●熊回香

全文检索中的汉语自动分词及其歧义处理*

摘　要　歧义处理是汉语自动分词的核心问题,汉语自动分词是中文信息检索的基础性课题。目前有基于词典的分词方法、基于统计的分词方法、基于语义的分词方法和基于人工智能的分词方法。自动分词的歧义处理,目前主要有:利用"长词优先"排歧,利用特征词消歧,利用"互信息"和"t-信息差"消歧,利用专家系统分词消歧。参考文献 15。

关键词　全文检索　汉语自动分词　歧义处理　t-信息差　专家系统

分类号　TP391

ABSTRACT　Processing of ambiguities is the key problem in the automatic segmentation of Chinese words, which is a basic issue in Chinese information retrieval. There are segmentation methods respectively based on dictionaries, statistics, semantics and artificial intelligence. In this paper, the author also introduces various kinds of methods for the processing of ambiguities. 15 refs.

KEY WORDS　Full-text search. Automatic segmentation of Chinese words. Porcessing of ambiguities. Expert system.

CLASS NUMBER　TP391

所谓全文检索,就是以各类数据诸如文本、声音、图像等为主要处理对象,根据数据资料的内容,而不是外在特征来实现的信息检索手段,其核心技术是将源文档中所有基本元素的出现信息记录到索引库中。由于中文的基本元素可以是单个汉字,也可以是词,因此,存在两种基本的索引结构,即基于字表的索引和基于词表的索引。基于词表的索引结构适于大规模应用,索引库较小,检索速度比较快,而且还可以实现同义词、反义词的概念检索[1],因而在全文检索系统中得到了广泛应用。而汉语自动分词技术无论是对于词表索引库的建立、全文检索的查全率和查准率,还是自然语言查询接口等都起着至关重要的作用。

1　汉语自动分词技术

词是最小的能够独立活动的有意义的语言成分,由于计算机内部存储的中文信息没有明显的词与词之间的分隔符,因此理解汉语的首要任务就是把连续的汉字串分割成词的序列,即自动分词。迄今为止,学者们已提出了多种分词方法,主要有:基于词典的分词方法、基于统计的分词方法、基于语义的分词方法和基于人工智能的分词方法。

1.1　基于词典的分词方法

该方法的基本思想是基于字符串匹配的机械分词:首先构建词典,词典中要尽可能包含可能出现的所有词。然后按照一定的策略将待分析的汉字串与词典中的词条进行匹配,若在词典中找到某个汉字串,则匹配成功(识别出一个词)。其匹配方法根据方向不同、字串长度优先次序不同,分为正向最大匹配、逆向最大匹配、双向匹配、逐词匹配、最少切分、全切分等匹配方法[2]。由于自然语言的复杂性,该类分词算法的最大弊端是无法避免分词歧义(即对同一段汉字串分词会出现不同的划分结果),因为机器词典不能提供可供进一步辨别切分结果的语法、语义知识,因此产生歧义。另外,对于词表中未能及时收录的新词,此类分词法无法予以正确切分。但这种方式回避了许多难度较大的语言自身信息的处理,且分词算法成熟,易于实现,是目前普遍使用的切分方法。

1.2　基于统计的分词方法

由于汉语词的定义的模糊性,有些学者利用统计方法,通过对大规模真实文本的统计,让计算机自己判断什么是词,这样就产生了基于统计的分词方法,又称为无词典分词。这类方法分词的依据和主要思想是:词是稳定的字的组合,因此在上下文中,相邻的字同时出现的次数越多,就越有可能构成一个词。因此字与字相邻共现的频率或概率能够较好地反映成词的可信度。也可以对语料库中相邻出现的各个字的组合的频度进行统计,计算它们的互现信息[3]。互现信息体现了汉字之间结合关系的紧密程度,当紧密程度高于某一个阈值时,便可认为此组字可能构成了一个词。

基于统计的分词方法优点在于,能够有效地自动排除歧义,能够识别新词、怪词,解决了基于字典的分词方法的弊病。但这种方法完全抛弃了汉语的词法、语法、语义信息,而只根据统计的结果来分词,因而具有一定的局限性,如会经常抽出一些共现频度高但并不是词的常用字组,并且对常用词的识别精度差,时空开销大。

1.3　基于语义的分词方法

* 本文是国家社会科学基金项目"基于中文 XML 文档的全文检索研究"(编号:04CTQ005)的研究成果。

该方法的基本思想就是：在语法分析的基础上，建立一个词库，其中包括所有可能出现的词和它们的各种语义信息，对给定的待分词的汉语句子，按照某种确定的原则取其子串，若该子串与词中的某词条相匹配，则从词库中取出该词的所有语义信息，然后调用语义程序进行语义分析，若分析结果正确，则该子串是词[4]。该方法的典型代表有：扩充转移网络法、知识分词语义分析法、邻接约束法、综合匹配法、后缀分词法、特征词库法、约束矩阵法、语法分析法等[5]。这类分词方法因为在分词过程中加入语法语义等信息，因而具有良好的歧义切分能力，但另一方面要对语言自身信息进行更多的处理，加大了实现的难度。

1.4 基于人工智能的分词方法

专家系统和神经网络是当前人工智能研究的两个热点，将两者应用到汉语自动分词中来，于是产生了专家系统分词法和神经网络分词法。

专家系统是一包含实例与规则的计算机程序，通常针对某一特定的范围可以以专家认定高水准的能力来协助解决问题。它是从模拟人脑的功能出发，构造推理网络[6]，将分词过程看做是知识推理的过程。专家系统具有显式的知识表达形式，知识容易维护，能对推理行为进行解释，并可利用深层知识来切分歧义字段。它的缺点是不能从经验中学习，且进行多歧义字段切分耗时较长，同时在知识表示、知识获取和知识验证等方面还有待进一步探索。

神经网络分词法旨在模拟人脑神经系统机构的运作机制来实现分词功能。它是以非线性并行处理为主流的一种非逻辑的信息处理方式，将分词知识分散、隐式地存入神经网络内部，通过自学习和训练修改内部权值，以达到正确的分词效果[7]。神经网络有很多优点，如联想、容错、记忆、自适应、自学习和处理复杂多模式等；但另一方面，神经网络的网络连接模型表达复杂，训练过程较长，不能对自身的推理方法进行解释，对未在训练样本中出现过的新词汇不能给予正确切分。

衡量一个自动分词系统的指标主要有3个：切分速度、切分精度、系统的可维护性。切分精度直接反映系统的正确性与科学性，是3个指标中最重要的一个。而要提高自动分词的切分精度，就必须有效地处理歧义字段。

2 切分歧义

由于汉字书写规则不同于西文，加上汉语词理解的多义性、复杂性，因而歧义字段在中文文本中是普遍存在的，歧义切分是自动分词过程中不可避免的现象，也是自动分词中一个比较棘手的问题。

分词过程中歧义产生的根源可归结为3类[8]。

(1)由自然语言的二义性所引起的歧义，称为第一类歧义。如："乒乓球拍卖完了"可切分为"乒乓球/拍卖/完了"，又可以切分为"乒乓球拍/卖/完了"。这两种切分形式无论在语法上还是语义上都是正确的，就是人工分词也会产生歧义，只有结合上下文才能给出正确的切分。

(2)由机器自动分词产生的特有歧义，称为第二类歧义。如："在这种环境下工作是太可怕了"用机器切分，可以切分为"在/这种/环境/下工/作/是/太/可怕了"，也可以切分为"在/这种/环境/下/工作/是/太/可怕/了"。对本句来说，只有第二种切分是正确的，用人工分词是不可能产生歧义的，歧义是由于机器机械切分产生的。

(3)由于分词词典的大小而引起的歧义，称为第三类歧义。如："王小二是一个农民"用机器切分被分为"王/小/二/是/一个/农民"，这里"王小二"是一个人名，在汉语中应是一个词，所以这个切分是错误的。由于机器自动切分是依据分词词典进行的，如果词典中没有这个词，此句就可能不被正确地切分。但分词词典不可能包括所有的词(如人名、地名)，另一方面，词典中所包括的词越多，也会产生新的歧义。例如"发展社会主义的新乡村"，新乡是一个地名，若词典中有该词，则"新乡村"，是一个歧义字段。因此，不论词典的大与小都可能产生歧义。

3 歧义字段的类型

根据歧义字段的构成形式，歧义字段又分为以下两类：交叉型歧义字段与组合型歧义字段。

交叉型歧义字段是指，如果一个汉字串包含A、B、C 3个子串，若AB、BC分别构成词，则该汉字串有两种切分形式：AB/C和A/BC。例如："这/糖/果真/好吃"与"这/糖果/真/好吃"都是符合切分规则的结果。交叉型歧义字段是由词与词之间的交叉组合产生的，占歧义字段的绝大多数，据统计达94%[9]。

组合型歧义字段是指，对于汉字串AB，既可以切分为AB，又可以切分为A/B。比如说"要紧的"，在句子"不要紧的，我想不会有事的"中，应切分为"要紧的"，而在句子"不要紧的，要松的"中，则应切分为"要/紧的"。组合型歧义字段，表现了汉语词可造词的现象，这类歧义字段为数不多。

4 歧义字段的处理

歧义性是汉语自动分词中不可避免的现象，排歧就成为自动分词中的一个核心问题。不同类型的歧义，产生的根源各不相同。应针对不同的歧义类型采取不同的解决方法。对于第一类歧义，由于它们本身就是汉语言中的歧义问题，解决这类歧义需要依靠上下文语义信息，即增加语义、语用知识的处理。处理第三类歧义目前主要有两种方法：一是增加构词知识，扩大词典；二是增加临时词典。还可以人工干预分词或人工分词与计算机自动分词结合，在遇到计算机解决不了的歧义时，借助于人工干预来完成。第二类歧义占歧义字段总数中的绝大多数，也是歧义处理中的难点，因而，本文重点讨论第二类歧义的处理方法。

4.1 利用"长词优先"排歧

目前,"长词优先"准则是解决组合型歧义的最常用也是最实际有效的一种切分准则。所谓"长词优先",就是尽可能地用最长的词来匹配句子中的汉字串。其基本思想是:假定分词词典中的最长词有i个汉字,则用被处理文档的当前字串中的前i个字作为匹配字段查找字典。若字典中存在这样的一个i字词,则匹配成功,匹配字段被作为一个词切分出来。如果词典中找不到这样的一个i字词,则匹配失败,将匹配字段中的最后一个字去掉,对剩下的字串重新进行匹配处理……如此进行下去,直到切分出一个词或剩余字串的长度为零为止[10]。

比如说"中国人民"、"中国"、"人民"、"国人"、"中国人"和"中国人民"都是词。当我们在句子中遇到"中国人民"这个汉字串时,就会用"中国人民"这个词来匹配它,使得切出来的词尽可能长,切出来的词条数尽可能少。"长词优先"准则在一定程度上模拟了人工分词的心理过程。对于绝大多数组合型歧义,"长词优先"准则是适用的。

4.2 利用特征词消歧

此方法是先建立包含各种具有切分特征词的词库,分词时先根据特征词库将待分汉字串分成较小的子串,再对各个子串使用机械匹配法切分。

根据特征词在句子中的不同语法作用,可将其分为结构性特征词和非结构性特征词。

结构性特征词是指在句子中能够标示句子整体或部分结构的词,主要分为:作谓语或谓语中心词的动词和形容词,其标示的主谓结构或动词结构特征非常明显;介词短语中的介词(单字介词优先);联合结构中的连词;"的"字短语中的"的"。

非结构性特征词没有明显结构特征,但在构词或构形方面有明显规律。这类特征词的处理规则包括固定切分类、固定结构、前后缀、重叠形式等。其中固定切分类特征词是指在给定条件下固定采用某种切分形式的一类词;固定结构包括成语、简略语和习用语等[11]。

利用特征词消歧就是在切分时要遵循特征词优先的原则。如:"文章/用/数字详细/列出/了人类对地球生态的影响"。进行切分处理时,首先匹配出动词"用"和"列出"后,这个句子的结构就很清楚了。

4.3 利用"互信息"和"t-信息差"消歧

在分词过程,两个汉字串应该是相结合组成一个词,还是应该被分开,这是与它们之间的结合程度有关的。"互信息"和"t-信息差"反映了汉字串之间的结合强度。

(1)互信息。

互信息是信息论中的一个概念。在自然语言处理中,它体现了词串两个部分的信息相关程度。对汉字串xy,汉字x,y之间的互信息(或称汉字x,y间位置的互信息)定义为[12]:

$$I(x,y) = \log_2 \frac{p(x,y)}{p(x)p(y)}$$

其中,p(x,y)是x,y的邻接同现率,p(x),p(y)是x和y各自的概率。

互信息可用于定量估计两个汉字间的结合力:其间互信息越大(超过某一阈值),两个汉字结合的紧密程度越高,则x与y相连;互信息越小(低于某一阈值),结合的紧密程度越低,则x与y断开。

汉字串x和y的结合程度不仅与串本身的概率有关,还受到它们所处的上下文影响。也就是说它们的结合程度还要受到x与其前面的汉字串的结合强度以及y与其后面汉字串的结合强度的影响。这种影响可用"t-信息"来表示。

(2)t-信息。

对汉字串xyz,汉字y相对于x及z的t-信息定义为[13]:

$$t_{z,z}(y) = \frac{p(z|y) - p(y|x)}{\sqrt{\delta(p(z|y)) + \delta(p(y|x))}}$$

其中,p(y|x),p(z|y)分别是y关于x,z关于y的条件概率,$\delta(p(y|x))$,$\delta(p(z|y))$是各自的方差。

如果$t_{x,z}(y) > 0$,则汉字串y倾向于与其后继汉字串z相结合(亦即倾向于与其前趋汉字串x断),且值越大,倾向性越强;如果$t_{x,z}(y) < 0$,则汉字串y倾向于与其前趋汉字串x相结合(亦即倾向于与其后继汉字串z断),且绝对值越大,倾向性越强;如果$t_{x,z}(y) = 0$,则无任何倾向。

(3)t-信息差。

在汉字串vxyz中汉字串x和汉字串y的t-信息差d(x,y)定义为:

$$d(x,y) = t_{v,y}(x) - t_{x,z}(y)$$

若d(x,y)>M,则汉字串x和汉字串y结合;若d(x,y)<M,则汉字串x和汉字串y分开。其中M为阈值,可通过多次实验得到最佳值[14]。

在计算x和y之间的t-信息差时充分考虑v对x的作用,以及z对y的作用,即t-信息差的计算时充分考虑了上下文的信息(动态地考虑vxyz四个字的耦合影响),这样就可以有效地消除歧义,提高分词精确度。

4.4 利用专家系统分词消歧

专家系统的自动分词是基于知识库的逻辑推理过程,所以知识库的组织、推理机制是专家自动分词系统的核心问题,知识库是使专家系统具有"智能"的关键性部件,推理机制规定了由事实出发推出目标的方法过程,分词过程是按照推理机制实现知识推理的具体操作过程[15]。

(1)知识库。

知识库又分为常识性知识库和启发性知识库。

常识性知识库中主要包括3个方面:现代汉语词汇体系中的词类知识;为可能产生歧义切分的歧义词所加的歧义标志及歧义类型编号;为消除语义歧义切分字段所需的部分语义知识。采用的知识表示方式是"语义网络",网络中的每

个节点表示一个基本元素，每个叶节点均描述与分词有关的一条知识。用这种方式组织词法知识库具有结构清晰，便于推理机调用、便于发现歧义字段，甚至可能发现部分新词等优点。

启发性知识则是指消除歧义切分所需要的，从实际经验中总结出来的词法、句法知识，甚至包括部分语义知识。启发性分词知识用产生式规则表示，这类规则专用于解决各类歧义字段的切分问题。关于某一词条（相当于常识性分词知识语义网络中的一个节点）的歧义切分规则，由于受到推理过程所产生的中间假设词语树中位置的约束而分为3种不同的形式：消除某一节点的歧义知识与其未分词的直接前趋（后继）节点有关；消除某一节点的歧义知识与其已分词的（直接）前趋节点有关；消除某一节点的歧义知识与其已分词的（直接）后继节点有关。启发性分词知识即按照这3种不同形式将词法切分规则、句法歧义切分规则以及语义歧义切分规则组成统一的歧义切分规则库。

（2）自动分词过程。

自动分词专家系统是基于知识的逻辑推理过程，知识推理过程就是分词过程。推理机根据原始数据，以知识为动力，逐步形成一系列的中间假设和决策（即一棵棵词语二叉树）。每次推理得到的一棵二叉树，都作为下一次推理的初始数据，随着推理的深入，词语二叉树越来越接近目标，当词语二叉树的每个节点都是推理机可识别的事实时，推理成功，推理机自动停止。

推理机在每进行一步推理的过程中，既启动常识性知识库又启动启发性知识库。运行推理机的具体步骤是：推理机把待分词或已分词的字符串视为词语树中的节点，利用常识性知识库进行顺向搜索匹配。若匹配成功，则该词把原字符串断为左右两段，以该词作为子树的根，左边一段为子树的左孩子，右边一段为子树的右孩子，来代替原字符串在词语树中的节点，形成一棵新的词语树。一旦子树的根节点（假定为A），满足条件 $A\rightarrow flag<>0$ 且 $A\rightarrow abg<>0$（前者表示该字符串是词，后者表示消除该词条歧义特征的规则号），则推理机根据 $A\rightarrow abg$ 的值（该字符串的歧义特征）启发相应的歧义切分规则，校正刚刚形成的这棵词语树，从而达到消除歧义的目的。

另外还有许多不同的歧义处理方法，如：基于WSD的向量空间的消歧法、基于SVM和k－NN结合的消歧法、利用语境上下文的松弛迭代算法消歧、利用“联想－回溯法”消歧等等。不同的歧义处理算法，有着各自不同的特点，“长词优先”准则对消除组合型歧义既简单又实用，但却无法检测到所有的交叉型歧义；利用特征词消歧，在一定程度上可以提高切分精度，但此方法必须建立具有大量特征词的特征词库；利用“互信息”和“t－信息差”消歧，不需要人工干预，能识别新词、怪词，且速度快，但其切分精度还有待于进一步提高。自动分词专家系统可充分利用词法知识、句法知识、语义知识和语用知识进行逻辑推理，实现对歧义字段的有效切分，切分精度是所有消歧方法中最高的，因而，自动分词专家系统是未来的发展方向。

5 结束语

汉语自动分词是全文检索中的“瓶颈”问题，而切分歧义的处理又是汉语自动分词中一个最困难也是最核心的问题。全文检索系统检索效率的提高，依赖于汉语自动分词技术的发展，依赖于对汉语的分词结构、句法结构、语义等语言知识的深入系统的研究，依赖于对语言与思维的本质的揭示；同时，在很大程度上还寄希望于人工智能技术的突破。

参考文献

1 曹元大，贺海军，涂哲明，王琴. 全文检索字索引技术的研究与实现. 计算机工程，2002(6)
2,3,7 曹倩，丁艳，王超，潘金贵. 汉语自动分词研究及其在信息检索中的应用. 计算机应用研究，2004(5)
4 董小芸，刘俊熙. 自动分词在中文信息检索中的应用. 情报杂志，2003(12)
5 文庭孝，邱均平，侯经川. 汉语自动分词研究展望. 现代图书情报技术，2004(7)
6 龚汉明，周长胜. 汉语分词技术综述. 北京机械工业学院学报，2004(9)
8,9 赵伟，戴新宇，尹存燕，陈家骏. 一种规则与统计相结合的汉语分词方法. 计算机应用研究，2004(3)
10 施彤年，卢忠良，荣融，王家云. 多类多标签汉语文本自动分类的研究. 情报学报，2003(3)
11 马光志，李专. 基于特征词的自动分词研究. 华中科技大学学报（自然科学版），2003(3)
12 李方平. 无监督建立分词词表方法. 高性能计算技术，2003(6)
13 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. 计算机学报，2004(6)
14 曹娟，周经野. 一种计算机汉字串之间相关程度的新方法. 中文信息学报，2004(4)
15 王彩荣. 汉语自动分词专家系统的设计与实现. 微机处理，2004(6)

熊回香 华中师范大学信息管理系副教授。通信地址：武汉。邮编430079。
（来稿时间：2004－12－27）