自然语言处理 文摘生成 计算机

计算机科学 1999Vol. 26№. 10

关于自然语言处理中的文摘生成及其相关技术*)

Text Summarization and Its Related Technologies in NLP

孙睿葵 钟义信
(北京邮电大学信息工程系 北京 100876)

Abstract As Internet is more and more popular, information resources are exploding quickly. People have to find all kinds of tools to search their interests in the network. So some technologies in natural language processing such as Text Summarization, Discourse Analysis and Information Extraction are attracting much attention from people. In this article, the above three technologies are introduced briefly and some associated problems in China are discussed.

Keywords Text summarization, Automatic abstracting, Discourse analysis, Information extraction, Natural language processing

一、引言

自然语言处理(简称 NLP)伴随着计算机的诞生而诞生。在自然语言处理中,开始人们关注的焦点是机器翻译(简称 MT),早在 1958 年已极为盛行。美国投入大量资金和人力组建各种研究小组进行机器翻译的研究。与此同时,自动文摘(Automatic Abstracting,即文摘生成,Text Summarization,简称 TS)也悄悄地诞生了,人们希望计算机能自动地对电子文本进行处理,生成文摘。这样可以节省大量人力进行手工摘要,而且节约时间,同时避免了人为因素而影响文摘的质量。

对文本信息的研究自然地形成了会话分析(Discourse Analysis,DA)和信息抽取(Information Extraction,IE)这两个分支。DA 也有人把它翻译成篇章分析,无论怎样翻译,它的基本内容主要分为两部分:一是篇章结构分析,二是对话分析。这两部分具有一些共同的特点,如句子的衔接(cohesion)与连贯(coherence)、照应(anaphora)等分析都是它们所共有的研究内容,这些研究对 TS 有着重要意义。早期 TS 的研究多是基于统计方法或信息抽取法,试图找出文章中的关键字、词、句,从而生成摘要。其中对关键信息的抽取方法逐渐发展演变至今形成一个独立分支:IE。目前 IE 技术主要应用于对大量同一类文献的信息抽取上。IE 从大量相关自然语言文本中提取出用户感兴趣的信息形成一个数据库供用户查看,为用户作出决策提供支持。TS,DA 与 IE 在 NLP 研究中所处地位如图 1 框架所示。

计算机的发展日新月异。特别是全球互连网的实现,骤然间地球一下子变小了,人们的联系变得触手可及,大量的信息从四面八方被放入网上以便让人们共享。面对这些铺天盖地的信息,人们迫切需要高效、准确的信息处理工具来快速定位到自己的兴趣点,因而各种信息处理的研究成为热点,最直接的就是网上的各种搜索引擎(SE)。

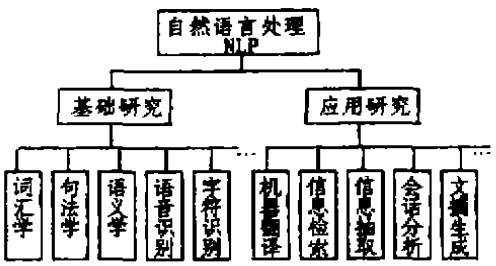

图 1 自然语言处理的基本构成

然而,目前的各种 SE 虽然部分能满足人们的需要,但还有许多问题。以人们使用最多的 YAHOO 为例,如果想了解有关 TS 的信息,当输入"Automatic Abstracting"或"Text Summarization"之后,它利用所谓智能搜索功能搜索到几百或上千个网页,这里面可能只有几个是你所需要的。但你必须花费许多时间去

*)本文工作得到了国家 863 计划的资助(编号:863-317-9601-06-03)。孙睿葵 博士生,研究方向:文摘生成、自然语言处理;钟义信 教授、博士生导师,主要从事人工智能、信息论、智能通信等研究。

查看才能确定，因为从目前的搜索结果来看，我们得到的东西多是文献的题目与地址，而仅从题目本身有时我们不能断定此篇文献是我们想要的。如果我们花费了时间和金钱查找并下载这篇文章后，却发现不是有用的，结果是多么令人沮丧。

现在让我们设想一下，如果有一种工具，它能对每篇电子文本进行自动生成文摘，精确把握住文献讨论的要点，并同时把摘要与某个搜索引擎有效结合起来，搜索的结果不仅有题目，而且还有摘要提示，那么我们就可以精确定位到我们想要的东西上。这是TS技术的应用之一。

目前TS系统面向应用的不多，基本上处于实验室阶段，与TS技术不同，IE技术已有较成功的实际应用系统，如用于人寿保险业MITA系统[1]；关于恐怖分子活动的信息搜集系统；关于芯片制造商的信息搜集系统以及关于企业发展动态的信息搜集系统等[2]，都具有商业价值。

从上面可以看出，TS，DA与IE在信息处理日益重要的今天都具有实际意义，随着它们的成熟与发展，必将大大促进NLP的研究，而且在电子信息逐渐成为人类生活中重要的获得信息来源的时代里，这些技术的应用前景极为广阔。

二、文摘生成(TS)技术的发展概况

前文提到TS开始于五十年代末。58年4月Luhn H. P. 在IBM研究杂志上发表了一篇题为"文章摘要的自动建立"[3]。文中首次介绍了用统计加权和部分语法信息的方法在计算机中生成文摘，当时先要把文章在纸带机上打出，成为计算机可阅读方式，通过纸带输入计算机成为电子文本，进行处理，最后生成摘要。他的实验系统是在IBM704计算机上实现的，它把纽约时报一篇关于化学技术用于精神疾病治疗的新方法一文输入计算机，对每个词出现的次数进行频度统计，并根据它们出现在文章中的不同位置进行加权计算，将权值高于某一阈值的词作为关键字，提取相应的句子加以组合形成摘要。实验结果表明效果不错，是一种可行的方法。同年10月Baxendale P. B. 也用类似的方法在IBM650机上实现了一个实验系统，在此系统中他更注重选取那些可能含有主题(Topic)的关键词，以此选取文摘句[4]。这两个实验系统拉开了TS研究的序幕。

然而，随后的发展并未像人们想象的那么顺利，人们在研究中遇到许多难题。Edmundson H. P. 在64年ACM通信杂志上总结了当时存在的问题[5]，首先是自动文摘的概念问题，然后是输入问题。当时电子文本的输入要靠纸带方式，很不方便。另外计算机的运算，输出以及文摘系统的评估标准都存在问题。因而致使当时美国政府对TS的研究热情不如对MT(机器翻译)的高。美国关于MT的研究小组有十几个共100多人，而TS的小组只有两三个总共12人；各大机构投资在MT上有几百万美元，而用于TS的只有几十万美元。因此，限于当时的条件和人们的认识，TS的研究发展从那时起至八十年代末都较为缓慢。这一点从发表的相关论文数量上也可以看出。在笔者查阅的180篇文献中，50至60年代的有8篇，占4%；70年代的有10篇，占5%；80年代的有26篇，占14%；90年代的有138篇，占77%。由此可见，TS的研究虽呈增长势头，可只有到了九十年代才得到飞速增长，这与互连网的普及和人们对信息的需求及认识密切相关。

七十年代前的TS研究，其技术方法主要为关键字法及统计方法。另外引入了会话分析的结构信息，特别关注一些线索词、大小标题、位置信息，以此来抓住文摘要点，提高文摘质量。

七十年代至八十年代除了美国，其他各国亦逐渐开始从事TS的研究，主要有英国、德国、加拿大、荷兰、日本、韩国及中国。这期间人们注重将会话分析技术引入TS中，此外还探讨了其他方法如用知识获取、心理学、神经网络等方法进行文摘的自动生成。这期间有代表性的系统主要有：耶鲁大学DeJong G. F. 等人建立的FRUMP系统[6]；剑桥大学Tait J. I. 建立的Scrable系统[7]；得克萨斯大学Alterman R. 建立的NEXUS系统[8]；日本Tamura N. 建立的棒球比赛新闻报道的文摘系统[9]等。

进入九十年代，TS的研究骤然升温，研究方法亦百花齐放，涉及的交叉领域也多起来。出现了一些面向应用的文摘系统及工具。Microsoft的Word Auto Summarization，新墨西哥大学的HyperGen，英国Telecom的ProSum，Xerox公司的LinguisticX，Intell。X公司的Summary Express以及Tetranet软件公司的Extractor等[10]。

有代表性的实验系统主要有：美国GE研究开发中心的SCISOR系统[11]；韩国Ulsan大学的ROSE系统[12]；德国Constance等大学的TOPIC系统[13]；加拿大Ottawa大学的TANKA系统[14]；这些系统大多为基于知识理解式的文摘系统，有的采用的技术不同，有的着眼点不同，如韩国ROSE的系统为面向读者型，根据用户要求生成文摘。另外，从认知学角度出发，德国Endres-Niggemeyer B. 等人开发了SimSum文摘系统[15]；Colubia大学的Radev D. R. 和McKeown K. R. 开发的SUMMONS系统[16]是从多个在线资源中提取相关报道，进行比较，指出这些报道的一致性、矛盾点等特征并生成文摘。

目前TS的研究项目主要有美国DARPA资助的TIPSTER PHASE 3，主要从事文摘系统的研究与评估；哥伦比亚大学的在线文档相关摘要的生成项目；德国Hannover大学的SumSum仿真摘要项目；英国Sheffield大学的摘要项目以及加拿大渥太华大学的文摘生成与分类项目[10]。

在我国从事自动文摘系统研究的主要有哈尔滨工业大学[17]，上海交通大学[18]及北京邮电大学[19]、东北大学[20]、山西大学[21]等。

三、文摘生成(TS)技术的研究内容

TS的研究对象是自然语言书写的文本、并以电子形式存放。它将文本中所表达的主要内容概括起来，形成简短的摘要。因而，TS的研究涉及NLP中的自然语言理解(简称NLU)和自然语言生成(简称NLG)两大方面。此外还涉及知识获取(KA)，机器学习(ML)及统计分析等研究内容。

从采用的方法上TS研究可分为两大类：基于统计的方法和基于知识(或称理解)的方法。基于统计的方法主要是机械式地计算各个词在文中出现的次数，结合在文中的位置赋予一定的权值，通过某种算法确定出关键字，进而找出相应的句子，最后进行润色形成文摘；基于知识的方法主要是利用含一定语法、语义的词典及相关句法规则等信息来抽取关键句子，形成文摘。

从应用的领域来看，TS分为面向受限领域和非受限领域两种。面向受限领域的文摘系统多为基于知识型，具有较强的理解能力，生成的文摘可读性较好，质量较高，但实现难度很大；面向非受限领域的文摘系统多采用统计方法，生成的文摘稍差一些，但应用范围不受限制；也有的系统采用统计与理解相结合的方法期望借助两者的长处，把系统做得更好。

TS的研究方法具体讲与如下一些分支关系密切[22]：

1. **篇章技术(Discourse)**

· 关于照应问题的算法研究，如用词语标注法代替句法分析器解决代词指代问题；

· 描述对象的终止与否的判定，如如何确定连接描述词与其前述词的关系；

· 篇章的分段，如利用词的内在联系来把带有注释的文本按子标题分成多段；

· 基于会话分析的摘要生成，如把连贯和衔接用于文摘生成的生成；

· 形式模型研究，如利用术语知识的表达和推理建立一种基于分类的模型从而生成文摘；

· 主题的确定，如根据文体及篇章结构的不同从语料中学习句子出现的位置规律进而确定题目。

2. **信息抽取(Information Extraction)**

· 基于统计的文摘生成，如利用对所有句中词加权建立一个矩阵方法从而生成摘要；

· 段落抽取，如利用一个基础语料及共同出现的高频率的词在各个段落中共现的相似度从而确定出最重要的段落形成摘要；

· 句子抽取，如利用各种加权特征给句子打分，选出最有代表意义的句子作为文摘句；

· 基于模板填充的信息抽取，如利用背景知识建立一时事新闻模式对多篇报道进行比较分析形成摘要；

· 文本分类，如通过语料学习建立一组相关语义词典为以后判别所读的文本类别提供依据。

3. **信息检索(Information Retrieval)**

· 信息检索，如利用标题法、位置法、词项共现信息对相关语句加权计算选出文摘句；

· 文本索引，如用于生成文摘的非受限领域的索引词选取。

4. **自然语言处理中的机器学习** 如为解决代词的歧义问题应用一种基于案例学习的方法进行特征的自动搜集。

5. **其他技术**

· 用于快速文本浏览的超文摘生成提取的图形用户接口技术；

· 基于心理学方法的摘要生成技术；

· 从多媒体资源中生成最新摘要技术；

· 利用数据库信息生成摘要技术；

· 摘要系统的评估和可移植技术等等。

由上述可以看出，TS与周边分支联系紧密，时有交叉。关于它的研究内容目前还没有一个统一的较科学的分支体系，我们只能就其侧重点不同罗列出来。

TS的研究内容根据Inderjeet Mani等人的观点(见AAAI98 Spring Symposium on Intelligent Text Summarization)主要有如下侧重点：

篇章模型(D=discourse model)；评估(E=evaluation)；人类摘要器模型(H=models of human abstractors)；信息抽取(I=information extraction)；基于知识(K=knowledge-based)；机器学习(L=machine learning)；多文档(M=multi-document)；叙述体摘要(N=narrative summarization)；概述(O=overiew)；弱知识(P=knowledge-poor)；统计(S=statistical)；句法分析(Y=syntactic analysis)。

此外一些TS文献也注重如下方面：

联结主义(C=connectionist)；所用输入格式(F=input formatting used)；文本类型(G=text genres)；生

成(Ge＝generation)；超文本建立器(H＝hypertext-builder)；强调(Hi＝highlighting)；多语言(Mu＝multi-lingual)；基于查询(Q＝query-based)；文本检索(R＝text retrieval)；语义表示(Se＝semantic representation)；形象化(V＝visualization)。

四、文摘生成研究在中国

在引言中曾提到TS研究在中国开始时间不长，基本处于起步阶段。由于汉语不同于印欧语系，其词与词中间没有空格，词的各种形态变化少、语序也极为灵活，因而对汉语的分析带来许多特有的问题，汉语理解的基础研究已经有很长一段时间，也取得了许多成果，但仍有许多难题等待攻克，如分词中的歧义问题、词义的解释问题、句法的鲁棒分析等等对TS的研究都有很大的影响。

有一些人避开汉语理解问题采用机械式方法进行文摘生成，通过统计词频及位置加权等方法确定文章的重点，从而提取文摘句，如哈工大的HIT-8631型系统[17]及北邮的GLANCE系统[22]。这种机械性文摘使用的有效信息提取手段有限，因而在可信度、准确度、自然度及逻辑一致性等方面质量要差一些，但它具有应用领域不受限、速度快、生成文摘长度可随意调节等优点。

另一些人意识到机械式文摘研究没有利用文本所提供的各种词、句的语法语义信息，造成文摘质量的不理想，因此投入面向理解的文摘生成研究中。他们从不同的角度分析篇章，力求生成质量较好的文摘。如哈工大注重篇章结构分析与部分词的语义关系，并结合加权计算法开发了基于意义的理解文摘系统MATAS[17]；北邮从语言心理行为角度出发，遵循人们表述观点的逻辑思维顺序结合语法语义分析开发了面向外事新闻报道的NEWS系统[24]。这些系统从理解的角度做了很有意义的尝试，虽然生成的文摘还存在这样那样的问题(如理解程度、范围有限，造成文摘的准确度，自然度等还不尽人意)，但随着研究的深入开展，相信高质量的文摘系统一定是基于信息理解的。

中文文摘系统生成与汉语语言理解是密切相关的。目前对汉语的分词、词性标注、词义排歧、语料库建立及句法分析等方面基础研究正在蓬勃开展。但也应看到我国与外国同类研究相比，还存在差距。首先是电子词典的缺乏，目前在国内很难找到一部带有基本词义或词性标注的汉语电子词典为大家所共享。各个研究小组各自为政，开发自己必需的信息词典，造成人力、物力资源的浪费；其次是大规模带标注的语料的匮乏，使得人们在进行研究的同时，不得不自己去花费许多时间进行手工标注，建立训练样本和测试样本。这些基础性建设对中文信息处理的各种研究是必需的。另一方面，我们在词汇、句式的自学习方面研究的不够。如何自动建立一部信息词典或自动归纳出短语、句式规则，哪怕是面向某种应用的研究课题(如文摘生成)的自学习研究，都将极大地促进中文信息处理的研究。

总之，TS研究不仅具有学术研究价值，而且是面向实际应用的研究课题，在信息爆炸、生活工作节奏日益加快的今天有着迫切的市场需求。

参考文献

1 Glasgow B. MITA-An Information-extraction Approach to the Analysis of Free-form Text in Life Insurance Applications. AI Magazine, 1998(Spring):59～71

2 Information Extraction. NLP Lab Of University Of Massachusetts. Available at: http://www.nlp.cs.umass.edu/～nlpgroup/nlpie.html

3 Luhn H P. The Automatic Creation of Literature Abstracts. IBM Journal of Research and Development, 1958, 2(2):159～165

4 Baxendale P B. Man-made Index for Technical Literature-an Experiment. IBM Journal of Research and Development, 1958, 2(4):354～361

5 Edmundson H P. Problems in Automatic Extracting. Communications of the ACM, 1964(7)

6 Dejong G F. Prediction And Substantiation: A New Approach To Natural Language Processing. Cognitive Science, 1979(3):251～273

7 Tait J I. Generating Summaries Using A Script Based Language Analyzer. In: Steels L, Campbell J A. eds. Progress In Artificial Intelligence. Chichester, Ellis Herwood, 1985

8 Alterman R. A Dictionary Based On Concept Coherence. Artificial Intelligence, 1985, 25:153～186

9 Tamura N. Formalization And Implementation Of Summary Generation. Journal Of The Japanese Society For Artificial Intelligence, 1989, 4(2):196～206

10 Summarizaion Resources. Available at: http://www.cs.columbia.edu/～hjing/summarization.html

11 Jacobs P S, Rau L F. SCISOR: Extracting Information From On-Line News. Communications of the ACM, 1990, 33(11):88～97

12 Bae J J. Another Investigation of Automatic Text Summarization: A Reader-Oriented Approach. Intelligent Information Systems. In: Proc. of the 1994 Second Australian and New Zealand Conf. 1994. 472～476

(下转第11页)

减少数据量来进行的。减少数据量的方法有两种，第一种方法是消除竖直视差，只考虑水平视差，这样一幅全息图就成为一系列竖直排列的条纹，其中每一条纹叫做一个全息直线。在竖直方向上的分辨率不用很高，2lp/mm 就可以达到要求，而在水平方向上的分辨率要足够高以能反映微小的变化，一般来说需要 1000lp/mm。这种方法可以使复杂度降低 1000 倍，问题是竖直方向没有立体感。第二种方法是降低观察角度，减小观察区域。从公式(1)中可以看到，在入射角 θ_2 一定的情况下，衍射角 θ_1 越大，则方程左边的值越大，而入射光波长 λ 是一定的，则 f 的值就会越大，要想不丢失全息信息，采样间距也要相应地减小。如果降低 θ_2，采样间距也会增加，使得总的数据量减少。如果衍射角接近 90 度，假设入射角是 0 度，那么根据公式(1)，最高采样频率应该是 $2/\lambda$。但是如果让衍射角只有 30 度，那么最高采样频率是 $1/\lambda$。而让衍射角是 3 度，那么采样频率只有 $0.1/\lambda$，比衍射角是 30 度的情况数据量降低了 10 倍。这样处理使全息视角变小。

与全息图的高分辨率相关的另一个问题是空间光调制器(SLM)。在上面的讨论中，假设入射角是 0 度，衍射角是 30 度，要求采样频率为 $1/\lambda$。采用 He-Ne 激光，最小的采样间距是 0.6 微米，要求 SLM 的分辨率应该小于或者等于 0.6 微米，这是很困难的。现在最常用的 SLM 是 LCD，其最小的元素是 50 微米，远远大于 0.6 微米。其它的 SLM 也存在分辨率低这个问题。SLM 分辨率低的直接后果就是使得衍射角度变小，三维效果受损。一般方法是在重现的光路中加一个透镜，使得衍射角度增大，但这种方法用在大物体上有困难。

4. 应用前景

计算机生成全息图使得人们可以在同一时刻、不同的地点看到同样事物的像，或同一时刻、同一地点看到同样的事物做不同的运动。这一特点必然会使计算机生成全息图技术在不少领域得到应用。

作为真正的三维图像，全息图比目前二维表达方法多一维信息。采用计算机生成全息图将计算和显示紧密结合在一起，这种信息表达技术的创新必然会对人类社会产生深远的影响。仅从目前的研究水平看，计算机生成全息在影视艺术、全息防伪、远程教育、医学成像、可视化等方面具有诱人的应用前景。

参考文献

1 于美文．光全息学及其应用．北京理工大学出版社，1996.8

2 Hilaire P S, et al. Electronic Display System for Computational Holography. In: Benton S A, ed. SPIE Proc. Practical Holography IV (Soc. Photo-Opt. Instr. Engrs., Bellingham, WA), 1990. 174～182

3 Hilaire P S, et al. Color images with the MIT holographic video display. In: Benton S A, ed. SPIE. Practical Holography VI. 1992. 73～84

4 大越孝敬[日]．三维成像技术．机械工业出版社，1982.12

5 朱自强，等．现代光学教程．四川大学出版社，1990.9

6 Lucente M. Diffraction-Specific Fringe Computation for Electro-Holography: [Ph. D. Thesis]. Dept. of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, September 1994

7 杨振寰，等[美]．光学信号处理、计算和神经网络．母国光等，译．新时代出版社，1997.5

8 Mark L. Interactive three-dimensional holographic displays: seeing the future in depth. Computer Graphics, 1997, 31(2): 63～67

(上接第 19 页)

13 Reimer U, Hahn U. Text Condensation As Knowledge Based Abstraction. 1988. 338～344

14 Copeck T, Delisle S, Szpakowicz S. Parsing And Case Analysis In TANKA. Available at: http: // www. site. uottawa. ca/tanka/ts. html

15 Endres-Niggemeyer B, Neugebauer E. Professional Summarizing. No Cognitive Simulation Without Observation. Journal of the American Society For Information Science, 1998, 49(6): 486～506

16 Radev D R, McKeown K R. Generating Natural Language Summaries from Multiple On-line Sources. Computational Linguistics, 1998, 24(3)

17 吴岩，刘挺，王开铸，陈彬．中文自动文摘原理与方法探讨．中文信息学报，1998，12(2)：5～16

18 新闻报道，中英文自动摘要集成系统通过鉴定．国际电子报，1998-5-25(A8)

19 杨晓兰，钟义信．基于全信息词典的自动文摘系统研究与实现．情报学报，1997(5)

20 麻志毅，姚天顺．基于情境的文本主题求解．计算机研究与发展，1998，35(4)：344～348

21 单永明．一类规范文本篇章结构的自动标引．中文信息学报，1998，12(4)：47～52

22 Text Summarization. Available at: http: // www. csi. uottawa. ca/tanka/ArtDB/bibliography. html

23 刘伟权．自然语言理解与汉语文本信息处理理论研究：[博士论文]．北京邮电大学，1997

24 郭祥昊．语言信息处理理论及自动文摘关键技术的研究：[博士论文]．北京邮电大学，1998