基于相似分类的文献理解及自动文摘系统研究*)

谈文蓉[1] 杨宪泽[1] 谈 进[2]

(西南民族大学计算机科学与技术学院 成都 610041)[1] (西南财经大学经济信息工程学院 成都 610065)[2]

摘 要 文献理解与相关处理技术,一直是自然语言理解研究领域的热点问题。本文在分析自然语言理解的概念与困难的基础上,提出了一种文献近似分类算法;讨论了汉语分词这一难点问题的研究思路及其在文献理解中的应用,给出了一种汉语文摘生成处理的方法。初步的测试表明,该方法具有较高的效率。

关键词 自然语言理解,近似分类,自动分词,汉语摘要

Study for Document Interpretation and Automatic Abstracting Based on Analogic Sorting

TAN Wen-Rong[1] YANG Xian-Ze[1] TAN jin[2]

(School of Computer Science and Technology, Southwest University for Nationalities, Chengdu 610041)[1]

(School of Economic Information Engineering, Southwest University of Finance and Economics, Chengdu 610065)[2]

Abstract Document interpretation and its relative processing technology are hotspots in the research domain of natural language understanding. In this paper, after analyzing the ideas and difficults of natural language, we propose an analogic sorting alogrithem for document interpretation. We discuss the ways of chinese word segmentaion and its application, which is the most difficult problem in chinese document interprtaion. Finally we develop a method of automatic abstracting, the elementary tests show that it has a high efficiency.

Keywords Natural language understanding, Analogic sorting, Automatic word segmentation, Chinese abstracting

随着电子文献的大量涌现和 Internet 网络的广泛应用,用户对海量文献进行智能化处理的要求越来越迫切。一般情况下,用户的专业素质决定了文献的利用效率,而专业素质的提高往往需要花费大量的人力、时间和费用对有关人员进行培训。即便如此,面对繁杂多样的各类文献,一些相关人员仍难以正确理解文献的含义,导致许多文献的使用价值得不到充分的利用。

一些传统软件的帮助系统大多采用应用分类检索的模式,但对于功能复杂的软件,用户使用这种模式来寻求帮助将十分烦琐,许多软件帮助系统难以有效发挥作用,无法满足用户的要求。若借助智能帮助系统,在系统构建时加入文献处理和智能语言理解技术,将极大提高文献的利用效率。用户使用智能问答系统时,只需直接在文本框中输入用自然语言描述的查询要求,系统即能在完成理解后,快速地查找出对应的答案,并提供与查询内容相关的一些知识。与传统的查询系统相比,智能问答系统的优势在于能够及时理解用户的问题并尽可能地为用户提供有关的文献,使众多的文献能够真正发挥帮助作用。

本文围绕智能问答系统中的核心问题,讨论了自然语言理解的一些关键技术[1~6]。针对为数众多的中文文献,提出了文献理解的近似分类算法。深入探讨并剖析了中文文献理解中的汉语分词问题,给出了一种基于关键词的汉语文摘生成方法,该方法能够实现文献的理解并快速地生成高质量的汉语文摘。

1 自然语言理解的概念与困难

要研究自然语言理解,首先必须对自然语言的构成有个基本认识。语言是音义结合的词汇和语法体系,是实现思维活动的物质形式。语言是一个符号体系,但与其他符号体系又有所区别。

语言是以词为基本单位的,词汇又受到语法的支配才可构成有意义的和可理解的句子,句子按一定的形式再构成篇章等。词汇又可分为词和熟语。熟语就是一些词的固定组合,如汉语中的成语。语法是语言的组织规律,语法规则制约着如何把词素构成词,词构成词组和句子,语言正是在这种严密的制约关系中构成的。

然而,对自然语言的理解却是一个十分艰难的任务。即使建立一个只能理解片言断语的计算机系统,也是很不容易的。这中间有大量的极为复杂的编码和解码问题。一个能够理解自然语言的计算机系统就像一个人那样需要上下文知识以及根据这些知识和信息进行推理的过程。自然语言不仅有语义、语法和语音问题,而且还存在模糊性等问题。

如果没有人工智能的参与,自然语言理解就无法实现。实现自然语言理解和自然语言生成困难的根本原因是自然语言文本和对话的各个层次上广泛存在的各种各样的歧义性或多义性。

一个中文文本从形式上看是由汉字(包括标点符号等)组成的一个字符串。由字可组成词,由词可组成词组,由词组可组成句子,进而由一些句子组成段、节、章、篇。无论在上述的各种层次:字(符)、词、词组、句子、段……,还是在下一层次向上一层次转变中都存在着歧义和多义现象,即形式上一样的一段字符串,在不同的场景或不同的语境下,可以理解成不同的词串、词组串等,并有不同的意义。一般情况下,它们中的大多数都可以根据相应的语境和场景的规定而得到解决。我们平时并不感到自然语言歧义。但是对于机器理解,为了消解歧义,需要极其大量的知识和进行推理。如何将这些知识较完整地加以收集和整理出来;又如何找到合适的形式,将它们存入计算机系统中去;以及如何有效地利用它们来消除歧义,都是工作量极大且十分困难的工作。

因此,自然语言的形式(字符串)与其意义之间是一种多

*)基金项目:四川省重点科技攻关项目(编号 05SG022-016),西南民族大学自然科学研究项目。**谈文蓉** 副教授,硕士,主要研究方向:自然语言处理,数据库。**杨宪泽** 教授,主要研究方向:自然语言处理,法与数据结构。**谈 进** 讲师,硕士,主要研究方向:数据库,计算机网络。

对多的关系。其实这也正是自然语言的魅力所在。但从计算机处理的角度看，必须消除歧义，这正是自然语言理解中的中心问题，即要把带有潜在歧义的自然语言输入转换成某种无歧义的计算机内部表示。

2 近似分类算法

N篇文献中的一些关键字可能同属于多门学科，但文献一般只认定属于一门学科，这样，这些文献具有不确定性，甚至属于交叉学科。因此，自动分类采用模糊法(近似法)较为合适。显然，分类恰当有利于使这些文献为各种需求服务。

2.1 算法构思

设N篇文献由 $X_1, X_2, \cdots, X_n$ 组成，n 篇文献中含有 m 个不同的关键字 $K_1, K_2, \cdots, K_m$，这样，一篇文献 x_r 可用 m 维向量来描述：

其中 $\Phi_{rj}=\begin{cases}1, Xr \text{ 中有关键字 } Kjr\\ 0, Xr \text{ 中无关键字 } Kjr\end{cases}$

对于每一关键字 $K_1, K_2, \cdots, K_m$ 事先标注了它是属于某一类或多类(某一门学科或多门学科)。例如，作者发表一篇文章，给出三个关键字，只认定这篇文章属于一门学科，那么三个关键字均属这门学科；然而，在另一篇被认定属于另一门学科的文献中，有一个关键字与前述文献相同，那么这一关键字将属于两类。

统计 K_j 在第 i 类中出现的概率

$P_{ij}=L_{ij}/L_j \quad (j=1,2,\cdots,m; i=1,2,\cdots,d)$

式中，L_j 是 K_j 在文献中出现的总次数；L_{ij} 是 K_j 认定属于 i 类的总次数，d 为分类数。

每一类 G_i 的模糊度量0～1之间的隶属函数可由下式算出：

$$\mu_i(x_r)=[\sum_{j=1}^{m}\phi_{rj}P_{ij}]/\sum_{j=1}^{m}\phi_{rj} \quad (i=1,2,\cdots,d; r=1,2,\cdots,n)$$

分类原则至少使每篇文献分到一类中去。这里采用扫描的方法，规定隶属函数 $\mu_i=0.9$，扫描一次，如果类集合中尚未包含全部 $X_1, X_2, \cdots, X_n$，降低隶属函数阀值使 $\mu_i=0.8$，再次扫描…，直至类集合中包含了 $X_1, X_2, \cdots, X_n$ 为止。

上述过程 P_{ij} 的确定是关键，但这并不难。因为我们实施的是近似分类，那么在一篇文献中出现 K_j 时，它可能被定义成某一类(如人工智能)；另一篇文献还可能出现 K_j，它可能被定义成另一类(如数据结构)，这样，K_j 共属这两类。只要 K_j 明确了，P_{ij} 就可以算出，其文献的近似分类同过上述过程就自动实现了。

2.2 算法描述

YJK1：确定分类数 d，每篇文献假定属于一类(由填表认定)。

YJK2：r 从1至 n，输入每篇文献 Xr 的关键字 $K_1, K_2, \cdots, K_m$(即建立索引)，有

(1) 若 K_j 所属文献为 i 类($i=1,2,\cdots,d$)，$L_{ij}=L_{ij}+1$ (L_{ij} 的初值赋0)。j 所属文献为 i 类($i=1,2,\cdots,d$)，$L_{ij}=L_{ij}+1$(L_{ij} 的初值赋0)。

(2) 若有 K_j 相同($j=1,2,\cdots,m$)，$L_j=L_j+1$(L_j 初值赋1)。

YJK3：r从1至 n，Xr 含有的关键字 K_j ($j=1,2,3,\cdots$) 对应 $\Phi_{rj}=1$(其余 Φ_{rj} 初值已赋0)。

YJK4：[初值 $j=1$] i 从1到 d，计算 $P_{ij}=L_{ij}/L_j$。

YJK5：$j\leftarrow j+1$，直至 $j=m$，重复 YJK4。

YJK6：[初值 $r=1$，$j=1$，$i=1$，$A=0$，$B=0$]，计算 $A\leftarrow A+\Phi_{rj}$，$B\leftarrow B+\Phi_{rj}P_{ij}$。

YJK7：$j\leftarrow j+1$，直至 $j=m$，重复 YJK6。

JK8：$\mu_{ir}=B/A$，$i\leftarrow i+1$，直至 $i=d$，重复 YJK6，YJK7。

YJK9：$r\leftarrow r+1$，直至 $r=n$，重复 YJK6—YJK8。

YJK10：[分类开始，每篇文献至少分到一类中，初值 $\mu=0.9$，$r=1$] i 从1到 d，若有 $\mu_{ri}\geqslant\mu$，$Xr\rightarrow G_i(j)$，$j\leftarrow j+1$(j 为类集合)，此时 $Pr=1$。

YJK11：$r\leftarrow r+1$，直至 $r=n$，重复 YJK10。

YJK12：r 从1至 n，若 Pr 均为1，分类结束；否则，$\mu\leftarrow\mu-0.1$，重复 YJK10，YJK11。

2.3 两点说明

(1)算法 YJK1—YJK9，是近似分类的一种计算机自动计算方法，关键在确定合适的隶属函数，这里只是一种探讨，也可以采用其它方法确定隶属函数。

(2) YJK10—YJK12 实现自动分类，最后文献被分到 $G_1, G_2, \cdots, G_d$ 中去。由于每篇文献至少分到一类中去，用 Pr 作标志，$Pr=1$，意味着 Xr 至少进入了一类。YJK12 首先判断是否有 $Pr=0$，若有，则还有文献还没有入类，这时降低阈值再分类，直到每篇文献至少进入一类为止。

3 自动分词

汉语自动分词是中文自然语言理解、自动翻译、电子词典等信息处理的基础性工件。所谓分词，就是把一句话，一篇文章甚至一部著作中的词逐个切分出来。汉语不象拼音文字那样有自然切分标志，且词长短不一，词的定义也不统一，语言学中对词的定义多种多样，造成切分的多样性，这自然给自动分词的同一性带来很大困难。汉语中词本身的词素、词、词组无明显的区分界限，没有一个统一的标准，易产生歧义，许多东西都是凭经验和语感来划分。这项工作如果全部交给计算机来作，就更复杂了。

尽管计算机自动分词存在许多困难，但由于自动分词是许多应用工作的第一步(也是汉语自动摘要的第一步)，促进了研究的持续不断，提出了不少方法，它们各有优缺点，也可能是基于特定环境的[6~8]。

3.1 正向最大匹配法和逆向最大匹配法

正向最大匹配法是最早提出的自动分词方法，其基本思想是先取一句话的前六个字查字库，若不是一个词，则删除六个字的最后一个字再查，这样一直查下去，至找到一个词为止。句子剩余部分重复此工作，直到把所有的词都分出为止。逆向最大匹配法也一样，不同的是它是从句子的最后六个字开始的，每次匹配不成功时去掉汉字串中最前面的一个字。

两法思路清晰，易于计算机实现，但由于试图用相对稳定的词表来代替灵活多变，充满活力的词汇，把词库搜索作为判词的唯一标准，因而具有很大的主观性和局限性。另外，这两种方法实际上否认了语言中的歧义现象。

3.2 自动分词逆向匹配法算法例

在应用中，方法有所变化。如下述算法我们初始不是取六个字而是取长度最短词的个数。

A1：一条汉语语句分划成单一字符 $X_1, X_2, \cdots, X$。

A2：决定语词中可能出现的词最大字符长度 L_{max}，最小字符长度 L_{min}。

A3：逆向匹配，取语句最后的 L_{min} 个字查词库，若查不到，加入一个字重复工作，直至字符数为 L_{max} 为止。

A4：若实施 A3 查不到词，去掉语句中最后一个字，再实施 A3，直至整个语句只剩下 L_{min} 为止。

3.3 歧义问题及初步处理

歧义在自动分词时就会出现，难点是歧义切分，而歧义切分主要可以分为四个方面：

(1) 词组的多义产生的歧义。

(2) 由自然语言的二义性产生的歧义。例如："在日本保留和尚使用的古典乐器很多"。这句若没有上下文辅助，连人也难理解其真实含义，计算机程序肯定出现两种分词情况 。

在/日本/保留/和/尚使用/的/古典/乐器/很多；

在/日本/保留/和尚/使用/的/古典/乐器/很多。

(3) 由计算机程序分词产生的歧义。这种情况虽然人可以正确分词，但计算机毕竟不是人，出现歧义难免。计算机程序分词产生的歧义一般有两种：组合型歧义。即，对于字串AB，可以分成AB，也可以分成A / B；交集型歧义。即，对于字串ABC，可以分成AB / C，也可以分成A / BC。

(4) 由词典大小产生的歧义。自动分词必须借助词典，若词典中没有的词，就不可能正确分词。

四种歧义中解决第二种最难，可以说目前还没有好的方法，好在统计表明这类歧义只占歧义总数的5%左右；第一种好的方式是分成专业词典；第三，四种怎样解决是研究的热点。我们给出一种初步研究的方法：

步骤1：待分词的句子用正向最大匹配法和逆向最大匹配法初步自动分词。

步骤2：比较两个分词结果，若结果一致，正确而分词结束；否则，继续步骤3。

步骤3：比较词数，若不等，选词数较少的一个作为分词结果；相等，继续步骤4。

步骤4：比较未登录词词数，若不等，选词数较少的一个作为分词结果；相等，继续步骤5。

步骤5：查找规则库，用规则进一步确定分词结果。

说明：步骤2中两个分词结果一致一般就是正确的；步骤3的情况是根据组成长词的情况可能性比例很高作为依据；步骤4的解释雷同于步骤3；步骤5的情况比较复杂，目前规则主要考虑具体词之间邻接的可能性、词类之间的邻接概率，还需要进一步研究。

4 汉语文摘生成处理的一个方法

4.1 假设

(1)汉语中常用词汇大都由2个、3个和4个汉字组成，5个和5个以上汉字组成的词汇使用频率较低，可以通过学习得到。因此，基本词库只包含2字、3字和4字词汇；

(2)汉语合式文本(去掉了控制字符)经与基本词库配合自动分词得到的词串中，人名、地名和实体名等最有可能是存在于下列词汇之间的词汇串：1)词长大于等于2的词汇；2)标点符号；3)单字停用词汇。

(3)经过人名、地名和实体名等识别后的字符串中，存在学习合成新词的可能性。这当中相互衔接的两个词汇合成新词的可能性最大。

(4) 如果词A在文章中出现的次数为N_1，词A与紧跟在它后面K个词B_1，B_2，…，B_k衔接出现的次数为N_2，且$N_2/N_1 \geqslant d$，则词$AB_1B_2\cdots B_k$为学习到的新词。具体实现时取$K=1$，如果词A在文章中出现的次数为N_1，词A与紧跟后继词B衔接出现的次数为N_2，且$N_2/N_1 \geqslant d$，则词AB为学习到的新词。

(5)汉语文章中，除去停用词后，词频与关键程度成正比。但按词频大小输出的前N个关键词不一定符合关键程度从大到小的顺序。

备注：停用词指汉语文章中频繁出现的一些助词、虚词和人称代词等(如：的、了、我、什么)。这些词虽然出现频率高，但它们并不是关键的词汇，可以把它们单独组成一个词库。

4.2 基于关键词汉语摘要思路

(1)简单摘要：以单个句子为最小摘要单位，统计一句当中总词数N_1和关键词数N_2，若$N_2/N_1 \geqslant d$，则在摘要中保留该句子。

(2)加权摘要：以单个句子为最小摘要单位，对词汇分类，新学习的关键词权值$W_1=10$，一般关键词权值$W_2=5$，非关键词权值$W_3=1$。统计一句中新学习的关键词数N_1，一般关键词数N_2，非关键词数N_3，若$(N_1*W_1+N_2*W_2)/(N_1*W_1+N_2*W_2+N_3*W_3)\geqslant d$。简单摘要是$W_1=W_2=W_2$的特例。

结束语 自然语言理解对于汉语来说，任务更为艰巨。在比较窄的应用领域或实用型系统中，从事汉语信息处理的研究者也曾构造了各种类型的计算语法模型。不过，可以看出这些模型几乎都是借用国外已有的语法理论(如上下文无关语法，扩充转移网络，语义语法，格语法，语义网络，广义短语结构语法，词汇功能语法，依存语法等等)解释一部分汉语的语法现象，许多问题还望有可突破性进展。

本文的近似分类算法能较成功地分类文献，该算法除了检索使用方便外，还可供有关部门参考各学科的差异、强弱。

关于自动分词，本文解决歧义的方法对交集型很有帮助，对其它的，没有满意的效果。这说明，汉语自动分词要在解决歧义方面有好的效果，还需作深入的研究，还有很长的路要走。

本文生成摘要的方法经过测试，5万汉字的文章在P4 2.2G的微机上运行，只要2分钟就可以生成文摘，当然这是与高效率的检索算法及分词方法的配合分不开的。这种摘要的方法特别适合网络应用，可以移植到互联网上，为网页快速预览服务。

参 考 文 献

1 Karov Y, et al. Similarity-based Word Sence Disambiguation [J]. Computational Linguistics, 1998, 24(1): 41～59

2 Bod R, Hay J, et al. Probabilistic Linguistics [M]. Cambridge, MA: MIT Press, 2003

3 Geoff W, Michael J P, Daniel B. Machine learning for user modeling [J]. UserModeling and User-Adapted Interaction, 2001, 11(2): 19～29

4 Hammond T. A domain description language for sketch recognition [M]. Cambridge, MIT Student Oxygen Workshop: MIT Press, 2002

5 谈文蓉，杨宪泽. MIS的智能处理的近似评判法及其算法研究[J]. 计算机科学，2005，32(3)：226～228

6 杨宪泽. 自然语言处理的句法分析和规则索引算法[J]. 科技通报，2002，18(6)：470～473

7 杨晓兰，钟义信. 基于文本理解的自动文摘系统研究与实现[J]. 电子学报，1998，26(7)：155～158

8 Matsumoto N T, Yuji. A New Approach to Unsupervised Text Summarization. In: Proc. of ACM SIGIR'01, 2001. 26～34

9 姚天顺. 自然语言理解——一种让机器懂得人类语言的研究(第二版)[M]. 南宁：广西科技出版社，2003