非受限域中文自动文摘系统的研究与实例

王开铸　李俊杰　吴　岩

哈尔滨工业大学计算机系，150001

摘　要：本文介绍了 HIT—863 1 型中文自动文摘系统的设计和实现，并通过大量事例阐述了它对不同体裁、领域、长短的文章，以不同的比例提取文摘的功能。

关键词：自动分词，自动文摘系统，无词典自动分词

Study and Instances for Unconstrained Automatic Abstracting System

Wang Kaizhu　Li Junjie　Wu Yan

Dept. of Computer Science　Harbin Institute of Technology　150001

Abstract　In this paper, the design and implementation of HIT-863I Chinese Automatic Abstracting System is introduced, and its abstracting for the text in arbitrary type, field and length with arbitrary abstracting rates is described by some instances.

1　引　言

自动文摘系统大致上可以分为两类：机械文摘[1-3]和理解文摘([4-6]。机械文摘系统是在词频统计和启发函数的基础上对非受限域文本进行文摘提取。理解文摘系统是基于语法、语义、脚本、CD 结构等知识表示，对受限域的语料提取文摘。HIT--863I 型自动文摘系统是在机械文摘技术的基础上，采用层次结构网络进行篇章表示，并在句子加权函数的设计和任意比例文摘提取算法等方面进行了大量的工作。技术鉴定表明，该系统在国内处于领先地位。

2　HIT—863I 型中文自动文摘系统的设计与实现

HIT--863I 型中文自动文摘是作者研制的非受限域中文自动文摘系统，它能对不同领域、体裁、长度的文章，以不同比例提取文摘。文摘采用原文中的完整句子，并且保持原文中句子的分段。因此，实际上是将原文以任意比例进行了压缩，且基本上做到了压缩后的文摘能较准确

地将原文中的重要句子优先提取出来。使文摘的质量得到了一定的保障。

自动文摘系统的总体结构，包括以下几个子系统：预处理子系统，文本接受子系统，无词典自动分词子系统，文摘生成子系统，后处理子系统。如图 1 所示。

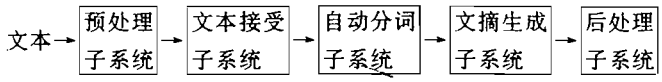

图 1　任意文本自动文摘系统的总体结构

各子系统功能如下：

(1)预处理子系统：将文章输入到计算机，并以文本文件的形式贮存在文件中，对于图表、公式等不可能进入文摘的内容删去。

(2)文本接受子系统：将预处理后的文本，以系统采用的层次结构网络的机内表示形式，转换贮存起来。

(3)无词典自动分词子系统：采用无词典自动分词系统，对文本的句子进行自动分词。

(4)文摘生成子系统：将分词后的文本，通过特征词提取，句子重要性动态测度，文摘提取等步骤，生成机器文摘。

(5)后处理子系统：将机器文摘进行冗余消除，分段，润色，并最后输出。

例 1　这是一篇研究生毕业论文的结束语，原文如下：

结　束　语

HOST 故障诊断专家系统采用了多层次的知识。既有深层知识(包括结构，功能和性能方面的知识)，又有浅层知识(现象与故障之间因果关系的知识)，此外，还采用了元知识，它负责启动相应的知识库(如自动诊断知识库和人工诊断知识库)。因此本系统基本上属于第二代专家系统。它明显优于基于浅知识的第一代专家系统。

因为本系统能够进行自动诊断，所以可以在此基础上，进一步发展实时诊断专家系统。首先总体结构可以大部分地得到继承；其次应该在通讯方面，采用较快速的并行通讯和共享内存等方式，并加强数据采集系统的预处理能力；第三采用多层次的知识组织，进行逐级诊断；第四采用中断机制，使重要的信息能够中断现行的诊断，使异常故障能够得到即时的发现。

在实时论断的基础上，可以很容易的实现实时控制，而且结构上基本一致，也是采用逐级控制的方式，各级之间既可以相对独立的工作，也可以以宏流水线的方式工作。

此外，在开发其它各种微机系统或机械系统的故障诊断专家系统时，本系统的框架可以基本保留，只是把知识库中的知识加以更换。因此本系统具有很强的通用性。

本系统的框架还可用于感知处理领域，如语音识别，图象处理，智能机器人和各种涉及到传感器信息处理的场合。它们的不同主要在于知识库中的知识和数据采集加工系统的软硬件结构和功能等。所以，本系统无疑为以上各类系统的设计和实现提供了一个可供参考的实例。

首先将它转换 HN 表示，每个字的使用情况用一个五元组表示(文章号，段号，句号，分句号，字号)，并将其存入对应于该字的外部文件中。另外，还生成一中间文件 Lsv。Lsv 的每一行是一个 n 元组(分句，分句的长度，文章号，段号，句号，分句号)。如下所示。

/ * //lsv * /

结束语 62111
HOST 故障诊断专家系统采用了多层次的知识 382211
既有深层知识 122221
包括结构 82222
功能和性能方面的知识 202223
⋮

第二步，将 Lsv 的各个分句进行分词，第一遍利用局部上下文信息进行分词，将文章中多次出现的专业词汇和特征切分出来。第二遍，利用背景语料信息，将文章中的常用词切分出来。对于实在切分不开的字段放在一起，并给一个权值 0。这里为了以后自动文摘提取时特征词加权的方便，将第一遍识别出来的词给一个正的权值，第二遍识别出来的词给一个负的权值，而对于英文和数量词给一个－1 的权值。而对于某些特殊字如“的，是，啊”等由于其组词能力有限，可以进行特殊处理，这类词的权值给一个固定值。经过上述几遍扫描分词的结果如下，形成一个中间文件 Lsg。

Lsg 中存放的是每一个分句的分词情况，其存放格式为

分句，分句长度，篇号，段号，句号，分句号，c_1 t_1 w_1 c_2 t_2 w_2…，0，0，0.0000

其中 c_1，t_1 表示抽出的开始字号和结束字号，w_i 表示其权值。由于采用递归算法，每次将最大权值的字符串优先抽取出来，因此，抽出的词不是按着原来句子中的顺序排列的。例如，对于第二句子“HOST 故障诊断专家系统……”，$c_1=1$，$t_1=4$，$w_1=-1.0$，它对应字符串“HOST”；$c_2=9$，$t_2=12$，$w_2=1280$，它对应字符串“专家系统”，等等。对英文和数词给予一个固定值－1；特殊字如“的”字给予权值 6；其它的字符串的权值计算由加权函数 $P=F\cdot L^C$。给出，其 $C=4$，$F\geqslant 3$，$L\geqslant 2$。

/ * lsg--middle file of segmentation * /

结束语 62111130.000000 0 0 0.000000 HOST 故障诊断专家系统采用了多层次的知识 38221114 －1.000000 9 12 1280. 000000 78160.000000 5680.000000 13 14 96.000000 15 15 0.000000 16 16 6 17 18 0.000000 19 19 0.000000 20 21 256.000000 0 0 0.000000 既有深层知识 12 2 2 2 1 5 6 256.00000 1 1 6 2 26 3 4 0.00000 0 0 0.000000 包括结构 8 2 2 2 2 3 4 80.000000 1 2 0.00000 0 0 0.000000 功能和性能方面的知识 20 2 2 2 3 1 2 0.00000 3 3 6 4 56 6 7 －64.000000 8 8 0 00000 9 10 256.00000 0 0 0.000000
⋮

第三步，进行句子加权，由于采用无词典自动分词系统，单词的加权实际上已在分词的过程中解决了。而对于使用其它自动分词系统，如各种机械匹配分词系统，还要进行单词的加权和特征词的提取。

句子加权后的中间结果以 n 元组（句子长度，文章号，段号，句号，句子权值）的形式存在一个中间文件 abs.sn 中

/ * abs.sn--middle file of sentences weighting. * /

6 2 1 1 0.000000 38 2 2 1 2.347500152 2 2 2 0.384756 34 2 2 3 5.556250 36 2 2 4 3.90000 70 2 2 5 1.544444 32 2 2 6 0. 758333 84 2 2 7 0.307609 36 2 2 8 0.863333 74 2 2 0.232143 2 2 10 0.157143 1080 2 2 11 1.204643 26 2 2 12 1.612500 92 2 2 13 0.341739 70 2 2 14 1.473611 64 2 2 15 0.784314 0 0 0 0 0.000000

第四步，根据用户提供的文摘比例 r，将 abs.sn 中权值最大的 K 个句子提出来，使 $|S_1|+S_2|+\cdots+|S_K|=T.r$，这里 $|S_i|$ 是句子 S_i 的长度，i=1，…，k，T 是原文的长度，r 是文摘的压缩比例。形成中间文件 a_1

/ * al--middle file of 34% abstract sentences setection &/

34 2 2 3 5.556250 36 2 2 4 3.900000 38 2 2 1 2.347500 26 2 2 12 1.612500 70 2 2 5 1.54444 70 2 2 14 1.473611 108 2 2 11 1.204643 0 0 0 0 0.00000

第五步，将 a_1 中的句子以原文中的顺序和分段情况，重新排列，并输出到文本文件 A 中，形成最后的机器文摘，用户可以对文件 A 进行阅读和查看。也可以进行人工润色。

/ * 34% abstract * /

HOST 故障诊断专家系统采用了多层次的知识。因此本系统基本上属于第二代专家系统。它明显优于基于浅知识的第一代专家系统。因为本系统能够进行自动诊断，所以可以在此基础上，进一步发展实时诊断专家系统。此外，在开发其它各种微机系统或机械系统的故障诊断专家系统时，本系统的框架可以基本保留，只是把知识库中的知识加以更换。因此本系统具有很强的通用性。它们的不同主要在于知识库中的知识和数据采集加工系统的软硬件结构和功能等。

3. 不同体裁、领域、长短的文章以不同的比例提取文摘举例

例 2　这是一篇毛主席写的“反对自由主义”的议论文，不同比例的摘要如下：

* 5% abstract * /

反对自由主义

自由主义者以抽象的教条看待马克思主义的原则。

我们要用马克思主义的积极精神，克服消极的自由主义。

/ * 12% abstract * /

反对自由主义

自由主义者以抽象的教条看待马克思主义的原则，他们赞成马克思主义，但是不准备实行之，或不准备完全实行之，不准备拿马克思主义代替自己的自由主义。这些人，马克思主义是有的，自由主义也是有的，说的是马克思主义，行的是自由主义。

自由主义是机会主义的一种表现，是和马克思主义根本冲突的。

我们要用马克思主义的积极精神，克服消极的自由主义。

从中可以看到以下几点：

(1)比例小的摘要内容被包含在比例较大的摘要之中。

(2)比例小的摘要中的句子是重要性最大的句子，随着比例的增大，次重要的句子逐渐被

收录进摘要，由此可以看出句子加权函数的准确性和有效性。

(3)对于用户来说，如果对某一比例中的局部内容不很清楚或比较感兴趣，则可以通过放大摘要比例，使该局部内容得到放大，使细节内容收录进去，直到用户满意为止。一般比例在30%左右的文摘已经足以令用户了解文章的主要内容和感兴趣的细节内容。同时对文章的写作风格和语言特色也已有了足够的了解。

(4)文摘的内容保持原文的分段和原句，因此，用户可以对摘要内容感到放心，一般不会窜改或者歪曲原文的内容。而且比例越大摘要的连贯性和一致性也就越好。

(5)任意比例的设定是由用户给出的，而摘要内容又是客观地抽取原文的内容，没有任何主观的修改，因此，用户会感到一种主动性和安全感，对摘要的内容的好坏和可接受程度给予一个正确的评价，不同的用户由于对文章的要求和理解不同，加之主观因素的影响，对摘要的评价可能会不尽相同。但本系统为每个用户提供了一个满足他们愿望和要求的手段，因此，会受到各种用户的欢迎。这比固定比例，固定摘要内容，以及按照固定的算法修改原文的内容的传统文摘系统要好得多了。

例 3　这是一篇中学课本上由马祖璋写的说明文“花儿为什么这样红”，由于篇幅有限下面仅给出 12%和 23%的摘要，原文省略。

/* 12% abstract */

花儿为什么这样红。

花儿为什么这样红。首先有它的物质基础。

花儿为什么这样红。还需要用物理学原理来解释。

花儿为什么这样红。还有它生理上的需要。

花儿为什么这样红。从进化的观点来考察，它有一个发展的过程。

花儿为什么这样红。从达尔文的自然选择学说来看，昆虫起到了重要的作用。亿万年前，裸子植物在地球上出现的时候，昆虫还不多。

花儿这样红，是大自然的杰作，更是人工培育成果。

/* 23% abstract */

花儿为什么这样红。

花儿为什么这样红。首先有它的物质基础。不论是红花还红叶，它们的细胞液里都含有由葡萄糖变成的花青素。当它是酸性的时候，呈现红色，酸性愈强，颜色愈红。

除了红花以外，还有黄色，橙色的花。

花儿为什么这样红，还需要用物理学原理来解释。太阳光经过三棱镜或水滴的折射，会分成红，橙，黄，绿，青，蓝，紫七种颜色。

花儿为什么这样红。还有它生理上的需要。光波长短不同，所含热量也不同，红，橙，黄光波长，含热量多。蓝、紫光波短，含热量少。

花儿为什么这样红。从进化的观点来考察，它有一个发展的过程。裸子植物的花是原始的形态，都带绿色，而花药和花粉则呈黄色。

花儿为什么这样红。从达尔文的自然选择学说来看，昆虫起到重要的作用。亿万年前，裸子植物在地球上出现的时候，昆虫还不多。花色素淡，传粉授精，依靠风力，全部是风媒花。

花儿为什么这样红。

花儿这样红，是大自然的杰作，更是人工培育的成果。

可以看出虽然摘要中具有一些冗余的句子，但原文的中心内容还是被准确地抓住了。

4 结　论

机械文摘系统所提出的摘要的评价标准，可以从(1)文摘的概括性；(2)文摘的可读性；(3)文摘的可理解性；(4)文摘的检索功能等方面进行，实验结果表明，本系统在这些方面具有较好的表现。

参 考 文 献

[1] H. P. Luhn, The Automatic Creation of Literature Abstracts, IBM Journal of Research and Development, 1955, NO. 2.

[2] A. Mathis, Techniques for the Evaluation and Improvement of Comptuter Produced Abstracts, Ohio State University, Dec. 1992, PB214675.

[3] H. P. Edmundson. New Methods in Automatic Abstracting, Journal of the Association for Computing Machinery, 1969, Vol. 21, NO. 6, PP. 264-285.

[4] G. Dejong, Prediction and Substantiation: Two Processes that Comprise Understanding. Proceedings of UCAI-79, Tokyo, 1979. 1.

[5] Danilo Fum. stal., Forward and Backward Reasoning in Automatic Abstracting, COLING82.

[6] Kenji Ono, Kazuo Sumita, Seiji Miike. Abstract Generation Based On Rhetorical Structure Extraction, COLING 94, Kyoto, Aug. 5-9, 1994, vol. 1, 344-348.