一种基于语音学的汉语自动分词算法

清华大学计算机系　蔡莲红　魏华武

【摘要】：

本文以汉语语音学为基础，通过对歧义词语段的分析，提出了一种实时高效的汉语自动分词算法。将这一算法运用于汉语文一语转换系统中，大大提高了系统输出语流的可懂度和自然度。

汉语自动分词是汉语自然语言理解的基础，是人工智能领域的重要课题。依照分词的目的，分词可划分为两类：一类是语法分词，应用于自然语言理解。另一类是语音分词，应用于文字－语音转换。语音分词是根据人们在表达上的习惯和语音流的停顿及强弱变化，在每个词之间插入长度不等的空语音符号（停顿）。提高语音流的节奏和自然度，以利于听者的理解。

建立一个好的自动分词系统，有两个关键：一是词库，二是分词算法。对于一个实时文－语转换系统来说，要求语音分词速度快，能处理岐义词串和未登录词。词库不宜过大。本系统的基本分词词库中的词条约8万，包括二字词、三字词和四字词。词条的语音学信息包括：音调、音变（儿化 轻声）、文本到语音的转换、语音到音库的索引。为标记一字多音的特性，还建立了多音字词典。

基于语音分词的目的，我们确定了匹配策略及岐义处理算法。本系统采用了正向扫描逆向极大匹配分词法（PSBM）。先从句首向后扫描匹配，将句子分隔成词语段，再处理长度大于四的岐义词串和连续单字词。寻找岐义词串使用从前到后的最长词匹配原则，而划分岐义词串使用从后往前的最长词匹配原则。

在分词处理中，遵循一些规则：<1> 坚持将句子分成一个个词。<2> 字粘法规则，如数词优先与其后的量词连接成词。<3>连续单字词的"二三原则"。分词时，未登录词常被分成单字词。考虑到汉语语流节奏性强的特点，本系统按"二三原则"把连续单字组词。

本分词系统的分隔速度快，每秒可分几千个词。在文－语转换系统中，以句子作为语声流输出单位。每次仅对一个句分词，故时间短暂，完全不影响语声流输出的实时性。

本分词系统已应用于我单位研制的汉语文－语转换系统 TH—Speech 中，这是一个高级的语音输出系统。它能实时把计算机内的文本转换成连续自然的语流。分词是文一语转换系统的重要组成部分，是提高语流自然度、可懂度、节奏感的关键，是音变处理的基础。

参考文献：

万里，赵立泰，<<汉语口语表达学教程>>，1990.4 北京师范大学出版社

刘开瑛，郭炳炎，<<自然语言处理>>，科学出版社

魏华武，蔡莲红，"汉语普通话全音词汇智能合成系统" 1992.9 计算机世界月刊