中 文 信 息 学 报
第 20 卷 第 1 期 JOURNALOFCHINESEINFORMATIONPROCESSING Vol. 20 No. 1

编者按 中国中文信息学会与兄弟学会于 2005 年 8 月在南京师范大学成功地召开了“全国第八届计算语言学联合学术会议(JSCL-2005)”。会议的程序委员会向本刊推荐出六篇论文,编辑部得到授权,在此发表,以飨读者。

文章编号:1003-0077 (2006) 01-0001-06

基于语料库的高频最大交集型歧义字段考察*

李 斌,陈小荷,方 芳,徐艳华
(南京师范大学文学院,江苏 南京 210097)

摘要:交集型歧义是中文分词的一大难题,构建大规模高频最大交集型歧义字段(MOAS)的数据库,对于掌握其分布状况和自动消歧都具有重要意义。本文首先通过实验指出,与 FBMM相比,全切分才能检测出数量完整、严格定义的 MOAS,检测出的 MOAS在数量上也与词典规模基本成正比。然后,在 4 亿字人民日报语料中采集出高频 MOAS14906条,并随机抽取了 1354270条带有上下文信息的实例进行人工判定。数据分析表明,约 70% 的真歧义 MOAS存在着强势切分现象,并给出了相应的消歧策略。

关键词:计算机应用;中文信息处理;最大交集型歧义字段;全切分;强势切分
中图分类号:TP391 **文献标识码**:A

CorpusBasedInvestigationonHighFrequentMaximal OverlappingAmbiguityStringinChineseWordSegmentation

LIBin,CHENXiao-he,FANGFang,XUYan-hua
(SchoolofChineseLanguageandLiterature,NanjingNormalUniversity,Nanjing,Jiangsu210097,China)

Abstract: OverlappingambiguityisstillanopenissueinChinesewordsegmentation.Thispapermakesadeepinvestigation onMaximalOverlappingAmbiguityString (MOAS).First,wediscussthedisadvantageofusingFBMMtodetectOAS. Then,bywordomni-segmentation,wecollect14906highfrequentMOASsfromPeople'sDailycorpuswhichcontainsabout 400Mcharacters.FortheseMOASs,1354270samplesentencesarerandomlyselectedandmanuallylabeled.Theresults showthatabout70%ofMOASswithtrueambiguityhaveastrongbiastowardsonesegmentation,andconsequently,adis-ambiguationstrategyfondealingwithoverlappingambiguitiesisputforward.

Keywords:computerapplication;Chineseinformationprocessing;maximaloverlappingambiguitystring;wordomni-segmentation;biasedsegmentation

1 引言

未登录词和切分歧义是汉语自动分词中的两大难题。据统计,在切分歧义中,85% 以上是交集型歧义[5]。目前已有的消歧策略主要是基于各种统计和规则进行的,如,基于词概率[6]、词的二元模型[3]、贝叶斯分类器[1]和混合策略[4]等方法。在中小规模开放测试中,消歧正确率

* **收稿日期**:2005-05-11 **定稿日期**:2005-10-31
基金项目:南京师范大学 211 资助项目(1240702504)
作者简介:李斌(1981—),男,硕士,主要研究方向为计算语言学.

可以达到 90% 以上。

此外,基于大规模语料的考察与分析也为交集型歧义的消解提供了重要依据。孙茂松等(1998)[8] 以 1 亿字的新闻语料为基础进行了交集型歧义字段的调查,并提出了最大交集型歧义字段(MaximalOverlappingAmbiguityString,MOAS)* 的概念。MOAS的特性在于,它不再与周围任何词语形成更大的交集型歧义字段,具有较强的独立性。该文使用 112967 个词条的词典和全切分的检测方法,从语料中提取出 23 万个不同的 MOAS。其中最高频的 4619 个段型,覆盖了全部段例的 50% 以上。通过语料观察和人工内省发现,绝大部分的高频 MOAS只有一种可实现的切分形式(称之为"伪歧义"),而具有多种切分形式的真歧义字段,数量十分有限。进而提出了通过查询数据库的方法对高频的伪歧义和真歧义 2(倾向于一种切分形式)进行消歧。

文献[1] 使用规模为 93700 个词条的词典和正逆向最大匹配法,从 6.5 亿字语料中采集了 73 万条不同的交集型歧义字段(OverlappingAmbiguityString,OAS),挑选出最高频的 47000 条,为每一条 OAS从语料中随机地抽取 20 个例句。经过人工判定后,形成 41000 条词例化的消歧规则。其规则如:"信心地 ⇒信心/地"。

这两项调查对于认识 OAS都有着重要意义,但也存在着一定的问题。前者主要是为了搞清楚高频真歧义和伪歧义字段的数量,建立用于消歧的伪歧义库。对真歧义的判定侧重于定性分析,没有对每个真歧义字段的各种切分形式之间的比例进行量化。后者则偏重于寻找针对伪歧义的词例化消歧规则。但每个 OAS的实例较少,很难统计到多少真歧义字段。此外,两者的具体做法也存在着一些差别。首先,对于 OAS的检测方法不同,前者是全切分,后者是正逆向最大匹配法。其次,使用词典的规模也不相同。在切分形式的判定方法上,前者是语料观察和人工内省并举,后者只观察抽样实例。这些差别也会在一定程度上影响到调查的结果。

本文在大规模真实语料的基础上对交集型歧义字段展开调查。在检测方法上,我们指出了正逆向最大匹配法的不足,同时观察了词典规模对 MOAS数量的影响。我们通过在语料中的大量抽样(100 例/条)来获取足够多的实例支持,然后借助抽样观察和人工内省判定歧义,从而建立一个高质量的高频 MOAS切分实例库进行统计分析,以服务于交集型歧义的消歧工作。

2 MOAS 的检测方法及相关问题

检测交集型歧义字段主要有 2 种方法。一种是正逆向最大匹配法(Forward&Backward MaxMatch,FBMM),即,对于一个汉字串,分别使用正向和逆向最大匹配法进行切词,二者切分结果不一致的地方则为 OAS。该方法检测速度快,但存在错检和漏检现象,参见表 1。

可以看出,FBMM不仅有漏检的情况,而且检测到的歧义字段未必是"最大"交集型歧义字段。如果仅根据其检测结果来建立伪歧义字段的消歧数据库,则无法给出正确消歧规则。

另一种是全切分的检测方法,即,对于一个汉字串,在由词表词构成的有向无环图中,找出各条路径上的 MOAS。一般来说,MOAS是大于词表词的,但也可能与词表词相同,如,在词表中有"同盟军"、"同盟"、"盟军",则"同盟军"本身即为一个 MOAS;或者小于词表词,如,在词表中有"别有用心"、"有用"、"用心",则可以在"别有用心"的内部发现 MOAS"有用心",使用全切分方法可以检测出所有的 MOAS,包括在词本身和词的内部检测到的。

我们利用了一部含 119487 个词条的词典(NjnuDic),对 1998 年 1 月人民日报语料(199801)

---

* 原文称为"最大交集型歧义切分字段",译为 MaximalCrossingAmbiguityString,MCAS[9]。限于篇幅,有关交集型歧义字段的相关定义,本文不再赘述。

进行了检测,比较了 FBMM和全切分两种方法的检测结果。

表 1　FBMM的检测错误类型

类型	FMM!=BMM且都错	FMM!=BMM且其一正确	FMM=BMM且都错	FMM=BMM且都正确
例子	正在以前所未有的力量拍击着中国的大门	如此刻骨铭心地进入她的内心深处	并发现有国际上尚未命名的新抗病基因	为广大人民群众服务
词表词	以前 前所未有 未有有的	如此 此刻 刻骨铭心	并发 发现 现有	广大 大人 人民群众
FMM	以前/所/未有/的(误)	如此/刻骨铭心/地(正)	并发/现有(误)	广大/人民群众(正)
BMM	以前/所/未/有的(误)	如此/刻骨/铭/心地(误)	并发/现有(误)	广大/人民群众(正)
FBMM	未有的	刻骨铭心地	NULL	NULL
MOAS	以前所未有的	如此刻骨铭心地	并发现有	广大人民群众
结果	错检	检测不完整	漏检	漏检

从表 2 中可以看到,使用全切分,可以检测到比 FBMM更多的 MOAS。另外,全切分检测到的 44880 个段例中,有 8043 个等于词表词,667 个小于词表词。可见,绝大多数的 MOAS是大于或等于词表词的,小于词表词的数量很少。

表 2　FBMM与全切分检测 MOAS的比较

检测方法	段例数(tokens)	段型数(types)
FBMM	21812	8337
全切分	44880	16992

我们使用全切分的检测方法,考察了不同词表对于检测结果的影响。仍以 199801 作为语料,使用了 6 部内容和规模各异的词典,分别为北大计算语言学研究所、北工大和中科院计算所的三部词典以及它们的合并版本(BBZDic)、199801 熟语料提取出来的词表和 NjnuDic。

表 3　词典规模对于 MOAS检测数量的影响

使用词典	词典规模(条)	MOAS段例(个)	MOAS段型(个)
199801熟语料提取	51286	33059	13069
北京工业大学	53735	34043	13534
中科院计算所	85602	37588	15159
北京大学计算语言学研究所	108750	44586	18113
NjnuDic	119487	44880	16992
BBZDic	128097	50559	20981
与词典规模的相关系数	——	0.972543	0.930326

从表 3 可以看到,随着词表规模的扩大,检测到的 MOAS的数量无论在段型还是段例上,都与词典规模基本成正比。通过上面的考察,我们认为,在较大规模的词表支持下,运用全切分的方法,能够较好地检测出真实文本中的 MOAS。

3　高频 MOAS 切分实例库的构建

3.1　MOAS的采集

如上所述,一个高质量 MOAS库的建立,首先需要一部规模大、质量高的词典。北大、北工大和中科院的词表通用性较好,质量较高,使用也较为广泛,我们采用了它们的合并版本(BBZDic)。采集语料为 1980 年至 2000 年(不含 1990 年)共 20 年的人民日报(RCorpus),共计 4 亿 2 千万字。使用全切分方法,检测出 MOAS段例 10844832 个,段型 782555 个,占 RCorpus总字数的 9.62%。MOAS的具体分布情况见表 4 和表 5。

表 4　MOAS的长度分布

长度	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	合计
段例数	5000087	4444736	858445	385365	89395	35131	21552	4527	3882	955	455	216	77	6	3	10844832
比例(%)	46.11	40.98	7.92	3.55	0.82	0.32	0.20	0.04	0.04	0.01	0.01	0.00	0.00	0.00	0.00	100
段型数	173185	339279	142185	86622	25351	10629	3360	1159	467	174	89	32	16	4	3	782555
比例(%)	22.13	43.36	18.17	11.07	3.24	1.36	0.43	0.15	0.06	0.02	0.01	0.00	0.00	0.00	0.00	100

表 4 显示,3 至 6 字长的 MOAS的段型数和段例数占绝对优势,与文献[8] 的统计结果相近。

表 5　MOAS的链长分布

链长	1	2	3	4	5	6	7	8	9	10	11	12	13	合计
段例数	6622317	3914448	193967	102199	7547	3778	366	176	16	6	9	1	2	10844832
比例(%)	61.06	36.10	1.79	0.94	0.07	0.03	0.01	0.00	0.00	0.00	0.00	0.00	0.00	100
段型数	262298	399841	72272	41091	4875	1879	195	81	10	6	5	1	1	782555
比例(%)	33.52	51.09	9.24	5.25	0.62	0.24	0.02	0.01	0.01	0.00	0.00	0.00	0.00	100

从表 5 可以看出,链长(链长是指 MOAS中,词语之间存在的交集关系的个数。如,"这时 - 时候",链长为 1)为 1 和 2 的 MOAS占绝对优势,接近孙茂松等[8] 的统计结果。

我们把 20 年语料按照月份顺序横向组合,平均分为 12 份(每份约 3300 万字),然后依次叠加起来,统计出随着语料规模的增大,MOAS段型和段例数量的增长情况,如图 1、2。

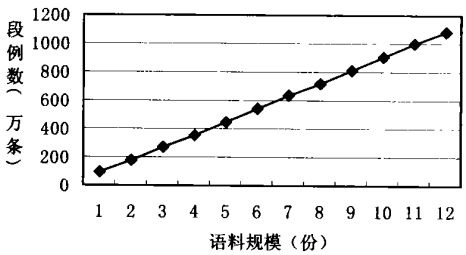

图 1　MOAS段例增量变化曲线

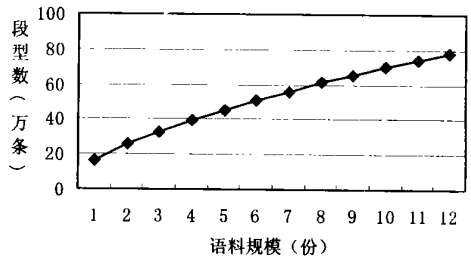

图 2　MOAS段型增量变化曲线

图 1、图 2 表明,随着语料规模的增大,交集型歧义字段的段例数量呈线性增长,段型数量的增长则呈曲线,增长相对趋缓。

3.2　切分实例库的构建

建立 MOAS切分实例库并进行人工判定和分析是本文的核心工作。我们从采集到的 MOAS中取出高频(频率不小于 100)段型 12653 个,作为建库的基本数据。再取出长度较大(6 -15 个字符)、频率相对较低(20-99)的段型 2253 个,作为补充。由此得到了一个规模为 14906 个段型(7685417 个段例)的高频 MOAS库,占 20 年语料中采集到的段例总数的 70.87%。

为得到这 14906 个段型的各种切分形式及相应比例,我们对每条频率不小于 100 的 MOAS随机抽取出 100 个带有上下文等相关信息的实例(对于频率为 20-99 的则全部提取),由此得到一个含 1354270 条实例的实例库,然后逐条进行人工判断。表 6 给出了部分样例。

表 6　MOAS切分实例库样例

上文	MOAS	下文	歧义构成	判定结果
缸头、缸套、活塞	等价值	20 多万元,	等价 - 价值	等/价值
防务安全问题也是	广大选民	普遍关心的问题。	广大 - 大选 - 选民	广大/选民
罗马尼亚	青年代表团	一百五十余人,	青年 - 年代 - 代表团	青年/代表团
本报讯	日本科学技术	基金会最近宣布,	日本 - 本科 - 科学技术	日本/科学技术

4 歧义字段数据分析

4.1 切分歧义统计

MOAS切分实例库的人工判定工作完成以后,得到了初步的统计结果。在 14906 个 MOAS 段型中,共有真歧义 224 个(1.50%),伪歧义 14682 个(98.50%)。表 7 给出了 14 个真歧义字段的统计数据。如,“个人中”,在 20 年语料中总共出现了 254 次,在抽样的 100 个例子中,切分为“个人/中”24 次,切分为“个/人/中”76 次。

表 7 人工判定得到的真歧义样例

MOAS	总频次	字段类型	第一种切分	freq1	第二种切分	freq2
个人中	254	非词	个人/中	24	个/人/中	76
后来到	256	非词	后/来到	56	后来/到	44
联合国内	148	非词	联合国/内	75	联合/国内	25
和平等	472	非词	和/平等	77	和平/等	23
干部会	233	词表词	干部会	88	干部/会	12
和美的	178	非词	和/美/的	92	和美/的	8
中青年	5238	词表词	中青年	93	中/青年	7
应用于	2031	非词	应用/于	97	应/用于	3
地面对	264	非词	地/面对	98	地面/对	2
好人才	183	非词	好/人才	99	好人/才	1
同学会	682	词表词	同学会	99	同学/会	1
赶上来	182	非词	赶/上来	99	赶上/来	1
赶上去	188	非词	赶/上去	99	赶上/去	1
开发区内	299	非词	开发区/内	99	开发/区内	1

4.2 真歧义字段的强势切分现象

通过观察,我们发现真歧义字段的多种切分形式之间的比例并不均衡。如表 7 所示,后 9 个真歧义字段的两种切分形式中,一种占到了抽样总例的 90% 以上。如,“应用于”,在其 100 个抽样总例中,有 97 例切为“应用/于”。这种比例悬殊的现象,我们称之为“强势切分”现象。即,“应用/于”是“应用于”的强势切分形式。在 224 个真歧义字段中,存在强势切分的多达 157 个,占 70%。详见表 8。

表 8 强势切分分布表

强势切分形式所占比例(%)	99	98	95-97	90-94	80-89	70-79	60-69	50-59	合计
MOAS数(条)	60	32	41	24	29	18	10	10	224
累计(条)	60	92	133	157	186	204	214	224	—
累计频率(%)	26.79	41.07	59.38	70.09	83.04	91.07	95.54	100	—

有了“强势切分”的概念以后,我们可以把真歧义和伪歧义看作一个自左至右的渐变过程,切分比例越均衡,越趋向于左端,其数量越少;切分比例越悬殊,越接近右端,其数量越大。由此,我们可以在 OAS 自动消歧的工作中,利用这一现象“分而治之”。由于绝大多数 MOAS都是强势切分的,对高频 MOAS可以采用基于记忆(查找强势切分形式)的方法进行自动消歧;对那些少量的、切分比例较为均衡的高频真歧义字段,则可以花费较大的代价去消歧;而对于数量庞大而频率低下的其他 MOAS,可以采用各种统计方法进行消歧。另外,“强势切分”现象可能造成真歧义字段的漏检。因为具有比 99:1 更为强势分布的真歧义字段,其数量也更大,随

机抽样时,很可能会漏掉处于劣势的切分形式。这说明,在 MOAS真歧义的采集工作中,语料的规模和领域性也很重要,在小规模或平衡性较差的语料中难以采集到大量的真歧义。即使是大规模语料,每个 MOAS也需要抽取相当数量的实例(如,100 条以上)才能基本确定究竟是真歧义还是伪歧义字段。

为了减少强势切分带来的影响和语料本身的局限,我们在人工判定切分形式的过程中同时采用了内省的方法。当然,纯粹内省出来的切分形式可能会比较刁钻,而且可能还是会漏掉一些存在的切分形式,但仍不失为一种重要的补充和参考。我们在高频 MOAS库里只有一种切分形式的段型中,找出具有潜在歧义的共 523 个。如,"内存在",在切分实例库中全部切为"内/存在",而人工给出其潜在的切分可能"内存/在",并给出相应的语境"最近/内存/在/涨价"。由此看来,单纯依靠观察抽样实例而得到的真歧义数量确实是有限的。

4.3 全切分方法的优化问题

本文第 2 部分曾经提到,全切分方法可以在词表词的内部检测到少量的 MOAS。在切分实例库中,共有 17244 个段例是在词表词内部检测到的,而它们的切分都应该服从于更大的词语单位。如,在"春风化雨"中检测出的 MOAS"春风化",应该服从整个词语的切分。因此,在使用全切分方法时,这类 MOAS是可以被忽略的。但是,与词表词相同的 MOAS是需要保留的。在切分实例库中与词表词相同的 MOAS有 222581 个段例。如,"中医学",在抽样的 100 个实例中,切为"中/医学"8 次,切为"中医/学"2 次,切为"中医学"90 次。可见,在使用全切分方法检测 MOAS的过程中,可以做一个简单的优化,只检测大于或等于词表词的字符串。

5 结语

本文通过对交集型歧义的大规模语料调查,得出了两方面的结论。首先,在 MOAS的检测问题上我们发现,常用的 FBMM除了会造成漏检之外,还会造成错检,而全切分能够检测出数量完整的且严格定义的 MOAS;后者检测出的 MOAS的数量与词典规模基本成正比。其次,使用全切分的方法,从大规模语料中采集了 14906 个高频 MOAS,并通过对切分实例库的分析,发现大多数真歧义字段具有强势切分的现象,从而提出了解决交集型歧义的"分而治之"的方法。下一步,我们将利用该数据库进行更加细致的统计分析和消歧实验,有关内容将另文探讨。

参 考 文 献:

[1] MuLi,JianfengGao,ChangningHuangetal.UnsupervisedTrainingforOverlappingAmbiguityResolutioninChinese WordSegmentation[A].In:ProceedingsoftheSecondSIGHANWorkshoponChineseLanguageProcessing[C].Sapporo,Japan,2003.

[2] 陈小荷. 现代汉语自动分析 ——VisualC++ 实现[M]. 北京:北京语言文化大学出版社,2000

[3] 陈小荷. 用基于词的二元模型消解交集型分词歧义[J]. 南京师范大学学报,2004, (6) :109-113.

[4] 戴新宇. 基于混合策略的机器翻译方法研究[D]. 南京大学博士论文,2004.

[5] 梁南元. 书面汉语自动分词系统 -CDWS[J]. 中文信息学报.1987,1 (2) :44-52.

[6] 刘挺. 歧义字段的最大概率切分算法[A]. 语言工程[C]. 北京:清华大学出版社,1997 :182-187.

[7] 刘开瑛. 中文文本自动分词和标注[M]. 北京:商务印书馆,2000.

[8] 孙茂松,左正平. 汉语真实文本中的交集型切分歧义[A]. 汉语计量与计算研究[C]. 香港:香港城市大学出版社,1998:323-338.

[9] 孙茂松,左正平,邹嘉彦. 高频最大交集型歧义切分字段在汉语自动分词中的作用[J]. 中文信息学报. 1999,13 (1) :27-34.