中文文本歧义切分技术研究*

郑家恒　　刘开瑛
（山西大学计算机科学系，太原，030006）

摘　要： 汉语文本的歧义切分问题是汉语自动分词的难点之一。本文采用语料库方法，从大量真实文本中，抽取歧义型词语字段，分别建造了交集型歧义字段库和多义型歧义字段库。分析了各类歧义型词语字段的特点和语言现象后，在统计数据的基础上，建立了歧义切分规则，我们对4646个歧义字段进行了测试，切分正确率可达87%。

关键字： 歧义切分、歧义字段、自动分词

The Research of Ambiguity Word—Segmentation Technique for the chinese text

Zheng Jiaheng Liu Kaiying
Dept. of Computer science Shanxi University (030006)

ABSTRACT: The ambiguity partition of Chinese text is one difficult problem in Chinese automatic word—segmetation. By adopting the corpus method to extract the ambiguous phrase from a very large number of real texts and to build ambiguous phrase base of overlap and combination, this paper analysed the characterstics and language universal of all kinds of ambiguous phrase, built the disambiguous rules which based on the statistic data. We tested 4646 ambiguous phrase. The accuracy of word—segmetation is above 87%.

Key words: ambiguity partition, ambiguous phrase, automatic word—segmetation.

一、引言

在汉语文本中存在着大量的歧义字段，解决歧义切分自然成为自动分词中一个重要的问题，对歧义字段切分的处理能力，将直接影响汉语自动分词系统的切分正确率。1995年，863智能机主题组对汉语书面语自动分词软件进行了评测。从评测结果看，交集型歧义切分正确率最高可达78%，多义型歧义切分正确率最高可达59%。由此可见，对歧义切分的研究仍需投入更大精力[1]。

* 国家863高技术资助课题(863—306—03—09—4)，国家自然科学基金资助课题(69673011)

目前，国内对歧义字段切分的研究，大多数都是以词性搭配，作为制定歧义切分规则的依据[3]。但由于词存在着兼类问题，在歧义切分的过程中，去解决词的兼类问题，势必增加歧义切分的难度。那么对歧义字段的切分，是否可以不完全依赖于词性，而将歧义字段的切分策略建立在统计数据的基础上呢？本文试图运用语料库方法，在对随机抽取的 180 万汉字新闻语料加工、统计的基础上，建立了歧义字段库，通过对歧义字段特点的分析，制定了歧义切分规则库。在对 4646 个链长为 1 的歧义字段的封闭测试中，正确率可达 87%；在开放测试中，正确率可达 81%。

二、歧义字段库建造

通常人们依歧义字段结构的不同，把歧义字段分为交集型歧义字段和多义型歧义字段。交集型歧义字段是指在字段 ABC 中，AB∈W，BC∈W。(A，B，C 为字串，W 为词表)如：当代表，“当代”，“代表”都是词。多义型字段是指字段 AB，AB∈W，A∈W，B∈W。如：个人，“个人”，“个”和“人”都是词。这两类歧义字段在构成歧义方式和语言特点都不尽相同，在切分过程中，发现歧义字段的途径和切分策略上也不相同。为了更系统地研究歧义字段的特点和规律，我们建立了交集型歧义字段库和多义型歧义字段库。

(一)交集型歧义字段库

1、抽取交集型歧义字段算法

在“信息处理用现代汉语常用词词表”[3]中，在 43570 总词条下，二字词有 3 3527 个，占总词条的 77%左右。这是对静态词表中占有的比例。对随机抽取的 120 万汉字语料切分结果的统计数据表明，词次大于或等于 1 的词条共有 20289，其中二字词有 13913，二字词占总词条数的 68%。我们认为不论从静态词表还是动态真实语料库中，两字词都占了绝对高的比例，所以建立交集型歧义字段库，首先应以二字词为主体。在语料中，逐词寻找二字交集型歧义字段。算法如下：

设汉字串 $N_1N_2\cdots\cdots N_n$(N_i 为字)。以二个汉字为切分单位，在字串 $N_1N_2\cdots N_n$ 中逐词扫描找到第一个为词的位置 i，即 N_iN_{i+1} 为词，从 N_{i+1} 开始，再逐词扫描找到不成词的位置 j，即 N_jN_{j+1} 不为词。则 $N_iN_{i+1}N_{i+2}\cdots\cdots N_j$ 为交集歧义字段。交集字串的个数称为链长。如：昨天下午。“昨天”、“天下”和“下午”都为词，交集字串为 2，所以链长为 2。

对于三个字、四个字为一切分单位的交集型歧义字段的获取问题和二个、三个、四个字为一切分单位的混合型交集型歧义字段的获取问题，完全可以用以上介绍的算法进行。

2、交集型歧义字段库结构及规模

为了研究歧义现象的规律，我们不仅抽取歧义字段，还抽取歧义字段前后两个汉字连同歧义字段作为歧义字段片语。同时统计了在 180 万语料中出现的频率。

交集型歧义字段库的结构由序号、歧义词语、类型、歧义词语前后关联片语、频率和链长等六项构成。

序号	歧义词语	类 型	频率	链长	歧义词语前后关联片语

其中，类型是指交集型歧义或多义型歧义；链长是指交集词语中字的个数；歧义词语前后关联片语是在交集型歧义词语前后各增加二至三个字。例如：

序号	歧义词语	类型	频率	链长	歧义词语前后关联片语
54	把手表	交集	1	1	赶紧把手表装进
3500	在世界	交集	54	1	20日在世界屋脊
3501	在朝鲜	交集	10	1	韩国在朝鲜核问题
5000	就是	多义	2		交点就是空中
5001	就是	多义	2		想到，就是国内

交集型字段库共收录了9500条交集型词语，其中以二字为一切分单位的歧义词语有8378条。

3、交集型歧义字段库的来源

我们从2000万新闻语料中，随机抽取180万汉字，作为抽取交集型歧义字段库的语料来源。以国标GB13715为标准，建立了4.5万常用词词表，其中39006词条来源于“信息处理用现代汉语常用词词表”，其余6000条是从对120万汉字语料的切分结果中收集的。如：夏季、显示器、项目、纪念等。在抽取交集型歧义字段时，是以4.5万常用词词表为依据，判断字串N_iN_{i+1}是否为词。

（二）多义型歧义字段库

我们在180万语料中，随机抽取了10万语料进行多义型歧义词语的统计。在10万语料中，二字词有5006条，其中多义型歧义词有27条，占1%；三字词有912条，其中多义型歧义词有10条，占1%；四字词有289条，其中多义型歧义词有4条，占1%。

例如：

不同　朝向/不同/的/出入口　　就是　交点/就/是/空中/飞机

准备/时间/的/不/同　　想到/,/就是/国/内……

由于多义型歧义字段只占歧义字段的16%[4]，解决这类歧义切分问题，需要语义知识和上、下文信息，我们将在今后作更进一步的研究。

三、交集型歧义字段的特点。

1. 我们对180万汉字新闻语料作了统计，结果如下：

链长	歧义字段数	歧义字段片语数	歧义字段次数
1	4646	10210	14581
2	3409	5790	7632
3	231	279	381
4	92	111	149
总计	8378	16390	22743

从统计数据可以看出：链长为1和链长为2的歧义字段分别占歧义字段总数的55%和

41%;歧义字段出现的次数,分别占歧义字段总次数的 64%和 34%。这二者合起来,占歧义字段总数的 96%,占歧义字段次数的 98%。如果解决好链长为 1 和链长为 2 的交集型歧义字段的切分问题,就能大大提高歧义切分的正确率。

2. 不同链长歧义切分的特点:

·对于链长为 1 的交集型歧义字段 ABC,切分结果有如下 4 种情况:

a)ABC 切分为 A/BC　如:出自己　切分成:出/自已

b)ABC 切分为 AB/C　如:出现在　切分成:出现/在

c)ABC 切分为 ABC　如:传染病

d)抽取的信息不够

180 万汉字新闻语料统计结果如下:

类 型	歧义字段数	歧义字段次数
A/BC	2330	6258
AB/C	1827	5505
ABC	374	2604
信息不够	123	217
总　　计	4654	14581

从表中可以看出,A/BC 和 AB/C 型占到 89%,而 AB/C 型和 ABC 型在正向大匹配扫描切分时,可以得到正确的切分。因此,解决链长为 1 的切分难点集中在解决 A/BC 型切分问题上。

·链长为 2 的交集型歧义字段 ABCD 中,切分的类型有如下几种情况:

类型	歧义字段	次数
A/BC/D A/B/CD	55	103
AB/CD	3331	7489
ABC/D A/BC/D AB/C/D	7	8
ABCD	5	5
信息不够	11	11
总　　计	3409	7616

从表中可以看到,AB/CD 型占到 98%,如:已经过去,切分结果:已经/过去。对链长为 2 的歧义字段切分,一般切分为 AB/CD。

·链长为 3 的交集型歧义字段 ABCDE 中,AB、BC、CD、DE 分别为词,从切分结果看出主要解决前三个字的歧义切分。

如:为人民工作,切分为:为/人民/工作。只要把"为人民"切分为:为/人民,就能正确切分

“为/人民/工作”。

·链长为 4 的交集型歧义字段 ABCDEF 中，AB、BC、CD、DE、EF 分别是为词，但在切分结果中均切为 AB/CD/EF。

如：QS 中国产品质量，切分为：中国/产品/质量。

3. 同一歧义字段，在不同歧义字段片语中的切分特点。

从统计结果可知，链长为 1 的歧义字段为 4646，其中只有 8 个在不同片语中切分结果不同，不到 1%；链长为 2 的歧义字段有 3409，同一个歧义字段，在不同片语中，不存在不同的切分结果。

如：歧义字段“出自己”，在下列片段中均切分为：出/自己。

歧义字段片语	切分结果
准备派出自己的优秀运动员	准备/派/出/自己/的/优秀/运动员
以实际行动献出自己的爱心	以/实际/行动/献/出/自己/爱心
许多人奉献出自己的爱心	许多/人/奉献/出/自己/的/爱心
为了打出自己的特色	为了/打/出/自己/的/特色
我掏出自己的钱交给他	我/掏/出/自己/的/钱/交/给/他

链长为 1 的交集型歧义字段中，存在不同切分结果的例子：

“从小学” 姐妹/三/人/从/小学/上/到/中学。

她/从小/学/戏剧/表演。

“以北约” 确立/以/北约/为/核心/的/军事/力量。

兴平/市/以北/约/十五/公里。

由于同一歧义字段，在不同片语中有不同切分结果的占很小的比例，因此，在研究歧义切分规则时，可以把重点放在解决不同歧义字段的切分上，而对极个别存在不同切分结果的歧义字段的切分按特例处理。

四、交集型歧义切分规则库

针对交集型歧义字段特点：交集型歧义规则库有以下几种情况：

·链长为 2 的 ABCD 型，均切分为 AB/CD。

·链长为 4 的 ABCDEF 型，均切分为 AB/CD/EF。

·链长为 3 的 ABCDE 型，执行链长为 1 的规则，解决前三个字的切分。

·链长为 1 的 ABC 型，是研究的重点，且 A/BC 型是切分难点。下面主要对这类歧义切分规则作讨论。

设：字串 $b_1b_2n_1n_2n_3c_1c_2$ 中 $n_1n_2n_3$ 为交集型歧义字段；b_ic_i 均为字。

W 为词表集。

W_1＝{1，2，3，4，5，6，7，8，9，0，一，二，三，四，五，六，七，八，九，零，两}；

W_2＝为字表集，此字表收集了 2330 个 A/BC 型义集歧义字段中第一个汉字 A。如：国、地、大、上、在等。

W_3＝为特例切分集，此集收集了 W_2 中对应字的歧义字段 ABC，应切分为 AB/C 的字段

AB。如:以“在”字打头的歧义字段有 33 个,其中 32 个应切分为“在/BC”的形式,只有一个歧义字段“在于是”应切分为“在于/是”,则把“在于”收入 W_3 中。

规则 1　若 $n_1n_2n_3 \in W$,则 $n_1n_2n_3$ 为词。

规则 2　若 $b_2n_1n_2n_3$ 或 $n_1n_2n_3c_1$ 是 AABB 重叠形式,则 $b_2n_1n_2n_3$ 为词,或 $n_1n_2n_3c_1$ 为词。

规则 3　若 $b_1b_2n_1 \in W$,则 b_1b_2 和 n_2n_3 分别为词。

规则 4　若 $n_3c_1c_2 \in W$,则 n_1n_2 和 $n_3c_1c_2$ 分别为词

规则 5　若 $b_2n_1n_2n_3 \in W$,则 $b_2n_1n_2n_3$ 为词

规则 6　若 $n_1n_2n_3c_1 \in W$,则 $n_1n_2n_3c_1$ 为词

规则 7　若 $b_2 \in W_1$,则 n_1 和 n_2n_3 分别为词

规则 8　若 n_1 的词性 XN1∈{动词,连词,介词,前接成分},则 n_1 和 n_2n_3 分别成词。

规则 9　若 $n_1 \in W_2$,且 $n_1n_2 \notin W_3$,则 n_1 和 n_2n_3 分别为词。

规则 10　若 $n_1 \in W_2$,且 $n_1n_2 \in W_3$,则 n_1n_2 和 n_3 分别为词。

不符合上述规则的,全部切分为 n_1n_2/n_3。

将规则运用到 4646 个链长为 1 的歧义字段,进行封闭歧义切分测试,正确率可达 87%;从 2 万新闻语料中抽取链长为 1 的交集型歧义字段,进行开放测试,正确率可达 81%。

五、结束语

实验证明:运用语料库方法进行歧义字段的切分是可行的。但还存在以下几个问题:

1、字表 W_2 和特例切分集 W_3 是建立在 180 万新闻语料基础上的。如果语料扩大或语料的类型改变了,如何通过自动学习的方法,补充、调整 W_2 和 W_3,是下一步研究的问题。

2、词表是抽取歧义字段的基础,到目前为止,国内还没有一部公认的供信息处理用的分词词表。本文介绍的基于语料库的切分方法,可以用于任意常用词表。

歧义切分仍然是自动分词中的难题,仅仅依靠统计数据是不够的,尤其是多义型歧义字段的切分,还要依赖于语义、语境,上下文等更多的信息。

参考文献

[1]. 刘开瑛,现代汉语自动分词评测技术研究,语言文字应用,1997. 1, P101－106.

[2]. 侯敏等,汉语自动分词中的歧义问题,计算语言学进展与应用, 1995. 10,P81－87.

[3]. 刘源等,信息处理用现代汉语分词规范及自动分词方法,清华大学出版社,1994. 6.

[4]. 白栓虎,汉语词切分及词性自动标注一体化方法,计算语言学进展与应用,1995. 10, P56－61.

[5]. 刘开瑛、郑家恒等,面向汉语篇章的句法分析器,国家七五重点科技攻关项目(68－05－10)鉴定材料,1991 年.