第 34 卷 第 23 期 计 算 机 工 程 2008 年 12 月
Vol.34 No.23 Computer Engineering December 2008
·软件技术与数据库· 文章编号：1000—3428(2008)23—0026—03 文献标识码：A 中图分类号：TP391

基于本体和句法分析的领域分词的实现

杨晓恕，蒋 维，郝文宁
(解放军理工大学工程兵工程学院，南京 210007)

摘 要：针对基于字符串匹配的分词方法、基于理解的分词方法和基于统计的分词方法所存在的缺陷，提出基于本体和句法分析的某领域分词方法，通过建立体裁本体进行句法分析，从智能化的角度进行查词，避免了传统方法不考虑上下文信息导致的语义丢失等情况。实验结果证明，该方法可以较大地提高分词的精度。

关键词：分词；本体；体裁；句法分析

Implementation of Field Word Segmentation Based on Ontology and Syntax Analysis

YANG Xiao-jia, JIANG Wei, HAO Wen-ning
(Engineering Institute of Engineering Corps, PLA University of Science & Technology, Nanjing 210007)

【Abstract】 According to the disadvantages of the method based on character string matching, the method based on comprehension and the method based on statistic, this paper presents a word segmentation method based on ontology and syntax analysis, which avoids the disadvantages about the lost of the semantic meaning with traditional method. It analyzes the grammar by building ontology, and inquires based on the intelligence. Experimental result proves that the accuracy is improved.

【Key words】 word segmentation; ontology; types of literature; syntax analysis

1 概述

英文以词为单位，词和词之间是靠空格隔开的，而中文以字为单位，句中所有的字连起来才能描述一个意思。例如，英文句子 I am a student，用中文则为：“我是一个学生”。计算机可以通过空格知道 student 是一个单词，但是不能很容易地明白“学”、“生”2 个字合起来才表示一个词。把中文的汉字序列切分成有意义的词，就是中文分词，也称为切词。

现有的分词算法可分为 3 大类：基于字符串匹配的分词方法，基于理解的分词方法和基于统计的分词方法。

基于字符串匹配的分词方法又称机械分词方法，它按一定的策略将待分析的汉字串与一个机器词典中的词条进行匹配，若在词典中找到某个字符串，则匹配成功(识别出一个词)。这种方法最大的弊端是要有一个“充分大的”机器词典，而且对词语歧义的消除和新词的识别几乎没有任何改善。

基于理解的分词方法是通过让计算机模拟人对句子的理解，达到识别词的效果。其基本思想是在分词的同时进行句法、语义分析，利用句法信息和语义信息处理歧义现象。这种分词方法需要使用大量的语言知识和信息。由于汉语语言知识的笼统性、复杂性，因此难以将各种语言信息组织成机器可直接读取的形式。

基于统计的分词方法无需切分词典，只需要对语料中的字组频度进行统计，因此，又称为无词典分词法或统计取词方法。但这种方法也有一定的局限性，会经常抽出一些共现频度高但并不是词的常用字组，例如“这一”、“之一”、“有的”、“我的”、“许多的”，并且对常用词的识别精度差，时空开销大。

结合上面 3 种方法的优点，本文提出了一种复合式的分词方法，针对歧义消除和非登录词识别这 2 个技术难点进行了改进。

2 总体框架

如图 1 所示，复合式分词方法的处理过程为：在取得待分词的文档后，先对文本进行初步的处理(包括基于词典的词语粗切分和基于隐马尔科夫模型的词性标注)，然后对处理过的句子进行自底向上的句法分析，最后进行排歧处理和未登录词的识别，得到最优分词序列。

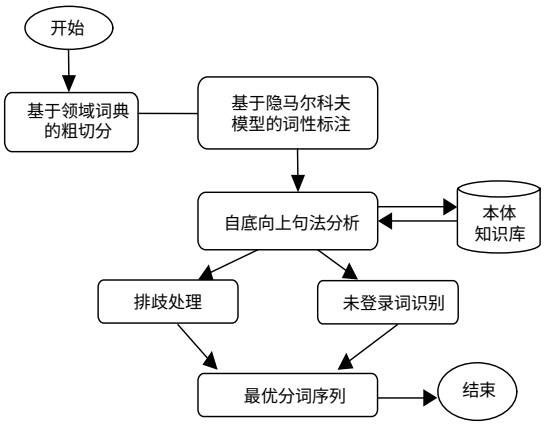

图 1 总体框架设计

基金项目：国家自然科学基金资助项目“合同战术训练评估系统”(70371039)

作者简介：杨晓恕(1976－)，男，博士研究生，主研方向：系统分析与集成；蒋 维，博士研究生；郝文宁，副教授、硕士

收稿日期：2008-04-20 **E-mail：**lijunling@nju.org.cn

句法分析通过形式化描述，使计算机能够模拟人去理解句子，对句子的合法性做出判断。句法分析的结束阶段包含了排歧阶段和未登录词的识别阶段，同时通过回溯和记录上下文信息，识别新的单词和排除有歧义的切分。

因为领域词典是有限的，所以对于一些有歧义的句子和未登录的新名词，在初切分阶段都可能造成错误切分。粗切分阶段采用的是机械分词方法，当根据词典对词语进行初步切分时，采用的是正反方向的 2 趟扫描，所以，文档中的每个句子最多可能有 2 个切分完毕的单词序列。本文通过它对文档进行初步处理，排除了许多切分可能性，提高了处理效率。词性标注是为了获取更多文档的结构信息，为之后的工作做准备。

本文对经过粗切分的单词序列进行词性的标注是通过统计的方法(隐马尔科夫模型)实现的。隐马尔科夫模型有 2 个随机过程：一个是潜在的随机过程，它是隐含的，但是该随机过程能通过另一个产生可观察随机过程序列进行观察。它可以形式化为一个五元组：$(S,\ O,\ A,\ B,\ \pi)$[1]，其中，$S=(S_1,S_2,\cdots,S_N)$ 表示状态的有限集合；$O=(O_1,O_2,\cdots,O_M)$ 表示观察值的有限集合；$A=\{a_{ij}\}$ 表示状态转移概率分布：

$$a_{ij}=P\{q_{i+1}=S_j/q_i=S_i\},1\le i,j\le N$$

$B=\{b_j(k)\}$ 表示状态 j 输出相应观察值的概率；$\pi_i=P\{q_1=s_i\}$，$1\le i\le N$ 表示初始状态分布。

对于词性标注任务来说，已知的单词序列 $w_1,w_2,\cdots,w_n$ 为观察值序列，词性序列 $c_1,c_2,\cdots,c_n$ 为隐含的状态序列。本文采用 Viterbi 算法进行角色自动标注。即从所有可能的标注序列中优选出概率最大的标注序列作为最终的标注结果。

3 分词实现

对文本进行初步处理后可得到已标注词性的单词序列，本文基于理解的方法对该单词序列进行句法分析。在句法分析过程中会使用到大量的语言知识和规则，如何将各种语言信息组织成机器可直接读取的形式是进行句法分析的关键。而本体是共享可重用的概念集合，利用本体捕获自然语言的语法知识，确定该领域内共同认可的术语(概念)，提供人和机器对该领域知识的共同理解，并给出这些概念之间相互关系的明确定义。所以，本系统分析和建立了分词阶段所需要的定义。

3.1 本体的分析

本文提出的本体定义具有多层次性，可以较好地实现本体的共享和重用，其层次如图 2 所示。

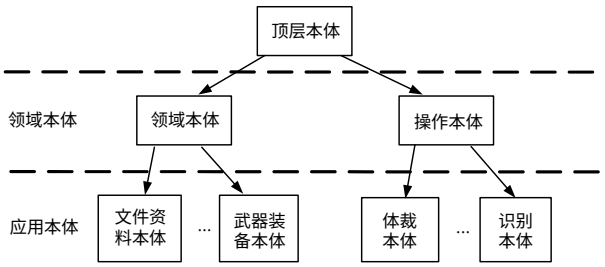

图 2 本体的层次

领域本体和操作本体的子本体是包含的关系。领域本体着重于术语的建立，它针对的是领域内对象属性的限定。而操作本体中体裁本体是对特定体裁的语法和词语的限定，识别本体是对未登录名识别中使用规则的定义。

3.2 体裁本体的建立

本文提出的体裁是指文章构成的一种规格和模式，文章的体裁一般分为一般文章和文学作品 2 大类。本文实现的环境是军事领域的一些基础数据，所以不包括文学作品这一体裁，仅含一般文章中的应用文、说明文和记叙文 3 小类，其中，应用文包含表格和公文 2 种类型[2]。文章体裁的限定对词语和句法有如下限制：

(1)对词语的限制

本文限定的体裁要求庄重、精确、严实，一般都是专业的术语(如出自军标通用语)，不采用修饰词和带感情色彩的词语，对于不同的体裁，有些同义词的意义的选择有了限定，同时它们的词性有了限定。在训练过程中通过统计可以计算出意义和词性的概率。

词语的词性可以通过形式化的语言进行描述，对它们之间的语法结构也可以形式化地描述。例如用 ADJ(形容词)，ADV(副词)，N(名词)，AUX(系动词)，Vt(及物动词)，Vi(不及物动词)，R(连词)表示词性，通过这些形式化的语言对语法结构进行描述。

一些常用的语法规则如下：

1)ADJ+N；

2)Vt+N；

3)ADJ+ADJ*+N (*表示该符号出现 0 次或 0 次以上)。

(2)对句法的限制

这类体裁的句形、句子比较单一，句形方面一般多用主谓句，且多数是完全句，排除倒装句和省略句等。

句形结构的描述：利用上文定义的形式化描述可以对句子的内部结构进行描述。

最常用的句形结构的形式化描述为：S 表示一个句子；SP 表示主语；PP 表示谓语；OP 表示宾语，由此可以定义一个句子的规则：

SP, PP, OP→S

该规则的意思是主语+谓语+宾语组成一个合法的句子。

另外一些常用的规则有

1)NP→SP；

2)VP→PP；

3)NP→OP；

4)PROP→SP；

5)N, R, N→NP；

6)V→VP；

7)Vt, N→VP；

8)COP, V→VP；

9)N, R, N →OP。

体裁对词语和句法的限制使本文分词方法的实现变得相对简单。

3.3 识别本体

建立识别本体是为了对未登录词进行识别。所以，识别本体最重要的一个属性是识别规则，该属性的常用属性值建立如下：

(1)N→NP；

(2)N, N*, PP→NP, PP；

(3)N, N*, R, N, N*→NP, R, NP；

(4)PP, N, N*, R, N, N*→PP, NP, R, NP；

(5)PP, Vt, N→PP, Vt, NP；

(6)PP, N, N*, R, N, N*→PP, NP, R, NP。

3.4 句法分析(自底向上算法)

本文采用的句法分析是自底向上的扫描分析[3]。该算法的原理如下：

定义 1 位置标记是指一个序列中的每个单词后用于表明次序的数字。

定义 2 符号标记是在句法分析过程中记录的分析结果，形式如：((词汇类符号或是短语符号)位置标记)。

算法步骤如下：

(1)设置初始状态，符号串为空，初始位置是句子的开始位置 1。

(2)按照符号串数字显示的位置标号读入词性。

(3)符号串分析。

对于刚读入的词性+最后符号，查询规则集，如果有符合的规则，则将规则读入，生成不同符号串，然后对生成的不同符号串重新从第(2)步开始进行循环操作。如果没有符合的规则，则把词性加入原符号串的尾部，生成新的符号串，重新从第(2)步开始进行循环操作，直至位置标号到句尾。

(4)排歧处理和未登录名的识别。

对生成的符号串进行体裁本体中规则属性值的匹配，生成新的符号串，对符号串进行句法合法性检查，从而消除歧义，再对剩下的符号串进行识别本体中规则属性值的匹配，进行未登录词的识别。之后将新的词语放入堆栈，最后在一篇文章结束时进行频率的计算，与阈值进行比较，将符合条件的加入索引。

3.4.1 歧义字段的处理

例：他的确切菜了。这句话被切分为 2 种情况：

情况 1：他$_{1,PROP}$/的确$_{2,ADV}$/切$_{3,Vt}$/菜$_{4,M}$/了$_{5,ADV}$。

情况 2：他$_{1,PROP}$/的$_{2,AUX}$/确切$_{3,ADJ}$/菜$_{4,N}$/了$_{5,ADV}$。

对情况 1 进行句法分析，进而消除歧义：

Step1 开始符号串()，位置标志为 0，记录位置 1 的单词的词性 PROP，符号串变为((PROP)1)。

Step2 记录位置 2 的单词的词性 ADV，检查符号串((PROP)1)，读取最后符号+刚记录的词性，查询定义的规则，没有符合，则符号串变为((PROP, ADV)2)。

Step3 记录位置 3 的单词的词性 Vt，检查符号串((PROP, ADV)2)，读取最后符号+刚记录的词性，查询定义的规则，没有符合，则符号串变为((PROP, ADV, Vt)3)。

Step4 记录位置 4 的单词的词性 N，检查符号串((PROP, ADV, Vt)3)，读取最后符号+刚记录的词性，查询定义的规则，有符合的规则 PP→Vt, N，所以，符号串改写成((PROP, ADV, VP)4)。

Step5 记录位置 5 的单词的词性 ADV，检查符号串((PROP, ADV, VP)4)，读取最后符号，则符号串变((PROP, ADV, VP, AVD)5)。

Step6 经检查，已经到最后位置，检查符号串，生成新的符号串((SP, ADV, PP, AVD))，判断符合句法，所以，情况 1 的切分是正确的。

对情况 2 进行分析：

最后生成的符号串((SP, AUX, ADV, N, AVD))不符合句法，所以，情况 2 的切分被驱除，从而消除了歧义。

3.4.2 未登录名的识别

例：中国人民解放军是人民子弟兵。

分词结果如下：

(1)中国人$_{1,,N}$/民$_{2,N}$/解放军$_{3,N}$/是$_{4,COP}$/人民$_{5,N}$/子弟$_{6,N}$/兵$_{7,N}$。

(2)中国$_{1,N}$/人民$_{2,N}$ /解放军$_{3,N}$ /是$_{4,COP}$ /人民$_{5,N}$/子弟$_{6,N}$ /兵$_{7,N}$。

2 个单词序列识别的符号串是相同的：

((N, N, N, AUX, N, N, N, N))

根据识别规则：

N,…,N, AUX, N,…,N→NP, AUX, NP

重新生成符号串((NP, AUX, NP))，因此，根据最后的符号串进行排歧处理，发现 2 个序列都是符合的，再进行未登录名的识别，最后得到的切分为：中国人民解放军/是/人民子弟兵。

现在有 3 种切分，可以将前 2 种切分中所有出现的单词进行记录，然后将第 3 种情况放入临时的堆栈。当一篇文章处理完之后对堆栈中的词语进行频率的计算，当大于设定的阈值，就为这个值建立索引，从而避免丢失文章的某些上下文信息。

4 实验结果

评测的结果选用了有标准分词答案的军事文本，采用准确率和召回率 2 个指标进行评测。其中，

正确率=识别出的正确词语数/词语总数×100%

召回率=识别出的正确词语数/标准结果集中的词语总数×100%

以下是本文算法和正向最大匹配算法分词结果的比较：

	正确率/(%)	召回率/(%)
正向最大匹配算法	86.76	86.58
本文算法	89.78	89.90

5 结束语

本文针对汉语进行领域分词，结合具体领域的特点，引入本体，通过句法分析对词语进行了识别和歧义处理，歧义的种类主要包括 2 种：交集型歧义(A/BC和AB/C)和组合型歧义(A/B和AB/)[4]。本文主要针对组合型歧义进行排歧，如何对交集型歧义进行有效处理是下一步的工作。

参考文献

[1] 俞鸿魁. 基于层次隐马尔可夫模型的汉语词法分析和命名实体识别技术[D]. 北京：北京化工大学计算机应用技术系，2004: 18-55.

[2] 方鹭飞. 中文文本体裁的自动分类机制[D]. 大连：大连理工大学计算机应用技术系, 2005: 32-70.

[3] 张晓淼. 基于神经网络的中文分词算法的研究[D]. 大连：大连理工大学控制理论与控制工程系, 2005: 26-68.

[4] 温　涛. 自适应歧义切分的汉语分词系统的设计与实现[D]. 苏州：苏州大学计算机科学与技术系, 2005: 30-67.