自动文摘系统EAAS

李小滨 徐越

(中科院软件所)

EAAS: AN AUTOMATIC ABSTRACT SYSTEM

Li Xiaobin and Xu Yue

(*The Institute of Software, Academia Sinica*)

ABSTRACT

This paper introduces an English automatic abstract system EAAS, describes its architecture, internal representations, algorithms and background knowledge.

摘 要

本文介绍了一个英文自动文摘系统EAAS，详细描述了其总体结构、各环节的内部表示和算法、以及背景知识的组织和表示。

§1. 引 言

自动文摘的广泛应用前景是不言而喻的。但自从十年前美国Yale大学首次开展对自动文摘的研究以来，越来越多的学者从事这方面的研究却至今未见十分成功的报道。因此，要探讨这样一个具有相当难度的问题，必须对其性质、背景以及困难进行客观的分析。

自动文摘是一类特殊的自然语言理解问题。语言的层面模型的观点指出，语言具有三个主要层面：结构层面、意义层面和功能层面。由于对语言各层面的研究至今尚很不充分，自动文摘就难免面临诸方面难以逾越的障碍。首先在意义层面上，由于语言可以有许多比喻性用法，对其意义进行了不同的引申，语句里各词的词义不是几个范畴能包括的，故准确地把握语言的意义十分困难；其次在功能层面上，由于语言的功能过于广泛致使歧义问题十分突出。因此，基于目前的语言研究水平，只有采取一些避开这些困难的有效对策才能使当前对自动文摘的研究不至于重蹈旧辙。

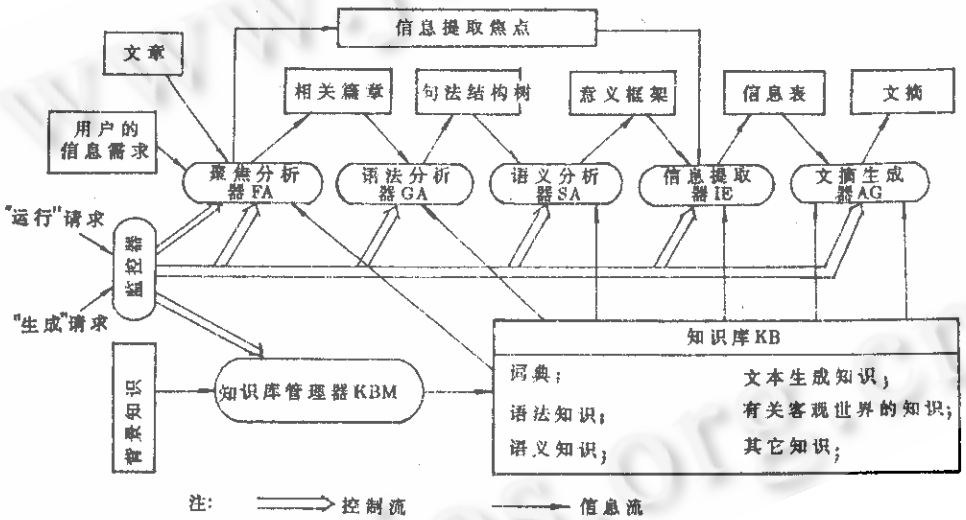
1989年8月25日收到，1990年3月26日定稿。

对策之一是将自动文摘局限在意义较单一的范围内进行，以避开意义层面上的困难。因特定领域里语言一般比喻、引申的用法较少，大多采用了准确的表达，易于被计算机系统所理解。对策之二是从合适的角度固定语言使用的客观环境之后才进行自动文摘，以避开功能层面上的困难。显然将自动文摘系统固定在文章的读者(即系统用户) 的角度上，按用户的需要从文章中摘取信息而不考虑文章中的其它内容，可以避开确定文章主题及处理文章中出现的新概念等目前难以圆满解决的问题，获得更大的成功。

基于上述切实可行的对策，两年来我们在研究自动文摘方面取得了初步的成果：利用基于知识的方法设计并实现了一个实验性的英文自动文摘系统EAAS。下文将介绍EAAS 系统总体结构、各主要环节的内部表示和算法思想、以及背景知识的组织和表示。

§2. EAAS 系统的总体结构及工作流程

EAAS 系统由监控器、聚焦分析器FA、语法分析器GA、语义分析器SA、信息提取器IE、文摘生成器AG、知识库管理器KBM 及知识库KB 组成(见EAAS 系统总体结构图)。



EAAS 系统总体结构图

EAAS 系统有三种工作状态：监控态、生成态、运行态。在监控态下，系统用户(包括知识工程师和文章的读者) 可以提出“生成”请求或“运行”请求，在监控器的调度下进入系统的生成态或运行态。在生成态下，知识工程师可通过KBM 以人机交互的方式生成KB(将背景知识装入KB) 或维护KB(对KB 中的知识进行增、删、改等)。在运行态下，文章的读者可以将一些文章和自己的信息需求输入系统，在KB 中各类知识的支持下，FA 在这些信息需求的驱动下确定文章所属领域和信息提取焦点并筛选出与焦点相关的篇章和段落，GA 和SA 接着进行语法和语义分析，产生出这些篇章段落的一种内部表示，IE 则在此基础上根据焦点进行信息搜索、推理及归纳，得出一个信息表，最后AG 将此信息表转换为一篇文摘。

下面将着重介绍运行态下系统各环节的工作。

§3. 聚焦分析

聚焦分析是为了了解用户对信息的需求，确定输入文章的应用领域和信息提取焦点集F，并根据F 对文章进行筛选，从而避免处理与F 无关的内容，达到提高系统效率的目的。

设用户所需信息项为$inf_1, inf_2, \cdots, inf_n$，则$F = \{f_1, f_2, \cdots, f_n\}$, $inf_i \in f_i$，f_i 实际上是包含inf_i 的最小信息范畴(这里称为焦点)，$i = 1, 2, \cdots, n$。

聚焦过程涉及KB 中以下几类知识:

1. 信息范畴体系ic

$ic ::= \{sic_1, sic_2, \cdots, sic_m\}$　　$sic_k ::= ic \mid <$ 终结信息范畴 $>$　$k = 1, 2, \cdots m$

2. 聚焦规则FR

$FR ::= P \rightarrow C, cf$　　$P ::= ((ca : se)|f)OP\{((ca : se)|f)\}^*|f$

$ca ::=<$ 逻辑格 $>$　　$se ::=<$ 语义参数 $>$

$OP ::=<$ 操作符 $>$　　$C ::= f|(ca : se)$

$cf ::=<$ 可信度 $>$

$f ::=<$ 焦点名 $>$

3. 细节约束框架CFR

$CFR ::= (<$ 焦点名 $> \{(SN \quad VA)\}^*)$

$VA ::= \{se_1, se_2, \cdots, se_m\}|n|\{OP \quad se\}^*$

$se_i ::=<$ 语义参数 $>$　　$n ::=<$ 整数 $>$

$OP ::=<$ 操作符 $>$

$SN ::=<$ 约束语义参数集 $> | <$ 关键词语义参数集 $> |$

$<$ 约束条件个数 $> | <$ 约束条件 $>$

聚焦分析器FA 的工作分为两个阶段:

1. 确定信息焦点: FA 通过交互获得一些待处理的文章和用户用自然语言提出的信息需求re，对re 进行语法语义分析之后得到其意义框架remf(关于语法分析和语义分析过程见 §4 和 §5)，然后在remf 之上运用聚焦规则fr 和信息范畴知识推出焦点集$F = \{f_1, f_2, \cdots, f_n\}$，最后利用细节约束框架cfr 产生$f_i$ 的相应焦点框架$ff_i, i = 1, 2, \cdots n$。

2. 筛选输入文章: FA 对输入文章各段落的首句进行语法、语义分析得到其意义框架smf，再利用fr 产生其涉及的焦点集F_1。如果$F \cap F_1$ 为空，则删除该段落；否则选择$F \cap F_1$ 中可信度最高的焦点fg $(1 \leq g \leq n)$ 为该段落的焦点，并采取关键字匹配法检查段落中的每个句子，看其是否满足fg 的焦点框架ffg 中的约束条件，从而决定该段落的取舍。

§4. 语法分析

语法分析器GA 以KB 中的词典及语法规则为背景对文章中的语句进行语法分析，确定文章中各词的词义、词的形态、词之间句法上的联系及语句的各语法成分，并以一棵层次结构的语法树描述之。

语句的句法结构树ST 定义如下:

$ST ::= \{< SU >, < MV >, < IO >, < DO >, < MOD >, < PD >\}$

$< SU > ::= NP|NC|ST$　　$< MV > ::= \{aux, verb\}$

$< IO > ::= NP|NC|ST$　　$< MOD > ::= PP|ST|adv$

$< DO > ::= NP|NC|ST$

$< PD > ::= NP|NC|ST$　　$NC ::= \{NP, conj\}$

$NP ::= \{< art >, < HEA >, < DES >, < QUA >\}$

$< HEA > ::= NP|NC|noun|pron|proper|ppn$

$< DES > ::= NP|ST|adj|num|noun|inf|ger$

$< QUA > ::= PP|adv$　　$PP ::= \{Pre, PRO\}$

$PRO ::= NP|NC$

$art ::= 'a'|'an'|'the'$　　$aux ::= 'do'|'have'|'be'|'can'|'will'|'should'$

注：{ } 为集合，< > 为句法成份，大写字符串表示短语范畴，小写字符串表示词法范畴，‘ ’为单词。

在语法分析器GA 的工作过程中涉及KB 中两类知识：词典(提供了单词的词性、词义及物性、助动性、人称等)，ATN 文法规则。由于ATN 的灵活性、文法描述与修改的方便性使得一些语义知识也能加入其中，这样，在利用ATN 文法进行文法分析时就增加了语义测试功能，方便有效地解决了语法分析中的歧义问题。GA 采用自顶向下的分析策略和深度优先的搜索方法。由于这种算法已广泛使用，这里不再赘述。

§5. 语义分析

要有效地从文章里提取出用户所需信息，文章的内部表示应能反映文章里各语句之间的联系、语句里各成分的语义及它们之间的内在逻辑关系。语义分析器SA 的任务就是在KB 中有关知识的支持下将语法分析之后产生的一系列语句的结构树转换为文章的一种深层结构— — 文章的意义框架amf。amf 定义如下：

$amf ::= (no\_csmf)$　　$no\_csmf ::= no : csmf|no : csmf; no\_csmf$

$no ::=$< 语句序号>

$csmf ::= (rel\_ssmf)$　　$rel\_ssmf ::= rel : ssmf|rel : ssmf; rel\_ssmf$

$ssmf ::= pred\ \ pmf$　　$rel ::=$< 复合句里各简单句的关系>

$pmf ::= (ca\_val)$　　$pred ::=$< 谓词语义参数>

$ca\_val ::= ca : val|ca : val; ca\_val$

$ca ::=$< 逻辑格关系> | < 语义范畴>

$val ::=$< 语义参数> $|pmf|ssmf$

注：csmf 为复合句意义框架；ssmf 为简单句意义框架；pmf 为短语意义框架。

语义分析以KB 中下列知识为背景：

1. 语义参数sep 和语义框架sef。在EAS 中词义用sep 和sef 描述之(sef 仅用于描述谓词的意义)。

$sef ::= sep : (ca_1 : properties_1; ca_2 : properties_2; \cdots; ca : properties_n)$

$ca_i ::=$< 逻辑格名>　　$properties_i ::=$< 语法约束>< 语义约束>

2. 语义范畴体系SC

$SC ::= \{SSC_1, SSC_2, \cdots, SSC_m\}$　$SSC_k ::= SC|$ < 语义参数>　$k = 1, 2 \cdots m$

3. 用于处理歧义问题的有关知识

SA 在语法分析之后得到的语句结构树ST 上工作。对于一ST，SA 先由KB 查得各词的sep 及谓词的sef，并根据语义范畴体系得到ST 中各短语的语义范畴，然后以sef 为模式

建立语句意义框架，按properties，递归地自上而下地搜索ST 的有关句法成分以确定出语句意义框架csmf 中各val，并确定出csmf 的pred(pred=sep)，从而建立了该ST 的csmf，最后在依次建立各ST 的csmf 的基础上再根据文章中语句的顺序、逻辑关系等建立起文章的意义框架amf。在此过程中的某些二义性问题往往借助意义框架里各成分之间语义制约关系或领域知识来引导解决。

§6. 信息提取

信息提取即按信息焦点集$F=\{f_1,f_2,\cdots,f_n\}$从文章的意义框架amf 里搜索、推理出有关信息，并进行归纳，产生出一个信息表il。il 定义如下:

$il ::= (inf\_items)$

$inf\_items ::= ic:imf|ic:imf;inf\_items$

$ic ::=<$ 信息范畴$>$ $imf ::=<$ 信息意义框架$>$

信息提取需借助KB 中下列知识:

1. 信息提取特征集iep, iep=$\{ief_1,ief_2,\cdots,ief_n\}$

$ief_i=ic_i:(ca_1:cond_1;ca_2:cond_2;\cdots;ca_m:cond_m)\quad i=1,2,\cdots n$

$ic_i ::=<$ 信息范畴$>$ $ca_j ::=<$ 逻辑格名$>$ $cond_j ::=<$ 语义约束$>$ $j=1,2,\cdots m$

2. 信息推理规则ir

$ir ::= p\rightarrow c$ $p ::= ief,sef\rightarrow op$

$ief ::=<$ 信息提取框架$>$ (同上) $sef ::=<$ 语义框架$>$ (同 §5)

$op ::=<$ 对意义框架进行修改的某种操作$>$

3. 精炼规则cr

$cr ::= p\rightarrow c$ $p=\{ic_{i1},ic_{i2},\cdots,ic_{in}\}$ $c=\{ic_{j1},ic_{j2},\cdots ic_{jm}\}$

这里$m<n$ 且$p\cap c$ 为空，$ic_{i1},\cdots,ic_{in},ic_{j1},\cdots,ic_{jm}$ 为信息范畴。

信息提取器IE 针对每个信息焦点f_i 依次进行信息搜索和推理。对于一个f_i，IE 先将其相应iep 中的一系列ief 与文章意义框架里的各子框架smf 进行匹配(即用ief 里各cond 检查smf 里的相应槽值)，如某smf_l 匹配上某ief_k，则ief_k 中的ic_k 就和smf_l 一道成为信息表il 中的一项。由于人类语言的字面意义之外还蕴含着大量信息，为提取这类信息IE 借助ir 对某些蕴含着与f_i 有关信息的smf 进行转换，推导出蕴含信息的显式意义框架smf，并和ir 中ief 里的ic 一道汇入信息表。在获得与各f_i 有关的信息后还需利用cr 对信息表里的信息进行整理和归并，最终得到一个较为简洁、与f_i 信息层次尽可能吻合的信息表il。

§7. 文摘的生成

文摘生成器的功能是由il 产生一篇用户易读的英文摘要。鉴于系统的目标是以文摘形式满足用户的信息需求，我们在文摘生成器AG 的设计中避开了自然语言生成中一些复杂的问题(如文摘的风格等)，着重于研究文摘的逻辑性和简明性。

AG 在工作过程中用及KB 中下列知识:

1. 信息范畴体系ic (见 §3)

2. 信息缺省说明nf, $nf ::= ic:sent$

$ic ::=<$ 信息范畴$>$ $sent ::=<$ 说明语句$>$

3. 文本生成规则gr (类似于一般文法规则，但其终结符不为单词，为逻辑格)

文摘生成分为两阶段：段章结构形成阶段和语句生成阶段。段章结构确定了il 里各信息项之间的关系，从而确定了各语句在文摘里的出现顺序，决定了文摘的逻辑性。这里采用段章结构树pst 描述文摘的段章结构。pst 定义如下：

$pst ::= (nodes)$　　$nodes ::= node|node, nodes$

$node ::= tnode|fnode$　　$tnode ::= ic : imf|ic : pst$

$fnode ::= ic : sent$　　$imf ::=$< 信息意义框架>

$ic ::=$< 信息范畴>　　$sent ::=$< 信息缺省说明语句>

在第一阶段中AG 以信息范畴体系的知识为背景，依次考查各焦点和il。对于一焦点f_i，首先建立起一个ic 值为f_i 的结点node$_i$。然后查il，如il 中存在与其相关的信息项($ic_j : imf_j$) $j = 1, 2, \cdots, m$，则据此建立起结点$node_1, node_2, \cdots, node_m$，并按$ic_j$ 之间的隶属关系rel_1 和平行关系rel_2 建立这些结点间的联系(如$ic_j\ rel_1\ ic_k$ 则node$_j$ 和node$_k$ 为父子关系；如$ic_j\ rel_2\ ic_k$ 则node$_j$ 和node$_k$ 为兄弟关系)，从而构造起一棵子树spst，作为node$_i$ 的子结点；如果il 中不存在与f_i 相关的信息项，则在node$_i$ 中放入该项信息的缺省说明。最后将各焦点对应的子树以兄弟关系连接起来就形成了文摘的pst。第二阶段AG 以深度优先的策略依次输入pst 上各项信息。如该信息结点有信息，就用gr 匹配结点中的imf，按匹配成功的句型依次输出imf 中的各成分；如该信息结点无信息，则输出结点的信息缺省说明。在此阶段AG 选择句型以简单明了为宗旨，在上下句的衔接上还考虑了代词的运用。

§8. EAAS 系统的实现及例示

EAAS 系统已在micro-vax 机上用C 语言实现，目前应用于一个就业机会介绍领域。系统可接受杂志Computer 中Career Opportunities 专栏的文章，按用户需求提取信息，以文摘形式输出。

下面给出两个小例子：

输入文章：

Texas A&M University

Applications are invited for faculty positions as Lecturer of Computer Science. Candidates from all areas of computer science will be considered. Experienced teachers who feel that their career would be invigorated by exposure to the computer science research community are invited to apply. Lecturers are not eligible for tenure.

The department of Computer Science has 15 full-time equivalent senior faculty members and additional teaching staff of 8 Lecturers. There is a tradition of quality instruction at the BS, MS and phD levels. The university has initiated a major commitment to develop excellence in Intelligent Systems, Artificial Intelligence, Software Technology, Simulation and other areas of computer science. This commitment is in support of the rapid build up of microelectronics and computer technology in Texas.

Computer Science at Texas A&M university is located in the College of Engineering. With nearly 10000 students, the College of Engineering is the nation's largest. Computer Science is expanding to become one of the larger and more comprehensive departments in the country. The university has the internal resources to achieve this goal.

Ability in teaching undergraduate is essential. An MS in Computer Science is required with a phD or comparable experience preferred. Applicants should submit a resume and three references to Bruce H. McCormick, Head, Computer Science Department, Texas A&M University,

Collge station, Texas 77843.

Texas A&M university is an equal opportunity, affirmative action employer.

例1:

What information do you want to know?

> I would like to know the information about education, speciality which the employee should possess.

> I want to get some information about the employer and pays.

ABSTRACT

Texas A&M University requires an MS in Computer Science with a phD or comparable experience preferred.

It will consider candidates from all areas of computer science.

Texas A&M university is an equal opportunity and affirmative action employer.

The salary which will be provied by employer does not be mentioned.

例2:

What information do you want to know?

> I want to know the position offered and requirement for the position.

> Please tell me the working environment.

ABSTRACT

Texas A&M University invites applications for faculty positions as Lecturer of Computer Science. Lecturers are not eligible for tenure.

Ability in teaching undergraduate is essential.

The department of Computer Science has 15 full-time equivalent senior faculty members and additional teaching staff of 8 Lecturers. The university has initiated a major commitment to develop excellence in Intelligent Systems, Artificial Intelligence, Software Technology, Simulation and other areas of computer science. There is a tradition of quality instruction at the BS, MS and phD levels. This commitment supports the rapid build of microelectronics and computer technology in Texas. Computer Science at Texas A&M university is located in the College of Engineering. The College of Engineering is the nation's largest with nearly 10000 students. Computer Science is expanding to become one of the larger and more comprehensive departments in the country.

注: 符号">"后为用户输入的信息要求

致谢: 在EAAS 系统的设计过程中我们得到了马希文教授的悉心指教, 在此向他表示深深的谢意。

参考文献

[1] Winograd, T, Language as Cognitive Process, vol, 1.

[2] De Jong, G. F, Skimming Stories in Read Time: A Experiment in Integrated Understanding, Yale University, Dept. Computer Science, Research Report 158 (May 1979).

[3] D. Roesner, Schemate for Understanding of Argumentation in Newspaper Texts, Universität Stuttgant.

[4] J. I. Tait, Generating Summaries Using a Script-based Language Analyser, Acorn Computers, Cambridge.