中 文 信 息 学 报
第 12 卷 第 2 期 JOURNAL OF CHINESE INFORMATION PROCESSING Vo1. 12 No. 2

中文自动文摘原理与方法探索①

吴 岩 刘 挺 王开铸
哈尔滨工业大学计算机系
陈 彬
哈尔滨医科大学基础医学院计算机教研室

摘要 本文首先介绍了自动文摘的研究情况及存在问题,然后给出了计算机自动文摘的一般模型,最后介绍了我们所研究的两种自动文摘的原理和方法,及其实验结果。

关键词 自动文摘 机械文摘 理解文摘

一、引 言

几乎从计算机诞生之日起,国外就开始进行自动文摘的研究工作。所谓自动文摘,就是由计算机将文本的中心思想,或用户所需要的内容用同于或不同于原文的句子自动提取出来。

迄今为止的自动文摘系统大体上可分为两类:基于统计的机械文摘和基于意义的理解文摘。前者主要根据线索词词典、词频、词和句子的启发函数进行模式匹配提取文摘,其代表人物为 Luhn,研究成果有 Luhn 系统和 Oswald 系统等八个系统[1]。后者则是利用句法和语义知识、或一阶谓词逻辑、CD 理论等对文章的内容,在理解的基础上提取文摘来。这类文摘的代表成果有 Yale 大学的 FRUMP 系统和 J. I. Tait 的 Scrable 系统等八大系统[2]。

中文自动文摘的研究起步于八十年代,1991 年上海交大的王永成教授领导的课题组研制成功了中文文献自动标引系统,该系统可将一篇关于科技情报的长达 7000 字的文章,自动摘出 300 字要点。从 90 年开始,在国家 863 的资助下,哈工大的王开铸教授领导下的课题组就一直进行自然语言理解—自动文摘的研究工作,迄今为止已推出了两种中文文摘实验系统:基于理解的 MATAS 文摘系统和基于统计的 HIT - 863 I型自动文摘系统,本文主要介绍这两种系统的原理及方法。

二、计算机自动文摘一般模型

由计算机自动提取出任意文本(除小说、诗歌外)的文摘,或文章的中心思想,其实质是

① 本文 1997 年 7 月 18 日收到

计算机对文章从某个角度进行分析、理解(理解的深度不一)的基础上得来的。实现这个过程的一般模型如图 1 所示:

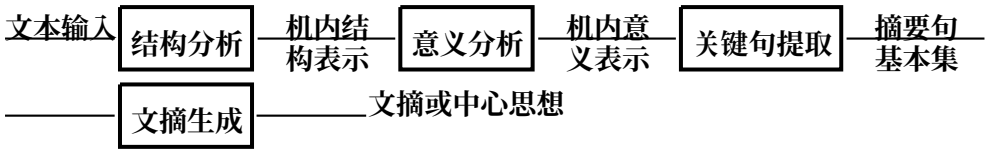


图 1. 自动文摘一般模型

上述模型反映出的全过程由四个主要映射函数有机组合而成。设任意文本 f ,经结构理解映射函数 SU ,获得结构机内表示 sr ,即

$$sr = SU(f) \tag{1}$$

在 sr 基础上,经意义理解映射函数 MU ,获得机内意义表示 mr ,即

$$mr = MU(sr) \tag{2}$$

第三个映射函数是关键句基本集映射函数 KS 对机内意义表示进行转换,获得摘要句基集 as ,即

$$as = KS(mr) \tag{3}$$

最后,经过文摘生成映射函数 AG ,获得文摘内容 a ,即

$$a = AG(as) \tag{4}$$

三、基于意义的理解文摘 MATAS[3]

(一)系统的实现过程

要想在理解文章的基础上提取文摘,不但要理解文章的词、语句、语句间关系,而且还要理解文章的总体结构,基于这一设想,MATAS 的摘要过程如下:

1. 源文分析:本过程对输入的汉语源文本(D)经过词处理阶段、句处理阶段、上下文处理阶段生成篇章意义机内表示 TMR:

```
TMR ::= text
text ⇒P-title:sentence-literal          文章标题
       P-style:style                    文章体裁
       P-author:author                  文章作者
       P-source:sentence-literal        文章出处
       M-title:sentence                 标题意义
       index:Index-List                 篇章检索链
       str-sum:Integer                  段数
       str-first-p:paragraph            段指针
Paragraph ⇒P-number:Integer             段号
          M-relation-p:PARAGRAPH-RELATION   段间关系
          str-first-s:sentence              句指针
```

str-next-p :paragraph　　下一段指针

Sentence ⇒P-srcops :source-position　　源位置

str-modality :seq of MODALITY　　情态结构

str-proposition :proposition　　命题结构

str-next-s :sentence　　句指针

M-relations :SENTENCE-RELATION　　句间关系

Proposition ⇒str-v :Verb　　动词修饰

str-c :seq of Case　　格结构

Verb ⇒V-literal :Verb-literal

Verb-modify :seq of MODV

Case ⇒case-type :CASE

n-literal :noun-literal

noun-modify :seq of MODN

其中,Verb-literal,noun-literal,sentence-literal,style,author,lndex-List 是物理属性私有类型,其表示依赖于实现。MODALITY,CASE,MODN,MODV,SENTENCE-RELATION,PARAGRAPH-RELATION 都是逻辑属性私有类型,其具体表示亦依赖于实现。Integer 是固有类型。

TMR 的最小操作集定义如下:

(1) make-node:创建树结点。(2) delete-node:删除树结点。(3) node-type:结点类型判断。(4) get-attribute:取结点属性。(5) fill-attribute:填入结点属性。

应用这一数学模型,源文的机内意义表示为分层加权有向图,图中的结点分三类:篇结点,段结点和句结点。结点属性有物理属性,逻辑属性。物理属性是文章的外观结构,如文章标题、物理位置等,逻辑属性是文章的意义表示。

〈1〉词处理阶段首先用最大匹配法进行分词,然后通过查词典获取有关词性及语义表达等信息:

〈2〉句处理阶段用短语结构 ATN 分析,最后通过候选排斥法确定句子的格关系。NP 定义如下:

NP + 的 + NP →NP　　n →NP

n + NP →NP　　NP + NP →NP

数词 + 量词 →NP　　Det + n →NP

Pronoun →NP　　Det + NP →NP

NP + conj + NP →NP

VP 定义如下:

V →VP　　V + VP →VP

VP + VP →VP　　VP + conj + VP →VP

adv + VP →VP　　NPloc + VP →VP

NPtime + VP →VP

介词短语 PP 定义如下:

方位词 →PP →Loc(方位词)

在 + NP + {"时"→PP →Time("时")}

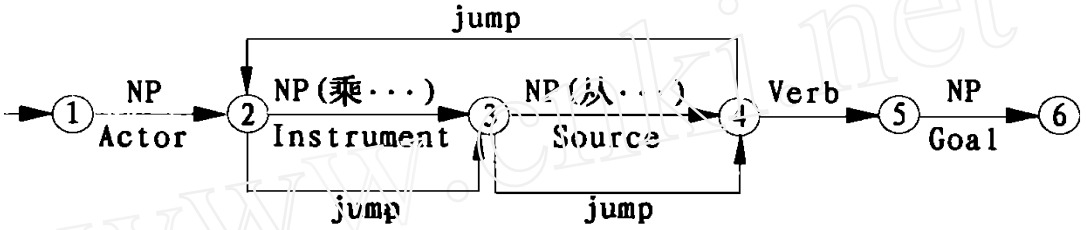
当 + NP +"时"→PP →Time

介词 + NP →PP

格关系的确定即是对句子中的所有 NP 确定其在具体的语句中的格角色,对每个名词我们将其格角色候选 N(C1,…,Cm)置入词典入口中,根据表序我们对每个 Ci 定义了格删除权,对 i > j,W(Ci) > W(Ci),也就是说 Ci 比 Cj 更易排斥掉,格关系的确定分如下几步进行:

①若 NP 在句中带有介词 P,其格标为 P(C1,…,Cn),那么 NC(C1,…,Cm) ∩P(C1,…,Cn) ⇒NC(C1,…Ck)为新的格角色候选,其中 k < m。

②语序即是句子的结构,我们使用转移网来表示语序,对每个词类我们创建一个转移



网,如动词"到达"类(到达,抵达,来到…,),其转移网如下:由转移网分析后,NP 的候选集为[NC1(C11,…,C1i) …,NCn(C11,…,C1j)],(i,j < m)

③对每个动词有格框架规定格是必需的、可选的、排斥的。排斥格充当过滤器来排斥掉候选格集的不允许格,FILTER(C1,…,Cl)为排斥格集:

NCi(Ci1,…,Cik) - FILTER(C1,…,Cl) ⇒NCi(Cl,…,Cj)

④本步使用语义信息来排斥候选格,如动词"喝"的语义限制 SR 如下:

SR(施事格)必须是动物

SR(客体格)必须是液体

SR(工具格)必须是容器

⑤选 C1 为 NC 的格角色

〈3〉上下文处理阶段经过由关联词制导的复句分析,和词间联系制约的一般句间关系分析,以及人机交互得到文章的体裁模式,最后得到扩充的篇章意义机表示 TMR,扩充的 TMR 是在〈1〉的 TMR 基础上生成的,两者的不同点在于结点 sentence 的机内表示上,〈1〉中的 sentence 是文章中的原句,而扩充的 TMR 的 sentence 则是〈2〉所得到的句子意义的格关系表示。

2. 信息压缩:本步对 TMR 进行两级信息压缩:句子级和上下文级,然后再根据词频等特性对句子重要性进行加权,最后由仲裁判断对句子进行删留处理。

〈1〉句子级信息压缩使用过滤函数 filter,删除句子的部分辅助成份,而得到句子的主干。filter 形式描述如下:

句子 S = V(Cv ……Cn) ……(意义格表示)

V = MODv(v) v ……(动词及其修饰成份表示)

Ci = MODn(ni) ni ……(中心名词及其修饰成份表示)

过滤函数 filter:

$$S1 = filter(S)$$
$$= (MODv(v) - AUXv(v))\,v$$
$$= (MODn(n_1 - AUXn(n_1))\,n_1$$
$$\cdots\cdots$$
$$(MODn(n_n - AUX_n(n))\,n_n$$

从上式可见,filter 主要是对 AUXv,AUXn 的研究,即名词和动词修饰成份的分析。

〈2〉上下文级信息压缩:上下文结构主要体现于句间关系和篇章结构。本系统定义了八种句间关系:并列(Par),承接(Con),选择(Select),递进(Dj),转折(Zhz),因果(Cause),假设(Prem),条件(Cond)。对于满足上述八种关系之一的句子 S1 和 S2,本系统规定了不同的加权标准,以为摘要作准备。对于段落的重要性,本系统按以下的结构模式加权:

①说明文结构模式:

分析型 —总说(1) + 分说(2) + …+ 分说(n) →

W(P1) > W(P2) = W(P3) = …= W(Pn)

综合型 —分说(1) + 分说(2) + …+ 分说(n - 1) + 总结(n) →

W(P1) = W(P2) = …= W(Pn - 1) < W(Pn)

分说型 —分说(1) + 分说(2) + …+ 分说(n) →

W(P1) = W(P2) = …= W(Pn)

②议论文结构模式:

归纳型 —分论(1) + 分论(2) + …+ 分论(n - 1) + 结论(n) →

W(P1) = W(P2) = …= W(Pn - 1) < W(Pn)

演绎型 —总论(1) + 分论(2) + …+ 分论(n) →

W(P1) > W(P2) = W(P3) = …= W(Pn)

演归型 —总论(1) + 分论(2) + …+ 分论(n - 1) + 结论(n) →

W(P1) > W(P2) = …= W(Pn - 1) < W(Pn)

分论型 —分论(1) + 分论(2) + …+ 分论(n) →

W(P1) = W(P2) = …= W(Pn)

③叙述文中分总叙章、分叙章、结尾章,其中 W(总叙章) > W(分叙章),W(分叙章) < W(结尾章)。

④描写文中分总描章、分描章、主从章,其中 W(总描章) > W(分描章),W(分描章) < W(主从章)。

⑤复合体系无统一模式。

3. 正文生成

本过程由篇章意义机内表示生成自然语言语句,它包括三步:

〈1〉预处理阶段进行介词扩充,并处理工具格、方位格、源格、目的格、时间格。

〈2〉格角色指派阶段将格成份转化成语法成份。

〈3〉语序重组阶段对语序进行重新调整,得到自然语言语句。

(二)系统的实验实例

以下是发表于 1985 年 12 月 4 日解放军报上的“战术导弹将改变传统的作战方式”一文,作者:方清良。

源文如下:

不断发展并日益成熟的战术导弹,已经成为现代战场角逐的重要角色。

在第四次中东战争中,双方损失飞机四百四十八架,坦克二千五百辆。其中百分之八十是被导弹击毁的。双方损失军舰五十艘,全部是被导弹击毁的。

在马岛战争中,阿空军“超级军旗”式战斗轰炸机用“飞鱼”式导弹,在三十多公里距离上发射,击毁了“谢菲尔德”号驱逐舰。同样,英伞兵第一营在进攻达尔文港时,用法制米兰“反坦克导弹”,一举摧毁了阿军的博克豪斯掩体阵地,为占领该港打开了通路。

基于上述事实,许多军事评论家指出,导弹正改变着现代战场的格局。传统的作战方式将受到冲击,新的作战方式将逐渐形成,多种类型的兵力布势和交战队形将相应产生。

摘要结果为:

战术导弹成为现代战场角色。军事评论家指出,导弹改变现代战场格局。传统作战方式受到冲击,新作战方式形成。

四、基于统计的机械文摘 HIT - 863 I型系统[4]

(一)系统的实现过程

HIT - 863 I型系统是个能处理任意文本的自动文摘系统,且文摘结果可以是任意比例。其摘要过程分如下几步:1. 源文接受过程用于将输入源文本转换成系统定义的机内表示形式 —层次结构网络(HN),HN 的篇章表示如图 2 所示:

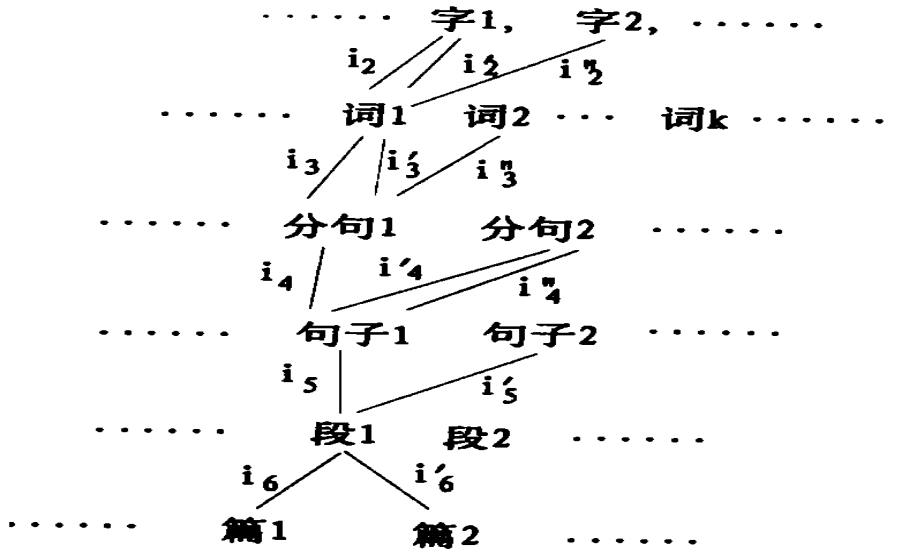


图 2

HN 方法基于自然语言本身的层次结构。具体来说,对于 Ti 层,其中的元素之间无关,而 Ti 层和 T_{i+1}层元素之间用加权有向边相联。设$\alpha_i \in T_i$,$\alpha_{i+1} \in T_{i+1}$,如果α_i是α_{i+1}的第 k 个组成元素,即$\alpha_{i+1} = \beta_1\beta_2 \cdots \beta_{k-1}\alpha_i\beta_k \cdots \beta_n$,$\beta_j \in T_i$,j = 1,2,…,n,则将$\alpha_i$与$\alpha_{i+1}$用一条加权有向边$(\alpha_i \xrightarrow{k} \alpha_{i+1})$相联。因此,对于篇 1 的第 16 个自然段的第 15 个句子的第 14 个分句的第 13 个单词的第 12 个字 C,可以用一个 n 元组(1,i6,i5,i4,i3,i1,C)唯一表示,它对应于层次结构网络中从 C 到篇 1 的一条通路。2. 基于部分封闭词的自动分词是出于这样的考虑:在一篇文章中,相邻共现次数越多的字,它们越可能是一个词,本过程首先使用封闭词将源文本中的句子分成一个个词串,然后应用加权函数为每个词串加权,加权函数为 $W = F * L^c$,F 是词频,L 是词长,c 是大于 1 的正整数。另外,考虑到字的随机共现情况以及单个汉字成词的不稳定性,F 应大于等于某一个下限 min,L ≥2。一个字符串的权值越大,越可能是一个词,那么在进行句子切分时就可用贪心算法将高权值的字符串挑出来,作为一个切分出来的词,而对剩下的部分再进行切分,这样递归进行,直到不可分为止。3. 文摘生成过程通过特征词提取,句子重要性动态测度等过程生成文摘。

〈1〉特征词提取:文章中的特征词具有以下的性质,特征词出现的范围有规律,且是名词或名词词组,并具有新颖性。综合这些考虑,可在分词的基础上进行特征词提取,并为特征词加权,本系统的特征词加权函数为 $P = K(F - min)(L - D)^c$,这里 P 是单词的加权函数,K = 1,表示该词是在上下文中多次出现,是特征词;K = - 1,表示该词在背景知识中多次出现而在上下文中很少出现,不是特征词。F 是词频,F > min,min 为频率下限,L 是词长,D 是词长下限,c 是一个常数。

〈2〉句子加权:本系统设计了一个句子重要性动态加权函数来为每个句子加权,该函数为:

$$T = K \cdot \frac{\sum_{i=1}^{N} Ti}{S \cdot S_1 \cdot S_2}$$

其中,K 为系数,一般为 1;T_i,i = 1,…,N,为特征词 i 的特征值;S 为句子中总的词数,S_1为分句个数,S_2为数字符号个数的 m 倍,m 为一常数。

由于该句子加权函数,不涉及语言的类型,文章的体裁,文章的专业内容,文章的长度等,因此可以满足非受限域文本的自动提取文摘的要求。

〈3〉按比例提取文摘:本过程根据句子权值的高低,按用户要求的比例提取出文章的摘要并输出。

(二)系统的实验实例

对于“战术导弹将改变传统的作战方式”一文,各种比例的摘要结果如下:

·10 %的摘要:

不断发展并日益成熟的战术导弹,已经成为现代战场角逐的重要角色。

·20 %的摘要:

不断发展并日益成熟的战术导弹,已经成为现代战场角逐的重要角色。

在第四次中东战争中,双方损失飞机四百四十八架,坦克二千五百辆。其中百分之八

十是被导弹击毁的。

·30 %的摘要:

不断发展并日益成熟的战术导弹,已经成为现代战场角逐的重要角色。

在第四次中东战争中,双方损失飞机四百四十八架,坦克二千五百辆。其中百分之八十是被导弹击毁的。双方损失军舰五十艘,全部是被导弹击毁的。

在马岛战争中,阿空军"超级军旗"式战斗轰炸机用"飞鱼"式导弹,在三十多公里距离上发射,击毁了"谢菲尔德"号驱逐舰。

五、结 论

下面我们根据两种文摘系统的实验结果从五个方面来讨论一个机械文摘和理解文摘的优缺点。

处理文本的范围:目前 MATAS 只是针对上述特定文章作了实验,若用 MATAS 对其他文章提取摘要,必须对 MATAST 系统内部进行完善,如增加词典词条,增加规则等。所以,作为理解文摘的 MATAS 具有处理文本领域受限的缺点。而 HIT - 863 Ⅰ型系统却能处理任意文本,这也是机械文摘的一大优点。

处理文本的输入形式:这两种文摘都能处理自然语言的文本文件,这是 MATAS 优于其它理解文摘之处,如 J. I. Tait 的 Scrable 理解文摘系统,必须将自然语言的源文由人工转换成 CD 结构表示才能处理。处理的不是自然语言的源文,是某些理解文摘的一大缺点。

采用的处理方法:理解文摘 MATAS 采用格文法、语法语义规则、知识库等处理手段,对输入的源文本进行基本理解的基础上提取文摘句,其实现过程复杂而费时,对上述文本,系统要耗费 2 分钟才能提取出文摘;机械文摘 HIT - 863 Ⅰ型系统利用模式匹配、启发函数、词频统计等方法提取摘要句,其实现过程简单而快速,对上述文本,系统只要 10 秒就可完成文摘的提取过程。

·文摘形式:理解文摘 MATAS 的文摘句完全是系统自动组织的语句,而机械文摘 HIT - 863 Ⅰ型系统的文摘句则是文本中的原句。

·文摘质量:从上述文摘结果我们可以看到,理解文摘 MATAS 的文摘句简洁精练,且反映了源文的中心思想。而机械文摘 HIT - 863 Ⅰ型系统的文摘中,20 %和 30 %的文摘结果都含有举例成份,如"在第四次中东战争中……"和"在马岛战争中……"都是文摘不应包含的内容。所以,HIT - 863 Ⅰ型系统的文摘结果不甚精练,应用此系统对大量语料的实验结果表明,HIT - 863 Ⅰ型系统的文摘结果亦存在机械文摘的语句冗余,上下文缺乏逻辑联系的问题。

从上面的分析可见,机械文摘的优点在其处理前端,缺点在其处理后端,而理解文摘与此相反。所以,目前我们设计了一个新的自动文摘系统。该系统是机械文摘和理解文摘的一个集成,即其处理前端用机械的方法,以吸收机械文摘的可处理任意文本且容易实现的优点。该系统的后端使用理解的方法,文摘结果为系统自动组织的语句,有效地吸收了理解文摘质量高的优点。该系统目前正在实验中,且效果良好,相信在不久的将来可达到实用。

参 考 文 献

[1] Luhm, H. P., The automatic creation of literature abstracts, IBM J. of Res. and Development, 1958, (2), 159 - 165

[2] Edmundson, H. P., Wyllys, R. E., Automatic abstracting and indexing - survey and recommendations, Comm of the ACM, May 1961, 4(5), 226 - 234

[3] 王建波, 王开铸, 自然语言篇章理解及基于理解的自动文摘研究, 《中文信息学报》, 1992, 6(2), 1 - 7

[4] 李俊杰, 王开铸, 任意文本文摘的自动抽取的研究, 哈尔滨工业大学学报, 1995, 2

[5] 姚天顺, 《计算机汉字信息处理》, 辽宁科学技术出版社

[6] 黄昌宁, 孙茂松, 汉语句法分析的一种多遍扫描确定性算法, 中文及东方语言处理国际会议, 长沙, 1990. 4

[7] 徐志敏, 自然语言处理进展, 《中国计算机用户》, 1988, No. 5, PP. 11 - 14

[8] A. Mathis, Techniques for the Evaluation and Improvement of Computer Produced Abstracts, Ohis State University, Dec, 1992, PB214675

[9] Jane Morris, Graeme Hirst, Lexical Cohesion Computed by Thesaral Relations as an Indicator of the Structure of Text, Computational Linguistics, Vo1. 17, No. 1, March, 1991.

[10] M. Rayner, A. Banks, An Implementable Semantics for Comparative Constructions. Computational Linguistics, Vo1. 16, No. 2, June, 1990

[11] Kathleen Dahlgren, etal, Knowledge Representation for Commonsense with Text, Computational Linguistics, Vo1. 15, No. 3, Sep. 1989

Research on the Method of Chinese Automatic Abstracting

Wu Yan Liu Ting Wang Kaizhu

Dept. of Computer, Harbin Institute of Technology

Chen Bin

Teaching and Research Section of Computer, Harbin Medical University

Abstract First, the paper introduces the research situation and problem of automatic abstracting, then gives general model of computer automatic abstracting, at last, introduces the principles and methods of two kinds of automatic abstracting of our research, and their experimental results.

Keywords automatic abstracting mechanical abstracting understanding abstracting