汉语自动分词的近邻匹配算法及其在QHFY汉英机器翻译系统中的实现

顾萍　吴涛　茅于杭

（清华大学自动化系）

摘要： 汉语自动分词的分词精度主要体现在对多义切分字段的处理上，不同的算法和词典设计对分词的速度影响也较大。本文介绍了在QHFY汉英机器翻译系统中实现的汉语自动分词的近邻匹配算法，主要叙述了多义切分字段的识别及校正处理方法。

关键词： 汉语自动分词，近邻匹配，算法，汉英机器翻译系统

THE ADJACENT MATCHING ALGORITHM OF CHINESE AUTOMATIC WORD SEGMENTATION AND ITS IMPLEMENTATION IN QHFY CHINESE-ENGISH SYSTEM

GuPing,WuTao,MaoYuhang

(Automation Department,Qsinghua University,Beijing 100084)

Abstract: The accuracy of Chinese automatic word segmentation is mainly reflected on the process of polysemantic ambiguous word group.Different algorithm and dictionary design will have great influence on the speed of word segmentation .This paper presents the adjacent matching algorithm implemented in QHFY Chinese-English Translation System and mainly describes the recognition and correction method of polysemantic ambiguous word group.

Key word: Chinese automatic word segmentation , adjacent matching , algorithm , Chinese-English Translation System.

一. 引言

汉语句子中词之间无明显的界限标志，因此汉英机器翻译第一步要解决的就是进行句子的自动分词。理想的自动分词算法需要综合词法、语法和语义信息，分词算法具有相当的复杂度。在机器翻译系统中，分词系统位于句法、语义分析系统之前，因此，我们不必为分词系统付出太多的语法、语义分析，在考虑分词精度和系统复杂度上需要一个折衷的方案。由于待翻译文本中歧义切分字段的出现不是很频繁，而且大都有一定的规律可循。因此，本文提出一种基于MM方法的正向最大匹配法，并辅以对多义切分字段的校正处理，可以处理绝大部分的多义切分字段，保证一定的分词精度与速度。

二. 分词中歧义切分字段的处理

目前，汉语自动分词已取得了不少的研究成果，但无论使用什么自动分词方法，都不可避免地会出现错误切分。要进一步提高分词精度，必须对原文中的多义切分字段采用正确的算法，同时与词典库的容量也有很大关系。

下面首先分析一下多义切分字段的类型及其产生原因:

1.交集型歧义切分字段（简称交集字段）

小写字母代表汉字，希腊字母代表汉字串。

汉字串 β 既是词 x1x2...xiβ 的后缀，又是 βy1y2...ym 的前缀，且句型"...x1x2...xiβy1y2...ym..."满足语法规则 P 并为人们所使用，则称 β 为交集字串，x1x2...xiβy1y2...ym 为交集字段。交集字段中交集字串的个数称为链长。例如:"不同情况"是一个链长为 2 的交集字段，其中"不同"、"同情"、"情况"都是词。

2.多义组合型歧义切分字段（简称多义组合字段）

设 $\beta=\alpha1\alpha2\in W$,W 是词的集合，且 $\alpha1,\alpha2\in W$,如果句型"...$\alpha1\alpha2$"在汉语句子中出现，我们称 β 为多义组合字段。例如:"把手"是一个多义组合字段，在不同的语境下，它可以分为"把／手"或"把手"。

据统计，汉语中交集型歧义字段占全部歧义字段的 90%以上。所以，处理好交集型歧义字段的正确切分就可以保证一定的分词精度。本文设计的汉语自动分词是针对汉英机器翻译系统的输入文本，在一般文本中，那些需要进行上下文理解才能正确切分的歧义字段，出现频率是很低的。因此，在近邻匹配分词算法中，主要针对如何正确识别交集字段和多义组合字段，以及对交集链长为 1、2 的交集字段的校正切分处理，并且应用规则知识来对多义组合字段的切分进行校正。

下面分别介绍对交集链长为 1，2 的交集字段的处理方法:

设一个句中出现...a1a2...aiβc1c2...cj...

其中 S=a1a2...aiβc1c2...cj,

并且 a1...ai$\beta\in$W, βc1...cj$\in$W,

1.链长为 1，这种情况又分为以下几个方面:

1) a1...ai$\in$W,但 c1...cj$\notin$W ——> a1...ai／βc1...cj

如:"用方块图形式"，其中"图形式"是一个交集链长为 1 的交集字段；按照以上规则，应切为"图／形式"。

2) a1...ai$\notin$W,但 c1...cj$\in$W ——> a1...aiβ／c1...cj

如:"实现在情报工作方面的自动化"，其中"实现在"是一个交集链长为 1 的交集字段；按照以上规则，应切为"实现／在"。

3) a1...ai$\in$W,且 c1...cj$\in$W ——> a1...ai／βc1...cj

如:"他看见一只白天鹅"，其中"白天鹅"是一个交集链长为 1 的交集字段，并且,"白天"、"天鹅"是二字词,"白"、"鹅"是一字词，此时，近邻匹配算法认为最近匹配成的词"天鹅"为确定的切分。即切为:"白／天鹅"。

当然，这就不可避免地会造成错误切分，比如:

例 1.他的确切地址在这。

例 2.这块肉的确切得好。

在例 1 中，“的确切”应切为“他／的／确切／地址／在／这”。

在例 2 中，“的确切”应切为“这／块／肉／的确／切／得／好”。

这里，“的确切”属于多义组合字段，因为这个三字字段，可以是“2+1”切分，也可以是“1+2”切分，需要根据上下文来决定其切分。后面将讲述对它的处理方法。

2.对于链长为 2 的交集字段，采用[1]中介绍的自然成词方法进行切分。

如：“昨天下午”切为“昨天／下午”。其中“昨天”、“天下”、“下午”是二字词，且交集链长为 2。

经过初步试验表明，通过对以上几类交集字段的校正处理，可以正确切分绝大部分交集歧义，错分率在 1／300 左右，达到实用化要求。

对于多义组合字段，主要利用规则库知识加以校正。

例 3.他将来上海。

例 4.将来的上海会有严重污染。

“将来”一词为多义组合字段，在例 3 中，应切为“将／来”；在例 4 中，应切为“将来”。在处理这类歧义字段时，我们针对个别词建立相应的规则，并在主词典中标记出其为多义组合字段，当分词扫视到该词时，发现其为多义组合字段词，则转而调用相应的规则知识进行判断，根据规则的指示来正确切分该字段。下面以“将来”一词为例，说明规则的格式：

汉字词	将来	
左模式	@	／＊ @ 表示可忽略 ＊／
右模式	Nd	／＊ Nd 表示地点名词 ＊／
标注	+1,+1	／＊+1,表示指针移动一个汉字，设立一个切分标记＊／

在程序中，解释执行这条规则，就可将例 3 正确地切分为：

例 3.他／将／来／上海。

类似地，对于“的确切”这个三字字段的多义组合字段，其规则如下：

汉字词	的确切		汉字词	的确切
左模式	Ne	／＊表示物的名词＊／	左模式	Nq
右模式	得		右模式	N
标记	+2,+1		标记	+1,+2

系统在第一次扫视中，会将例 2 错误地切分为“这／块／肉／的确切／得／好”。当系统进行第二次扫视中，发现“的确切”为多义组合字段词，因此，就会检查其是否符合规则中的条件。在例 2 中，“的确切”左面为名词，右面为“得”，符合规则中的描述，因此，应将其切为“／的确／切”。因为“的确切”有两条规则，因此存在规则的优先顺序问题。系统将约束力强的规则放在前面，并且认为最后一条规则为缺省的规则：即当某词串与任何一条规则都不匹配时，选用该规则进行匹配切词。在例 1 中，“的确切”左面为代词“他”，右面为普通名词“地址”，可与第二条规则匹配，即将其切为“的／确切”。这样做的优点是：规则与程序分开，用户可不断地增加规则知识，而不必改变程序，就可以提高系统的切词精度；同时，规则库也拥有其相应的索引库，查询起来非常方便。

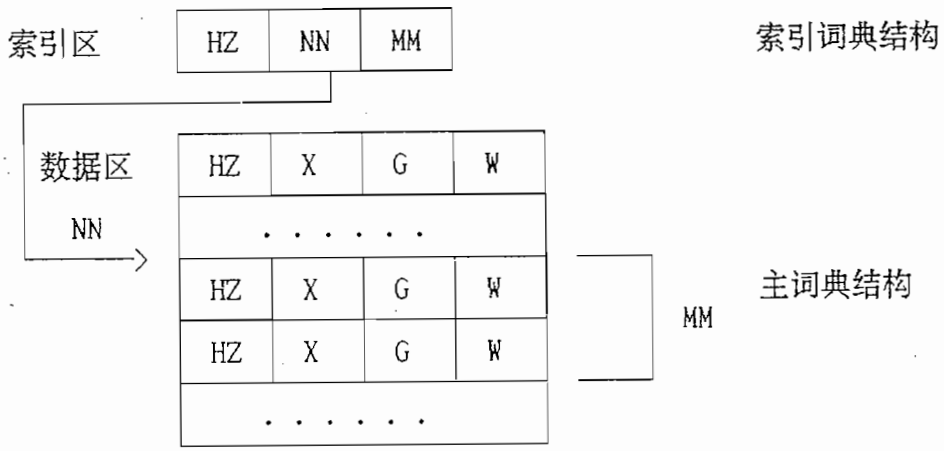
三.词典结构设计

由于在分词过程中，会用到少量的词法信息。因此，本系统的分词词典就是整个翻译系统的主词典，其中包括了词条、语法信息以及一些特征控制信息。词典中相同首字的词条按长度

从大到小排列，词典以数据库文件存放，同时也是个知识库。

主词典容量很大，词条约 6 万条左右，因此，必须对其建立一个索引词典。这样，索引可以全部读入内存，主词典的数据存在外存，当需要哪个部分的信息时，再读入相应的词条信息。

分词所用的索引词典也采用数据库文件格式，存放词首字、以该字为首的词条在主词典中偏移量以及以该字为首的词条个数。



索引区中的 NN 用来帮助确定以 HZ 为首的词条在主词典中存放的位置，MM 表示以 HZ 为首的词条个数。这样组织词典结构有以下特点:

1.对分词算法程序的支持能力强，一次读盘即把一个首字下的所有词条读入内存，减少了访问词典的次数，提高了分词速度。

2.词条按从大到小的顺序排列，对分词算法中的最长减字匹配法的操作十分便利。

3.同一时间只有一个首字的词条进入内存，大量节省内存空间。

四.近邻匹配分词算法及其实现

所谓近邻匹配分词算法，即在 MM 方法基础上，通过词尾字向后搜索构词，进行匹配；若最近一次词尾字不能与后面的词串构成词，则确定当前切分的字段。其基本思想是"机械匹配分词+歧义校正"。具体讲，它包括以下几个方面:

1.在作最长匹配时，根据词典中首字下所有词条的最大长度来与待切分字串匹配；并记录匹配成功的最长词条长度及次长词条长度。

2.将扫描指针移向最长词条的词尾字，继续向后搜索匹配，看是否可构成词。

3.第二次扫描，扫描指针扫视有无多义组合字段，若有，则激活该词的多义组合字段切分校正规则，通过规则解释器对第一次切分的结果加以校正；若无，则转 4。

4.输出切分结果。

这种算法的特点在于，在分词过程中，检查出错误后，才予以校正，即算法针对性较强，可以处理大部分交集歧义以及多义组合歧义；与其它的完全利用规则知识库进行分词的系统相比较，它不会因为规则顺序的排列不当，而引起错误的切分。经过试验测试，发现产生错误切分的原因主要有:

第一，词典中无该词条，即词典不完善。

第二，由于算法只处理交集字段中的链长为 1，2 的歧义字段，因此，当超出此范围时，可能会出现错误切分。

第三，对多义组合字段的切分，由于规则知识不全而引起的错误切分。

目前，此算法已在 IBM-PC 机上实现，采用面向对象的设计语言 BORLAND C++编程。由于 C++特有的性质，使得系统的数据封装性好，程序可读性强。对词典、规则采用统一的类的形式设计并实现，这样，可与机器翻译系统其它几个模块取得统一，便于系统的完善。

参考文献:

[1] 梁南元，汉语计算机自动分词知识，《中文信息学报》，1990.2

[2] 何克抗等，书面汉语自动分词专家系统设计原理，《中文信息学报》,1991.2

[3] 姚天顺等，基于规则的汉语自动分词系统，《中文信息学报》,1990.1

附录:

待切分文本:

1.研究生一般年龄较大。
2.研究生命起源。
3.这个研究所不大。
4.这项研究所涉及的问题很复杂。
5.实现在情报工作方面的自动化。
6.不同情况下有不同解释。
7.用方块图形式来描述。
8.他看见一只白天鹅。
9.让位移小于 10 毫米。
10.独立自主和平等互利原则。
11.这支歌太平淡无味了。
12.产品需求和规格说明。
13.其实也是看中和中国大陆做生意的机会。
14.战事已经有了结局。
15.发展中国家的经济状况很好。
16.发展中国家庭副业。
17.使用户外天线要注意避雷。
18.使用户满意的做法。
19.昨天下午他不在。
20.他将来上海。
21.将来的上海会有严重污染。
22.他从马上下来。
23.老师叫你马上去。

切分后结果如下:

1 . 研究生 一般 年龄 较 大 。
2 . 研究 生命 起源 。
3 . 这 个 研究 所 不 大 。
4 . 这 项 研究 所 涉及 的 问题 很 复杂 。
5 . 实现 在 情报 工作 方面 的 自动化 。
6 . 不同 情况 下 有 不同 解释 。
7 . 用 方块图 形式 来 描述 。
8 . 他 看见 一 只 白 天鹅 。
9 . 让 位移 小于 10 毫米 。
10 . 独立 自主 和 平等 互利 原则 。

11. 这 支 歌 太 平淡无味 了 。
12. 产品 需求 和 规格 说明 。
13. 其实 也 是 看中 和 中国 大陆 做 生意 的 机会 。
14. 战事 已经 有 了 结局 。
15. 发展中国家 的 经济 状况 很 好 。
16. 发展 中国 家庭 副业 。
17. 使用 户外 天线 要 注意 避雷 。
18. 使 用户 满意 的 做法 。
19. 昨天 下午 他 不 在 。
20. 他 将 来 上海 。
21. 将来 的 上海 会 有 严重 污染 。
22. 他 从 马 上 下来 。
23. 老师 叫 你 马上 去 。