MIS 智能接口中汉语分词系统的设计与应用*)

谈文蓉[1] 杨宪泽[1] 谈 进[2] 刘 莉[1]
(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 提供汉语检索接口是MIS应用的一大趋势,其主要困难在于如何让计算机理解汉语检索用语,为此本文构建了MIS智能检索接口中的汉语分词系统,并提出了分词策略。对汉语切分中的歧义问题进行了深入的探讨,应用互信息和t-信息差完成了消歧算法的设计。实验表明,该系统具有较高的切分正确率与效率。

关键词 MIS,自动分词,切分歧义,交集型歧义,组合型歧义

The Design and Application of Chinese Word Segmentation System in a MIS Intelligent Interface

TAN Wen-Rong[1] YANG Xian-Ze[1] TAN Jin[2] LIU Li[1]
(College 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 It is a trend that MIS provide a Chinese indexing interface, its main difficulty is how to let the computer comprehend Chinese. In this paper we set up a word segmentation system of Chinese intelligent indexing interface and propose its word segmentation strategy. After disussing the ambiguity problems of Chinses word segmentation, we give a diambiguity alogrithem based on mutual information and difference t-test. The experimental results show that it has a high segmentation accuracy rate and efficiency.

Keywords MIS, Automatic word segmentation, Word segmentation ambiguity, Crossing ambiguity, Combination ambiguity

近年来,随着计算机在我国的普及与应用,MIS作为一种能进行信息收集、传递、存储、加工、维护和使用的软件系统被应用到了社会生活的方方面面。传统MIS的应用设计,常使用菜单或命令对数据库进行查询,导致MIS的检索功能较为单一,无法适应多层次用户及不同应用环境的需要,因此在MIS的设计中使用智能技术是必然的选择[1,2]。对用户素质参差不齐、应用频率较高的MIS而言,衡量其性能的最重要的指标是它与用户接口的友好性与实时性。我们从客观上要求一个好的MIS能接受灵活多样、内涵丰富的汉语句子。具体来讲,MIS应能提供汉语检索接口,能理解常用的检索用语,在检索功能方面具有求解复杂问题的能力,即具备汉语检索的智能接口。

为MIS设计汉语智能检索接口,要解决的首要的问题是让计算机理解汉语。如何让计算机理解汉语,一直是我国人工智能工作者不断研究的核心问题。它涉及到一系列语言学知识和计算机科学理论,难度极大。由于中文的输入都是以汉字为单位,词与词之间并没有像英语那样以空格分开,因此从计算机处理的角度来看,MIS中的汉语检索用语就是表示为汉字的字串,理解它们的第一步就是处理汉字串之间的关系,即进行汉语分词。汉语的词法约束很不规范且千变万化,给分词造成了很大的麻烦,长期以来汉语分词作为中文信息处理中最困难的问题之一,受到广大研究者的广泛关注。

针对MIS汉语检索用语的特点,本文探讨了分词词典的构建思路,提出了检索用汉语的分词策略;讨论了组合型歧义和交集型歧义的统计消歧方法,提出了一种应用互信息和t-信息差来消除检索用汉语切分歧义的方法,给出了消歧算法。初步的测试和实验表明,分词策略和消歧算法具有较高的切分正确率和排歧效率。

1 检索用语的歧义问题

1.1 汉语中的切分歧义

中文形式的检索用语没有类似英文空格之类的表示词的边界标志,其分词的任务就是要由分词软件在汉语检索用语的词与词之间自动地加上空格,再根据检索要求完成相关的处理。切分歧义是指汉语句子中的某些字段,如果纯粹根据词表作简单的字符串匹配,则它可能存在多种切分形式。切分歧义根据词与词之间的组合关系又可分为交集型歧义和组合型歧义两种类型。

定义1 在汉字字段AJB中,AJ是词并且JB也是词,则AJB为交集型歧义字段。其中A,J,B为字串。

定义2 在汉字字段AB中,A是词,B是词,AB仍是词,则AB为组合型歧义字段。其中A,B为字串。

定义3 由交集型字段或组合型字段的自身嵌套或交叉组合的汉字字段则为混合型歧义字段。

1.2 基于统计的消歧方法

统计模型具有鲁棒性和概括性,在含有错误的数据和新数据中性能优异,可以在分析自然文本的大规模系统中成功地消除歧义问题[3],因此基于统计学习的消歧方法正逐渐成为排歧的主流技术。基于概率统计的机器学习方法具有较好的实用性,可以避免规则方法的许多缺陷。它利用的知识主要是统计数据,这些数据可以从语料库中利用有指导或无指导的学习方法得到,从而避免了人工获取规则的繁琐过程。

*)基金项目:四川省重点科技攻关项目(05SG022-016),西南民族大学自然科学研究项目(05NY003)。谈文蓉 副教授,硕士,主要研究方向:自然语言处理,数据库。

一般情况下，基于统计的消歧方法从语料库(标注或未标注的)中统计支持不同歧义组合、不同词性或词义的上下文证据，这些证据可以用来对新的输入句子中的歧义进行消解，因此基于统计的方法经常使用词与词、词义与词义的搭配等上下文特征来消除歧义。

2 基础工作

2.1 语料的准备

统计消歧自动从语料库中学习词汇和结构偏向性信息，以此来探寻解决歧义的有关方法。统计消歧方法不仅需要海量的存储空间来存放语料，而且还经常需要从语料库中获取大量的统计信息。语料库统计分析结果的有效性以及是否能解决该类问题的典型样本等是决定统计消歧结果优劣的最主要的因素之一，为此我们收集了大量 MIS 检索用语，通过对它们的特点的分析和归纳，我们发现用于检索的 95%以上的汉语句子含有所属领域或专业的关键词，41%在切分时存在歧义现象。其中，绝大多数切分歧义是组合型歧义，一部分歧义为交集型歧义，极少数为混合型歧义，这些主要是由检索用语的固有的特点决定的。

切分歧义所需要的词频信息必须具有相当的规模，而若采用预经人工分词的语料作为训练样本，则其人工或机器分词的代价太大。因此，我们直接从未经加工的生语料库出发，通过字的统计信息模拟词频，并据此设计汉语检索句子的分词算法。我们对 5000 多句常用的检索用语建立了目标语料库供统计消歧使用，通过语料库获取汉字的统计信息，消除了人工切分可能引起的种种缺陷，保证了数据的准确性和一致性。

2.2 分词词库

针对 MIS 汉语检索用语中专业术语多、领域性强的特点，我们在为分词算法准备词库时有以下的构建思路：(1)建立两个分词用词库，一个为用于存放与具体学科专业相关的关键词词典，另一个为用于存放普通词汇的常用词词典。(2)根据各个领域的用词特性来构建关键词词库，词库按专业类别进行分类，专业代码由系统事先定义。检索时，MIS 首先提示用户输入目前需要检索的专业类别，智能接口系统将根据不同的专业类别选择不同的专业词库。

由于检索用语具有长词多、关键词多、平均词频大的特点，因此采用“先关键词，后普通词”的双向最大匹配分词策略，本身就能够很好地发现和消除组合型歧义，这样的分词策略符合 MIS 检索系统专业性、领域性较强的特点。同时，双向最大匹配的分词策略还能发现大部分的交集型歧义和组合型嵌套的混合型歧义字段，使用基于统计的消歧方法可以消除大部分此类歧义。

3 歧义字段的发现及处理

3.1 基于双向最大匹配法的分词策略

对用于检索的 MIS 汉语句子，采用“先关键词，后普通词”的分词策略。在对检索句子进行正向最大匹配扫描切分时，首先在关键词词典中查找相匹配的最长汉字字串，若找到，则进行切分后重新开始进行下一次的扫描匹配；否则，则要到普通词词典中去查找相匹配的最长字串，完成切分后重新开始下一次的扫描匹配。当正向匹配扫描至句尾时，又利用逆向最大匹配法进行回溯切分。如果正向匹配与逆向匹配的切分结果不一致，则可能存在交集型歧义和混合型歧义。MIS 检索用汉语句子的切分扫描过程如下：

(1)输入待切分的汉语检索句子 $S=A_1A_2\cdots A_n$，其中 A_i ($1\leqslant i\leqslant n$)为一个汉字或标点符号；

(2)进行正向最大匹配扫描切分(先关键词，后普通词)，得到正向切分结果 $Z=W_1W_2\cdots W_m$；令 $j=m$；

(3)进行逆向最大匹配扫描切分(先关键词，后普通词)，得到逆向切分结果 $L=C_1C_2\cdots C_t$；令；$k=t$；

(4)将 Z 的 j 个词与 L 的 k 个词进行逐词比较，如果出现不相同的切分，则根据词首字确定歧义字段的范围，将其加入切分候选结果集中，待下一步进行消歧处理。

3.2 相邻字间的互信息

所谓互信息是指对有序汉字串 xy，有 $I=(x:y)=p(x,y)/p(x)p(y)$，其中 $p(x,y)$ 为汉字串 xy 作为二字词出现的概率，$p(x)$，$p(y)$ 分别代表 x 和 y 可作为单字词独立的概率。如果 $p(x,y)=0$，表明 x 和 y 不成词；如果 $p(x,y)\neq 0$，$p(x)=0$ 或 $p(y)=0$，表明 x 和 y 至少有一个不能作为一字词单独出现，x 和 y 归并在一起的可能性较大；如果 $p(x,y)\neq 0$，$p(x)\neq 0$ 或 $p(y)\neq 0$，表明 x 和 y 既能作为一字词单独出现，也可归并在一起。

互信息能够反映出汉字对间结合关系的紧密程度，假设 xy 在训练用语料中作为二字词出现的次数为 $f(x,y)$，而 x 和 y 作为单字词独立出现的次数分别为 $f(x)$，$f(y)$，则可利用公式 $p(x,y)=f(x,y)/N$，$p(x)=f(x)/N$，$p(y)=f(y)/N$ 计算出互信息 $I(x:y)$。虽然互信息是两个汉字之间结合力的绝对度量，但在一些区域内，仅依靠互信息难以确定两个字是否应结合，还需通过上下文字对之间的比较进一步寻找依据。

3.3 相邻字间的 t-信息

对有序汉字串 xyz，$t_{x,z}(y)=(p(y|x)-p(z|y))\sqrt{P^2(y|x)-p^2(z|y)}$ 是汉字 y 相对于 x 及 z 的 t-信息，其中 $p(y|x)$ 表示当 x 可作为单字词出现的条件下，xy 作为二字词出现的条件概率；$p(z|y)$ 表示当 y 可作为单字词出现的条件下，yz 作为二字词出现的条件概率，有 $p(y|x)=p(x,y)|p(x)$，$P(z|y)=p(y,z)/p(y)$。t-信息能反映出字 y 与 x，字 y 与 z 结合的紧密程度，以此作为判断应该是 x 与 y 归并还是 y 与 z 归并。$t_{x,z}(y)>0$ 时，字 y 有与后继字 z 相连的趋势，值越大，相连趋势越强；$t_{x,z}(y)=0$ 时，不反映任何趋势；$t_{x,z}(y)<0$ 时，字 y 有与前趋字 x 相连的趋势，值越小，相连趋势越强。

3.4 相邻字间的 t-信息差

对有序汉字串 $vxyw$，$\Delta t(x:y)=t_{v,y}(x)-t_{x,w}(y)$ 是汉字 x 和 y 的 t-信息差，$t_{v,y}(x)>0$，$t_{x,w}(y)<0$，此时 x 和 y 之间相互吸引，必有 $\Delta t(x:y)>0$，x 和 y 之间倾向于连。$t_{v,y}(x)<0$，$t_{x,w}(y)>0$，此时 x 和 y 之间相互排斥，必有 $\Delta t(x:y)<0$，x 和 y 之间倾向于断。$t_{v,y}(x)>0$，$t_{x,w}(y)>0$，此时 y 吸引 x，同时 w 吸引 y，产生竞争：$\Delta t(x:y)>0$ 倾向于连，$\Delta t(x:y)<0$ 倾向于断。$t_{v,y}(x)<0$，$t_{x,w}(y)<0$，此时 x 吸引 y，同时 v 吸引 x，产生竞争：$\Delta t(x:y)>0$ 倾向于连，$\Delta t(x:y)<0$ 倾向于断。互信息反映了 x 和 y 之间的静态结合能力，t-信息差则动态考虑了 $vxyw$ 四个字的耦合影响，反映出字 x 和字 y 是相互吸引还是相互排斥。这两个参数具有一定的互补性，结合起来可以形成更趋合理的统计判断依据。

3.5 消歧算法

假设歧义字段有两个可能的断点 s1，s2，所在的位置表示为 b_mc_1 和 a_ib_1，α 和 β 是通过实验确定的阈值，则歧义字段的排歧算法流程如下：

(1)$i=0$；

(2)第 i 个字段已是候选结果集中的最后一个歧义字段

吗？是，转(9)；

(3)否：$i=i+1$；读入候选结果集中的第 i 个歧义字段；

(4)利用从训练语料中得到的数据对第 i 个歧义字段，计算每个二字应成词的 $I(x:y)$、$t_{x,z}(y)$和 $\Delta t(x:y)$；

(5)$I(b_m:c_1)>I(a_i:b_1)$且 $\Delta t(b_m:c_1)>\Delta t(a_i:b_1)$或 $I(b_m:c_1)<I(a_i:b_1)$且 $\Delta t(b_m:c_1)<\Delta t(a_i:b_1)$吗？是，$I$ 和 Δt 的判断一致，按的判断结果切分，转(2)；

(6)否：$|I(b_m:c_1)-I(a_i:b_1)|\geqslant\alpha$ 吗？，是则由 I 的判断结果切分，，转(2)；

(7)$|\Delta t(b_m:c_1)-t(a_i:b_1)|\geqslant\beta$ 吗？是则由 Δt 的判断结果切分，转(2)；

(8)否：由 I 的判断结果切分，转(2)；

(9)算法结束。

4 实验与讨论

歧义字段切分查准率和查全率是评价切分歧义消除效果的两个重要指标，定义如下：

查准率＝

$$\frac{\text{正确切分的歧义字段的个数}}{\text{正确切分的歧义字段个数＋错误切分的歧义字段的个数}}\times 100\%$$

查全率＝

$$\frac{\text{正确切分的歧义字段的个数}}{\text{正确切分的歧义字段个数＋未发现的歧义字段的个数}}\times 100\%$$

用本文的方法对 200 句 MIS 检索用语进行了测试，从分词结果来看，歧义切分查准率达到 84%，查全率为 77%。说明应用互信息和 t-信息差这两个统计量对检索用语的歧义字段进行切分，提高了分词精度，本文的分词策略和歧义消除算法是可行的。

结束语 切分歧义问题是中文分词处理中的一个难点，本文提出了 MIS 汉语检索接口中分词系统的构建方法，给出了对应的分词策略。应用互信息和 t-信息差等统计消歧方法设计了切分歧义的消除算法并对算法进行了测试，取得了较好的分词效果。分词策略与消歧算法适用于 MIS 智能检索等专业性和领域性较强的应用环境。受词典及训练语料规模及领域的限制，分词系统在测试时发现了部分切分错误，说明本文方法对部分交集型歧义的发现还不够完善。同时系统需要进行大量的数据计算，时间开销较大，消歧效率还有待提高。下一步将进一步扩大训练语料的规模和覆盖范围，降低消歧算法的计算复杂度，提高分词效率。

参 考 文 献

1 Michael R G, Nils J N. Logical Foundation of Artificial Intelligence. Morgan Kaufmken Publishers, Inc, 1987
2 Nguyend T, Vindrow B. Nearal networks for Self-Learning Control Systems. IEEE CSM 1990, 10(3): 18～23
3 Christopher D. Manning, Hinrich Schutze. 统计自然语言处理基础[M]. 苑春法，等译. 北京：电子工业出版社，2005. 143～163
4 孙茂松，肖明，邹嘉彦. 基于无指导学习策略的无词表条件下的汉语自动分词[J]. 计算机学报，2004，27(6)：736～742
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 曹娟，周经野. 一种计算汉字串之间相关程度的新方法[J]. 中文信息学报，2005，18(4)：55～59
7 孙茂松，黄昌宁，等. 利用汉字二元语法关系解决汉语自动分词中的交集型歧义[J]. 计算机研究与发展，1997，34(5)：332～339
8 杨宪泽，谈文蓉，唐向阳，等. 一种混合式机器翻译方法及其算法[J]. 计算机应用与软件，2005，22(9)：142～146

(上接第 146 页)

4.3 扰乱与扩散性能分析

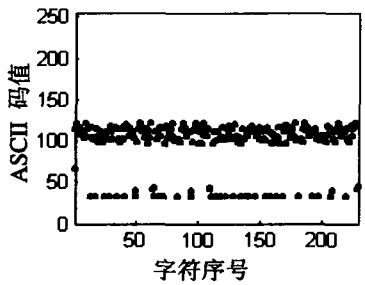

图 6 明文

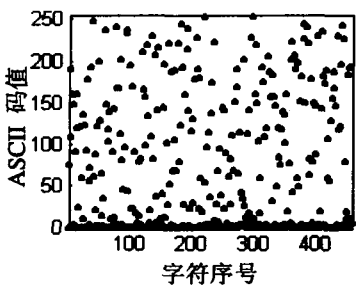

图 7 密文

扩散是将每一位明文的影响尽可能地作用到较多的输出密文位中去，同时，还要尽量使得每一位密钥的影响也尽可能迅速地扩展到较多的密文位中去。其目的是有效隐藏明文的统计特性，这也就是混沌系统的初始条件敏感依赖性。扰乱，是指密文和明文之间的统计特性的关系尽可能地复杂化，这也就是混沌映射通过迭代，将初始域扩散到整个相空间。为更清晰地描述这一特性，我们使用二维图形来表达。在图 6 和图 7 中，横轴代表信息中字符出现的序号，纵轴代表对应字符的 ASCII 码值(范围 0～255)。从明文和密文的图形来看，明文的码值比较集中，而根据本文算法所得到的密文在整个密文空间的分布都非常均匀。也就是说，通过扩散、扰乱等作用后，密文中不包含明文的任何信息(包括明文的统计概率信息)。这正是我们想要达到的加密效果。

结论 本文在分析三阶细胞神经网络和 Chebyshev 映射混沌特性的基础上，提出了基于神经网络和混沌映射的序列密码算法，给出了系统实现原理和算法描述，对混沌序列进行模拟实验和计算机仿真，同时对该系统产生的安全性能进行了分析。结果表明，这种方法设计的序列密码具有随机性好，在较低精度下序列的相关性能好，这大大降低了实现了成本。

参 考 文 献

1 van Schyndel R G, Tirkel A Z, Svalbe I D. Key independent watermark detection. IEEE International Conference on Multimedia Computing and Systems, Florence, Italy, 199: 580～585
2 Eggers J J, Su J K, Girod B. Public key watermarking by eigenvectors of linear transforms. European Signal Processing Conference, Tampere, Finland, 2000・428～435
3 丘水生，陈艳峰，吴敏，等. 混沌加密的若干问题与新的加密系统方案. 见：2002 中国非线性电路与系统学术会议论文集，中国，深圳，2002. 174～179
4 王育民. 混沌密码序列使用化问题. 西安电子科技大学学报，1997，24(4)：560～562
5 Chua L O, Roska T. The CNN paradigm. IEEE Trans. CAS-I, 1993, 40: 47～156
6 Civalleri P P, Gilli M. On dynamic behaviour of two-cell cellular neural networks. Int. J. Circ. Th. Appl., 1993, 21: 451～471
7 何振亚，张毅锋，卢宏涛. 细胞神经网络动态特性及其在保密. 通信中的应用. 通信学报，1999，20(3)：59～67
8 Tohur K, Akio T. Pseudonoise sequence by chaotic nonlinear and their correlation properties. IEICE Trans commun, 1993, E97-B (8): 855～862