基于选择生成文摘法的自动文摘系统研究与实现

杨晓兰　宋帆　钟义信

（北京邮电大学信息工程系 186#，100088）

摘要： 本文根据文摘与文本结构的关系，提出基于选择生成法的自动文摘系统模型。通过文本选择分析器对与文摘有关的文本部分进行分析和理解，把与文摘生成有关的概念提取出来，统一保存在既定的文摘框架中。文摘生成器根据文摘框架的填充情况生成完整、简洁、可读性好的文摘。基于上述理论的研究，设计并实现了计算机病毒领域的自动文摘实验系统，验证了基于理解的自动文摘系统的设计思想。

关键词： 自然语言理解　自动文摘　选择生成法　选择分析器

Design and realization of automatic abstracting system based on selection and generation method

Yang XiaoLan , Song Fan , Zhong YiXin

(Dept. of Information Engineering , Beijing Univ. of Post and Telecommunication)

Abstract: This paper presents an automatic abstracting model based on selection and generation method which utilizes the knowledge acquired about structure of source texts . Selective parser is specifically designed to make a careful and thorough analysis of the portions of text which may contain the main concepts in the abstract-frame and to fill in as many slots as possible .The completed abstract-frame is then run through a abstract-generator which uses an abstract-template to produce a coherent , cohesive abstract .

Keywords: natural language understanding , automatic abstracting , selective parser selection and generation method ,

一、 引言

文摘是以简洁的形式来表达原始文献的主要内容。如何从原始文献出发，编制出高质量的文摘，始终是没有得到很好解决的问题。由于文摘能反映文献的基本内容，编制文摘工作量大，对文摘员的要求也高。所以过去不少文献没有文摘。尽管现在许多刊物要求作者提供文摘，但其规范性不高。而且作者的文摘可能是不可靠的，带有主观偏见。Evan 、Pollock 和 Marine 等人对不同领域作了调查，发现至少有 20 %的作者文摘是不合适的，未起到文摘的作用。另一方面，在当今一天几乎发表二万篇文献的信息爆炸的年代里，仅仅靠手工文摘已难适应时代发展的要求。于是，自动文摘的研究应运而生。

文摘分为指示性文摘和报道性文摘两大类。我们的研究工作是围绕报道性文摘开展的，它需要对文献作深入的语义分析和理解，并在此基础上对原文作完整的浓缩。

本文分四部分，第一节介绍自动文摘的必要性和文摘的分类。第二节介绍传统的文摘文摘方法。第三节根据文摘与文章结构的关系，提出了基于选择生成法的自动文摘系统模型。第四节是围绕该模型建立的计算机病毒领域的自动文摘系统的实验结果和评价。

二　传统的文摘方法

试图用计算机实现自动文摘的工作始于五十年代 Luhn 的工作。目前人们主要采用的文摘方法有三种：简单提取法、提取重组法和理解生成法。

提取重组法是针对简单提取法所产生的文摘进行润色以提高文摘的可读性。这两种方法最关键技术在于借助于关键词词典及一定的规则，计算句子的权值，根据权值大小从文献中的有关部分自动地抽取候选文摘句，在此基础上选出文摘句。通常采用如下加权抽取方法：提示词加权法，标题加权法，方位加权法，关键词加权法。

理解生成法采用人工智能和专家系统的智能信息处理技术，从文本中提取语义信息，并以一定的中间形式进行表示。然后，将这种表示作为自然语言生成器的输入，产生文摘。该法最先在 DeJong 的 FRUMP 系统被使用，该系统成功地给出了来自众多领域的新闻的文摘。其实现技术核心是一组概要脚本，每个脚本含有一系列的预期事件。概要脚本实际上是对一特定事件类型可能发生的各种活动的详尽顺序列表。

根据评价文摘质量的标准，我们对这三种文摘方法作了比较，如图表一所示。

文摘标准	简单抽取法文摘系统	提取重组法文摘系统	FRUMP 系统
可读性	否	可能	是
简明性	可能	可能	是
完整性	可能	可能	可能
可移植性	是	是	可能

表 1　几种文摘方法的性能比较

三　文摘新法——选择生成法

从表 1 可以看出传统文摘方法不尽人意，但它们性能上有互补之处。本节介绍的文摘

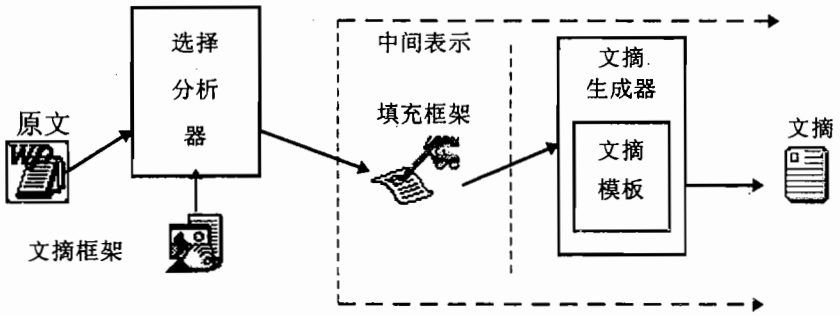

图 2 自动文摘系统流程图

系统是建立在文本理解系统的基础上，采用传统文本理解系统的处理技术，并结合文本结

构与文摘的关系，采用选择生成法，由文本选择分析器选择与文摘有关的文本部分进行详细的语义分析，填充文摘框架，根据文摘框架的填充结果，文摘生成器在文摘模板的基础上生成风格一致、忠实原文、高质量的文摘。这种文摘方法可适用的领域广，可移植性好，而且对复杂的语义分析要求较传统的文摘方法大大降低，利用选择分析器分析文本提高了系统处理效率，增强了系统的鲁棒性。图 2 是文摘系统工作流程图总览。

§ 3.1 文章结构与文摘的关系

文章结构、体裁和语言是文章的形式。早在 1958 年有人研究发现不同的作者在写同一主题的文章时，总是尽可能的使用相同的文章结构。因此，这一类文章可以看作是由若干意义基元（Meaning elements）组成，每个意义基元在层次上、句法上和其他意义基元相联系。以科技文献为例，每篇文章包括研究的背景、目的、方法、结论等意义基元。国外已经有不少学者开始研究在人们在构建文本结构过程中如何使用意义基元组成文本的。同一领域的文章可以用一个由若干意义基元组成的通用模板概括。作者把每个意义基元实例化，按一定的文本结构形式，组成一篇有意义和主题的文章。

我们的文摘模型是从模拟人的文摘编制过程，首先由主题过滤器分析文章的主题，然后根据主题选用合适的文摘框架，在文摘框架的引导下，选取合适的文本部分进行分析，提取出合适的信息填充文摘框架。因此，文摘框架的设计是实现该模型的核心技术。

§ 3.2 文摘框架

许多研究表明，每篇文摘都有其内在结构。若是科技文献，文摘中应包含研究的对象、目的、方法、实验结果和结论等部分。由于不同领域的作者在写文章时，总是尽可能地和该领域文章的通用文章结构相一致，这就给自动文摘系统提供一个契机。如果对每一个领域的文章都能用一个类似文摘框架的东西去描述，对这些文章作文摘只需要有针对性地提取文摘框架中的这些概念。可见选择生成法是与领域无关的，减弱了与领域知识的关系。

我们选取计算机病毒领域的文本作为文摘实验系统的研究对象，通过对有关病毒理论的研究以及具体描述病毒文章的结构分析，决定采用图 3 所示的文摘框架来表示从文本中获取的概念。从图中可看出，组成文摘框架的每个槽部件代表计算机病毒的一个特征，框架中的每个槽可能是短语或文本中的句子。

```
病毒 {
    (病毒名称：（category : NB; value : String ))
    (病毒传染对象：（category : NC *;  value : String ))
    (病毒类属：（category : NK ; value : String ))
    (病毒攻击对象：（category : NA ; value : String ))
    ……
}
```

图 3 计算机病毒类文章的部分文摘框架

文摘框架是实现选择生成法的关键。经研究认为，设计一个好的文摘框架应遵循如下原则：文摘框架对文章的描述是充分的、清晰的、确定的、单调的、易接受的。

§ 3.3 选择分析器

目前有许多分析器可以对非受限领域的文本进行句法分析，但是至今没有哪个分析器能够真正理解一篇文章，哪怕是一篇短的新闻。造成这种情况的部分原因是因为语言现象太复杂，语法分析器的语法模型不可能概括所有语法现象，但最根本的原因是出现在语义上和概念层次上。因为计算机程序很难从文本中获得真正的语义，它缺少必要的概念知识和指代关系。但是，可以设计一个选择分析器，它专门去分析、挖掘某一类特定信息，而对其它主题或概念不进行处理，则可以避开这个矛盾，使问题得以简化。因此，对面向某一任务（如文本分类，信息检索及自动文摘等）的自然语言理解系统采用这种分析器可以提高其处理的有效性和准确性。但选择分析器必须是精确的，即保证即使它的处理范围增大，它以前对特定文本部分的分析结果不会发生变化。

我们希望在对文摘实例化的过程中，尽可能地避免复杂的句法分析和语义解释。毕竟文摘与文本不同，文摘包含文本的中心内容，而文本中总是含有许多“噪声”，在填充文摘框架的过程中，分析器应有能力区分哪些句子是与框架的填充有关的，哪些句子是不需要进行分析的。正是考虑文摘只是文本的一部分，分析器应该而且必须是选择分析器，它可以只对文本重要部分的关键句子甚至是短语进行分析。

以计算机病毒领域的文摘系统为例，假设它对计算机病毒的特征比较了解。选择分析器第一遍扫描将过滤出可能含有与文摘框架信息提取有关的句子。第二遍进一步过滤出值得分析的短语，将分析得到的信息填入的文摘框架的相应槽中。如下面这句话“该病毒是一种恶性病毒”，句中“该病毒”肯定是指某一种计算机病毒，与医学上的病毒无关，至于病毒的特性“恶性”肯定是指该病毒的破坏意图是恶性的，将导致计算机系统不能正常运行等等。之所以可以产生这些解释，是因为建立了该领域的领域模型，文摘系统有该领域的背景知识，分析器一旦看到这些概念，它便会产生合理的解释，而不会产生其它歧义。例如下面这句话：“在染毒程序运行时，病毒常驻内存”。由于它是描述病毒的驻留特性的，故作为与提取信息有关的句子被保留进行下一步分析，此时分析器不对“在……时”引导的时间介词短语进行分析，仅对后半句分析，得到其具体语义“该病毒驻留内存”。至于分不分析“在……时”短语对系统的最终目标并不会有什么影响。

因此选择分析器实现的关键在于如何从原文中提取信息来填充文摘框架。为此我们设计了全信息词典来解决该问题。这是一部对词语进行全方位（语法、语义、语用）描写的词典。它将多种知识源有机地组成一体，用统一的文法进行描述，很好地解决了自然语言中知识表示和知识运用的问题。全信息词典由三部分组成：概念词典，概念联用规则库，和效用规则库。概念词典是词典的静态部分，它其实是以分类语义场为基础的多层次、多类型静态语义网络，用来分析概念的聚合关系。概念联用规则库是词典的动态部分，它是基于框架的动态语义网用来分析概念间的组合关系。效用规则库是与面向具体应用的程序有关，把词和任务有机地联系在一起。词不再是一个抽象的意义单元，而有其语用意义。因此，该词典一方面可以引导句子分析器进行语义分析，另一方面它为选择分析器的部分

工作机理提供了依据。不同的领域有不同的概念，对应不同的概念联用规则和效用规则。

全信息词典的数据结构为：

词	语法语义范畴	概念联用规则	映射规则	语义类限制规则	效用规则

语法语义范畴是指词的分类方法是按语法语义分类的。概念联用规则主要是针对动词、形容词、名词和介词制定的，体现了概念之间互相制约互相影响的语义约束关系，它是和映射规则、语义类限制规则联合作用引导句子进行语义分析，得到语义解释。概念联用规则中定义了六种语义关系，如施事、受事。映射规则中有八种语法成分，如主语、谓语等。

词汇的概念联用规则和效用规则库，它是选择分析器算法的基础，也是填充文摘框架的基础。每个词的效用规则是根据它们与文摘框架槽的填充关系编写的。因此利用词或短语的效用规则来选择合适的句子进行分析，将分析的结果填入文摘框架中。词或短语的效用规则的制定是基于语料库的技术，找出与文摘框架中槽的填充有关的短语或词，制定相应的判定规则，选择正确的信息填入文摘框架。目前规则的制定方法综合利用了传统文摘方法中的文摘句的四种加权方法，不同之处在于它把权值反映在词语和特征短语上，而不是句子级了。

有两种类型的效用规则：特征短语和关键词的效用规则。句子中如出现特征短语就意味着这句话含有重要的信息，可以填充文摘框架的某一个槽。例如，根据有特定意义的短语，如“感染……”，该短语表示病毒的感染目标，而文摘框架中要求提取病毒的感染目标信息，选择分析器根据这一线索对这句话进行分析，提取出合适的信息填入相应的槽中。上述效用规则是基于特征短语。还有一种效用规则是基于特征词的，如“恶性”，若分析器遇到这个词，它则去分析这个句子，检查其主语是否是一种计算机病毒，若是，这在病毒的文摘框架的破坏意图槽中填入相应的值。

可见，选择分析器是把传统的自顶向下（又称基于期望的驱动分析方法）和自底向上（又称基于语言驱动的分析）算法结合在一起，共同作用于待处理的句子。首先，选择分析器在文摘框架的基础上，根据句子中的某些线索来决定是否有分析的必要，其本身的运行是基于期望驱动的。如果发现句子可能含有重要信息，则用自底向上方法对语句进行详尽的分析，得到意义解释，决定是否有信息填入文摘框架。

选择分析器可以同人一样有针对性地挑出感兴趣的句子进行最少的必要的分析，故它效率是较高的。这种智能的实现是靠全信息词典和文摘框架的应用。另外，利用概念联用规则和上下文寄存器较好地解决了代词回指、话语实体确定和省略句等语言学问题。鉴于篇幅受限，全信息词典的生成方法和选择分析器实现算法将另文介绍。

§ 3.4 文摘生成器

文摘生成器有两个功能，一是对文摘框架的再填充（补充填充），二是在填充的文摘框架的基础上生成文摘。由于文摘框架中的某些槽具有逻辑关系，因此尽管有些槽没有被填充，但可以根据其它槽的填充情况，文摘生成器根据规则尽可能补充填充更多的槽。如根据“如果病毒的感染对象是文件，那么它属于文件型病毒”的规则，可以确定病毒的类属。通过制定与算法分开的推理规则库，可以增强系统的可移植性、智能性。

文摘模板是一个包含文章重点内容的框架，图 4 是计算机病毒领域的文摘模板。文摘模板中的空白部分的名字总可以找到与之对应文摘框架的槽名。因此，把文摘框架的填充结果对应填入文摘模板的空白部分，便可以产生一篇风格良好、可读性好、质量高的文摘。

正如文摘框架是同--领域文章内容的超框架，文摘模板也是一个超文摘框架。

句子编号	控制	文摘片断
1	OR	本篇文章介绍了 病毒名称 的特性。
2	OR	这是一种病毒破坏意图病毒，它属于病毒类属 。
3	OR	它感染 病毒传染对象 。
4	OR	它的破坏性有 病毒破坏性 。
5	OR	它的攻击对象是 病毒攻击对象。

图 4 计算机病毒类文章的部分文摘模板

四 结论

利用上述文摘系统模型，我们设计并实现了计算机病毒领域的文摘系统。本系统已经用十篇计算机病毒类的文章进行了测试，均成功地提取出文摘。由这十个框架实体自动生成了一个计算机病毒知识库。现将其中一篇文章分析结果提供如下。

原文： Sunday 病毒是攻击 IBM PC 及其兼容机的。有时，这种病毒叫"快乐的星期五"病毒。该病毒入侵系统之后驻留于系统的内存，监视系统的运行，并寻找系统中运行的 COM 文件及 EXE 文件以及系统的覆盖文件。受其传染的系统程序长度增加 1636 字节。该病毒是一种恶性病毒。它影响系统的运行，破坏系统执行的程序和系统的覆盖文件。

文摘： SUNDAY 病毒是一种恶性病毒。本文叙述了该病毒的破坏性和传染性。该病毒驻留内存。它属于文件型病毒。

病毒文摘框架填充略

参考文献

【1】 Edmundson , H.P. New Methods in automatic Extracting . Journal of Association fo Computing Machinery . 16(2) : 264-285 , 1969

【2】 Baxendale , P.B. Man-made Index for Technical Literature - An experiment I.B.M journal of research and Development . 2(4) : 354-361 , 1958

【3】 DeJong , G . " An overview of the FRUMP System " In Strategies for Natural Language Processing , Lehnert and Ringle (eds.) 1982 . Hillsdale , NJ : Lawrene Earlbaum Associates

【4】 吴应天，《文章结构学》，中国人民出版社，1986 年

【5】 赖茂生，徐克敏，《科技文献检索》，北京大学出版社，1994 年