反映文章主题的词, 可以称为有效词 (Significant Word), 有效词集中的句子就是能概括文摘主旨的句子, 如关键句 (Key Sentence), 关键句集构成摘要。严格地说, 根据这种方法提取出来的文本只能称为摘录, 而不是真正意义上的文摘^[5]。

机械文摘的生成通常是通过分析文本的六种形式特征来确定的。这六种特征是: 词频、标题、位置、句法结构、线索词、指示性短语等, 它们从不同的角度揭示文本主题。由此也可以看出, 机械文摘所使用的方法主要是依靠统计 (如统计词频) 和经验 (如分析关键词关键句出现的位置) 获得的。

机械文摘相关的研究开始得比较早, 从 50 年代末 Luhn 所做的最初的研究开始, 机械文摘就一直是自动文摘领域最重要的解决方法之一。其中较为成功的例子有: IBM 公司的文摘自动生成程序 ACSI-matic^[6], 该系统以 Luhn 的研究为基础, 通过计算句子在文献中的权重来进行文献的摘录, 其在权值的计算

收稿日期: 2003-07-14; 修返日期: 2004-03-18
基金项目: 国家自然科学基金资助项目(60073036); 教育部博士点基金资助项目(20020141005)

方面对 Luhn 的研究进行了改进; 美国 GE 研究与开发中心的 Rau 等人^[7]实现了 ANES 系统, 该系统采用相对词频作为词的权值来分析文献, 并生成摘要; 70 年代初, 俄亥俄州立大学的 Rush 教授^[8]开发的 ADAM (Automatic Document Abstrcting Method) 系统, 该系统强调的是排斥句子的标准, 而不是选择句子的标准, 是利用从文献中删除句子的方法进行文摘生成。由于机械文摘的本身性质所决定, 其质量一般不高, 所以为了保证文摘效果, 许多自动文摘系统都综合利用了分析文本的多种特性。如新加坡南洋大学研制的图书馆新闻删节系统^[9], 提供了题名法、位置法、关键词法和指示性短语法四种自动摘录方法供用户选择。另外, 1997 年, 日本的 Nomoto 等人^[10]提出的一种基于语料库的自动摘录方法。它是让计算机自动地从训练集中提炼出各个特征的结合函数, 为多种形式特征的综合利用开辟了一条新的道路。

通过传统的机械文摘技术给出的文摘虽然通常能够抓住文献的关键所在, 并用作者的原句加以概括, 而且获得了一定的实际应用, 但它的缺点也是明显的, 其中最为突出的有以下几个方面: 摘要的质量不稳定, 缺乏句间的连贯性, 有时摘要内容冗余等。

3 理解文摘

基于上述机械文摘的缺陷, 人们探索了利用自然语言理解技术进行自动文摘的方法。由于受到知识不足的限制, 基于理解的文摘技术只能适用于某个狭窄的领域, 如用于处理有军事情况的新闻等, 但摘要的质量明显优于传统文摘^[11]。

基于理解的文摘方法是以人工智能, 特别是自然语言理解技术为基础而发展起来的文摘方法。该方法与机械文摘的明显区别在于对知识的利用, 它不仅利用语言学知识获取语言结构, 更重要的是利用领域知识进行判断、推理, 得到文摘的意义表示, 最后从意义表示中生成摘要。基本原理是: 在某一特定领域的文章中, 必然存在着特定的信息焦点, 即读者感兴趣的内容, 如军事情况报道必然包含有关的地点、人数、伤亡情况等内容。利用语言学手段将文章中代表这些信息焦点的文字识别出来, 用话语加以组织即可形成一篇连贯的高质量的文摘。

基于理解的文摘方法实现时主要分以下几个步骤: 语法分析、语义分析、语用分析、信息提取和文本生成等。语法分析和语义分析统称文本分析过程, 其目的是要寻找最能代表原文内容的成分; 语用就是语言的实际应用, 主要是进行交际对话, 就是用语言进行信息交流和交换, 语用分析是指分析语用的特点, 即静态变动态、共性变个性、多义变单义、意义与语境相关等; 信息提取即转换过程, 即通过概括等方法压缩文本; 最后一步重组原文内容, 生成文摘。

目前, 理解文摘主要的方法有脚本、概念从属结构、框架、一阶谓词、关联网络、修辞结构以及语用功能等。大多是从文章结构出发, 有局限性地理解文章的内容和结构。

理解文摘系统的相关研究的主要成果有: 70 年代末 80 年代初, 美国耶鲁大学的 Schank^[12]在脚本的基础上研制的 SAM (Script Allicer Mechanism) 系统, 该系统应用脚本分析简单的文献, 并在此基础上总结出摘要。美国耶鲁大学的 DeJong^[13]

于 1979 年研制的著名的 FRUMP (Fast Reading Understanding and Memory Program) 系统, 该系统用于快速阅览英文新闻资料, 是理解文摘系统的样板, FRUMP 由预言器和验证器组成, 预言器利用预先设置好的梗概剧本预测文献中可能出现的事件, 验证器则去证实这些被预测的事件, 并给出实际信息。FRUMP 系统创造了理解文摘的典范, 但由于内部存储的剧本限制, 如果文章中没有该系统所期望的内容则会被误导, 从而出现歧义。美国的 Tait^[14]对 FRUMP 系统进行了改进, 称为 Scrable 系统, 它要求输入的资料在处理前先转换成 CD (Conceptual Dependency) 结构, 在此基础上分析和确定已预测的信息与未预测的信息之间的关系, 并将这两部分信息合理地组织成一篇完整连贯的文摘。意大利 Udine 大学的 Fum 等人^[15]在 80 年代初研制了 SUSY (S Ummarizing System) 缩写系统, 该系统以一阶谓词逻辑为基础, 取得了较好的效果, 体现出了逻辑方法的潜力。德国康斯坦茨大学的 Kuhlen 等人^[16]研制了 TOPIC 系统, 该系统与框架作为知识表示的基础, 通过全文的语法语义分析生成不同长度的摘要, 其处理对象主要是针对微处理器领域的科技文献。80 年代末, 美国 GE 研究与开发中心的 Rau 等人研制了 SCISOR (System for Conceptual Information Summarization, Organization and Retrieval) 概念信息缩写、组织和检索系统^[17,18]。该系统采用关键词过滤和模式匹配等方法对待处理的文献进行分析, 然后采用自底向上(完全的句法分析) 的分析器识别句子的结构, 最后运用自顶向下(部分的句法分析) 的分析器提取句子结构中的内容, 是典型的理解文摘, 处理对象是关于“公司合并”的新闻报道。

目前的理解文摘同样有其不足, 主要在于领域严格受限。造成领域受限的原因有: (1) 面向大规模真实语料的语法语义分析技术尚未完全成熟, 因此如果想获得高质量的语言分析结果, 就必须将待处理的语料限制在某个范围之内; (2) 理解文摘方法的基础是框架等知识表示, 框架需要根据领域知识预先拟定, 因此如果想把适用于某个领域的理解文摘系统推广到另一领域, 则需重新拟定框架, 这种填充和组织领域知识的沉重负担使理解文摘难以移植。

4 中文自动文摘技术的研究

我国大约从 1985 年开始介绍国外自动文摘方面的研究情况, 从 80 年代末开始研究自动文摘实验系统, 至今也有 10 余年的历史了。但目前的技术水平尚不成熟, 问题主要是在中文本身的语言特点以及自然语言理解方面的困难。

从识别角度来说, 汉语和西文的句子主要区别在于汉语词之间没有空格, 而真正负载信息的是词而不是字, 因而中文自动文摘就存在分词的问题。同时, 汉语的词汇极为丰富, 同一个概念可以用很多不同的词汇表达, 这给词频统计带来了很大的困难。同时, 汉语词汇与西文相比, 在句子中词形没有变化。对中文文章的理解则更是短期无法完成的问题。目前中文自动文摘的研究主要集中在以下几个方面, 即分词、消歧义、词频统计以及理解等。

上海交通大学王永成教授从 80 年代末就开始研究自动摘录技术, 1997 年研制了 OA 中文文献自动摘要系统^[19]。该系

统集成了位置法、指示短语法、关键词法和标题法等多种方法, 是一个较为实用的系统。目前, 依托于上海交通大学的上海交大纳讯高新技术应用有限公司的中英文自动摘要系统就是在此基础上完成的^[20]。

80 年代末, 东北大学姚天顺教授和香港城市理工大学联合开展了“中文全文自动摘要系统”的研究^[21], 该系统采用脚本知识表示, 通过与用户交互获取文摘。

1990 年前后, 中科院软件所的李小滨、徐越, 在北京大学马希文教授的指导下, 对英文自动文摘进行了研究, 并研制了一套实验系统——EAAS (English Automatic Abstract System)^[22]。该系统是一个标准的理解文摘系统, 它局限于“就业机会介绍”这样一个领域。系统首先通过与用户交互获得信息焦点集, 然后对文章进行语法语义分析, 接着按照信息焦点集从框架中搜索推理出有关信息, 最后生成具有一定逻辑性的文摘。

哈尔滨工业大学王开铸教授于 1992 年开始研制基于自然语言理解的文摘系统, 目前已经有了基于词频统计、关键词提取的 HIT863 系列自动文摘系统^[23]和基于理解文摘的实验系统 MATAS^[24]。其中 HIT863 系列机械文摘系统适用于任意领域、任意题材、任意长度的文章, 通过词频等方法实现摘要的自动生成; 理解文摘实验系统 MATAS 用中文意义表示法分析处理输入的文章, 再进行信息压缩, 从而生成摘要。

北京邮电大学信息工程系钟义信教授等人^[25]采用基于多 Agent 技术的文摘方法, 类似于 Paice 的选择与生成文摘法, 目前主要针对计算机病毒方面及新闻报道方面的相关文章, 开发出了 Glance 自动文摘系统及 News 自动文摘系统等^[26]。山西大学郭炳炎教授等人^[27]也在开展自动文摘的研究, 他们采用了基于统计的方法分析文本结构。复旦大学吴立德教授^[28]研制的自动文摘系统分析了篇章段落之间的联系, 建立了语义网, 具有一定的篇章理解能力, 能给出任意长度的摘要。

据悉, 目前一些国际性软件公司也在开发中文自动摘要系统。比如 IBM 中国研究中心、微软亚洲研究院、摩托罗拉中国研究中心等都在研制中文自动文摘系统的产品。

5 结束语

自动文摘技术在几十年中得到不断地发展, 从单纯依靠字面分析的机械文摘过渡到了依靠理解来进行自动文摘生成的理解文摘, 或者是两者相互融合。

目前自动文摘技术应用已经不仅仅限于自动文摘系统软件, 在信息检索、信息管理等领域都得到了应用。比如北京网际创华软件技术有限公司的第六感信息管理系统, 北京通惠利华教育发展中心的教学资源库系统, 北大青鸟网络信息查询系统等, 都是基于自动文摘技术的应用实例。可以预见, 在自然语言处理高速发展的今天, 自动文摘技术的应用范围将更加广阔。

参考文献:

[1] 桑良至. 文献学概论[M]. 北京: 中国书籍出版社, 1993.

[2] 杨建林. 自动文摘的逻辑解释[J]. 情报理论与实践, 2002, 25 (2): 112-115.

[3] Luhn H P. A Statistical Approach to Mechanized Encoding and Searching of Literary Information[J]. IBM Journal, 1957, 309-317.

[4] 吴岩, 刘挺, 王开铸. 中文自动文摘原理与方法探索[J]. 中文信息学报, 1998, 12(2): 8-16.

[5] 刘挺, 等. 机械文摘的方法与实例[J]. 电脑学习, 1996, (4): 1-4.

[6] Luhn H P. The Automatic Creation of Literature Abstracts[J]. IBM Journal of Research and Development, 1958, 2(2): 159-165.

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

[8] Mathis B A, Rush J E. Abstracting[M]. New York: Encyclopedia of Computer and Technology, Marcel Dekker Inc, 1975. 102-142.

[9] Hui S C, Goh A. Incorporating Abstract Generation into an Online Retrieval Interface for a Library Newspaper Cutting System[J]. ASLIB Proceedings, 1996, 48(12): 259-265.

[10] omoto T, Matsumoto Y. Data Reliability and Its Effects on Automatic Abstracting[C]. Proceedings of the 5th Workshop on Very Large Corpora, 1997. 113-126.

[11] 李蕾, 等. 面向特定领域的理解型中文自动文摘系统[J]. 计算机研究与发展, 2000, 37(4): 493-498.

[12] Schank R, Abelson R. Scripts, Plans, Goals and Understanding [M]. Hillsdale, NJ: Erlbaum, 1977.

[13] DeJong G. An Overview of the FRUMP System[M]. London: Lawrence Erlbaum, 1982.

[14] Tait J I. Automatic Summarizing of English text[R]. Technical Report 47, University of Cambridge Computer Laboratory, 1983.

[15] Fum D, Guida G, Tasso C. Forward and Backward Reasoning in Automatic Abstracting[C]. COLING82, 1982. 83-88.

[16] uhlen R. Some Similarities and Differences Between Intellectual and Machine text Understanding for the Purpose of Abstracting[D]. Representation and Exchange of Knowledge as a Basis of Information Processes, 1984. 87-109.

[17] au L F, Jacobs P S, Zemik U. Information Extracting and Text Summarization Using Linguistic Knowledge Acquisition[J]. Information Processing & Management, 1989, 25(4): 419-428.

[18] Jacobs P S, Rau L F. Extracting Information from Online News[J]. Communication of the ACM, 1990, 33(11): 88-97.

[19] 王永成, 许慧敏. OA 中文文献自动摘要系统[J]. 情报学报, 1997, 16(2): 128-132.

[20] Wang Yongcheng, et al. The Development of Automatic Abstracting on Chinese Documents[C]. NLPRS '97 Proc. of the Natural Language Processing Pacific Rim Symposium, 1997. 641-644.

[21] 姚天顺. 自然语言理解——一种让机器懂得人类语言的研究 (第二版)[M]. 北京: 清华大学出版社, 2002.

[22] 李小滨, 徐越. 自动文摘系统 EAAS[J]. 软件学报, 1991, (4): 12-18.

[23] 吴岩, 等. HIT-97I 型英文自动文摘系统[J]. 情报学报, 1998, 17 (5): 358-365.

[24] 王开铸, 等. 基于理解的自动文摘系统设计[J]. 电脑学习, 1996 (2): 4-7.

[25] 胡舜耕, 等. 基于多 Agent 技术的自动文摘系统的研究和设计 [J]. 电子学报, 2001, 29(2): 247-250.

[26] 郭燕慧, 等. 自动文摘综述[J]. 情报学报, 2002, 21(5): 583-591.

[27] 薛翠芳, 郭炳炎. 汉语文本结构的自动分析[J]. 情报学报, 2000, 19(4): 319-326.

[28] Wu Lide, Wei Xionguan, et al. Fudan Abstract System of Chinese Text[J]. Communication of COLIPS, 1996, 6(1): 35-39.

作者简介:

金博(1978-), 男, 博士生, 研究方向为计算智能、自动控制、机器写作; 史彦军(1973-), 男, 博士生, 研究方向为人工智能、协同设计、机器写作; 滕弘飞(1936-), 男, 教授, 博士生导师, 研究方向为人工智能、CAE、布局、人机协同; 艾景波(1979-), 男, 硕士生, 研究方向为人工智能、机器写作。