面向信息检索的排除词识别研究

章成志 苏新宁

(南京大学信息管理系 南京 210093)

【摘要】 针对信息检索中存在的词语排除关系问题,给出排除词的定义并说明排除词在信息检索中的作用。指出排除词实质上是最大准交集型歧义切分字段的伪歧义切分所导致的,描述排除词的识别方法,并给出识别的结果,并在实际的信息检索平台上对排除词词库进行应用测评。

【关键词】 信息检索 中文信息处理 交集型歧义 除词识别 伪歧义 【分类号】 TP391 G252

Recognition Mutually Exclusive Words for Information Retrieval

Zhang Chengzhi Su Xinning

(*Department of Information Management, Nanjing University, Nanjing 210093, China*)

【**Abstract**】 This paper introduces the phenomenon of mutually exclusive words and poses the role of mutually exclusive words identification in information retrieval system. The authors indicate that mutually exclusive words can be identified from the pseudo - ambiguity results after segmentation of maximal quasi overlapping ambiguity string, then present the method and result of the mutually exclusive words identification. Also the application of mutually exclusive words is provided.

【**Keywords**】 Information retrieval Chinese information processing Overlapping ambiguity Mutually exclusive words recognition Pseudo - ambiguity

1 引 言

中文自动分词是中文信息处理的基础研究内容之一,其中面临的两大难题便是切分歧义和未登录词问题。当前和今后一段时间里,对切分歧义的相关研究包括:基于 Web 和专业领域核心词表的分词歧义穷尽式调研、非受限的通用分词歧义表构造、各领域的常用分词歧义表构造等。其中,交集型歧义切分字段又占全部歧义切分字段的绝大多数[1]。对交集型歧义切分的研究主要集中于各种消歧算法的设计上,目前已有的方法主要是基于各种统计和规则进行的[2],如基于词概率[3]、词的 Bi - Gram 模型[4]、Bayes 分类器[5]等方法。同时,研究者还将消歧算法用于构建消歧实例库,在自动分词时调用切分实例或规则进行中文的分词[1,5]。这些方法对提高中文信息检索的质量起到了一定的提升作用。

本文将说明在信息检索中存在的一种特殊的准交集型歧义切分现象,即词语排除关系现象,给出了排除词的定义并论述了排除词在信息检索中的作用。

收稿日期:2006 - 11 - 20
收修改稿日期:2006 - 12 - 06

2 排除词及其在信息检索中的作用

2.1 相关概念

相关说明如下:

Ω:中文字符串集合;

S:中文字符串,$S = c_1 c_2, \cdots, c_n$,即 $S \in \Omega$;

Ψ:分词词典,用于中文分词;

T:已分词的训练语料。

下面给出与交集型切分歧义相关的基本定义。

定义 1:交集型歧义切分字段。对于字符串 S,$S = c_1 c_2, \cdots, c_n, S \in \Omega$,$S = c_1 c_2, \cdots, c_n$ 为汉字,如果存在整数 $i_1, i_2, \Lambda, i_m, j_1, j_2, \Lambda, j_m (m \geq 2)$,满足:

(1) $S \notin \Psi$;

(2) $w_1 = c_{i1} \Lambda c_{j1}, w_2 = c_{i2} \Lambda c_{j2}, \cdots, w_m = c_{im} \Lambda c_{jm}$,且 w_1、w_2、$\cdots$、$w_m \in \Psi$,且 S 中不存在包含 w_1、w_2、$\cdots$、w_m 的词;

(3) w_1、w_2、$\cdots$、w_m 构成相互交叉,即:

$1 = i_1 < i_2 \leq j_1 < j_2, i_2 < i_3 \leq j_2 < j_3, i_3 < i_4 \leq j_3 < j_4 \cdots,$

$i_{m-2} < i_{m-1} \leq j_{m-2} < j_{m-1}, i_{m-1} < i_m \leq j_{m-1} < j_m = n;$

则称字符串 S 为交集型歧义切分字段(Overlapping Ambiguity String, OAS)[1]。

例如,字符串"日本体育",其中"日本"、"本体"、"体育"

均为词,并构成交叉,因此"日本体育"是一个交集型歧义切分字段。

定义 2:伪交集型歧义切分形式。对于给定的交集型歧义切分字段 S,对于某一形式"$c_1/\cdots/ w_i/\cdots/c_n$",其在 T 上的出现概率,即:$p(c_1/\cdots/ w_i/\cdots/c_n|T)$,低于给定阈值 θ,则称该切分形式,即:"$c_1/\cdots/ w_i/\cdots/c_n$"为 S 在 T 上的伪交集型歧义切分形式,其中 w_i 为伪交集型歧义切分形式下的词语之一。

例如,"市政府"虽然存在"市/政府"和"市政/府"两种不同的切分形式,但在训练语料中,切分形式为"市/政府",而 $p(市政/府|T)=0<\theta$,因此,"市/政府"为"市政府"在 T 上的伪交集型歧义切分形式。识别出伪交集型歧义切分形式后,分词系统就可以采用查表的方式直接确定分词形式,而不再参与后续的分词处理过程[6],即歧义消解通过直接查表即可实现[7]。

定义 3:准交集型歧义切分字段。对于交集型歧义切分字段 S,若 S 为复合词,且 S 可加入到词表 Ψ,此时,S 将不再是严格定义下的交集型歧义切分字段。则将 S 称为准交集型歧义切分字段(Quasi Overlapping Ambiguity String,QOAS)。

例如,"人民法院"开始不在词表 Ψ 中,其中"人民"、"民法"、"法院"均为词,并构成交叉,因此"人民法院"是一个交集型歧义切分字段。但"人民法院"为复合词,可将"人民法院"加入词表 Ψ 中,此时,"人民法院"不再是交集型歧义切分字段,而是准交集型歧义切分字段。显然,若将"人民法院"从词表 Ψ 中删除,则"人民法院"又成为交集型歧义切分字段。

定义 4:排除词。对于给定的准交集型歧义切分字段 S,存在伪交集型歧义切分形式,即 $c_1/\cdots/ w_i/\cdots/c_n$,则称 w_i 与 S 构成排除关系,即 w_i 与 S 互称排除词(Mutually Exclusive Words,MEW)。例如,"电脑科学"为准交集型歧义切分字段,存在伪交集型歧义切分形式"电/脑科/学",因此,"脑科"与"电脑科学"构成排除关系,即"脑科"为"电脑科学"的排除词。同理,"民法"为"人民法院"的排除词。

2.2 排除词在信息检索中的作用

(1) 全文索引时降低索引膨胀率

当全文检索系统后台采用的是词索引时,通常情况下,系统会对词语所有出现的位置进行索引,例如,对"动机"建立词索引时,会将"电动机"、"永动机"等关键词的文档列入到索引中,这样一来,当用户检索"动机"时,返回结果中自然会包含"电动机"、"永动机"等不相关的文档。若事先建立排除词词典,在对排除词建立索引时,考虑类似"动机"与"电动机、永动机"等排除关系,即在该排除词词典的干预下,可以过滤这种语义不相符的索引,降低了索引膨胀率。

另外,在进行词索引时,若进行自动分词,借助于排除词词典,可加快分词质量和速度,这是基于记忆的伪交集型歧义切分处理过程[7]。

(2) 信息检索中的缩检作用

利用排除词词典可以排除与查询式不相关的结果,如检索"脑科",排除包含"电脑科学"的结果,检索"个性化学习",排除包含"化学"的结果,这样起到缩检作用,并提高检准率。

(3) 信息检索中的扩检作用

当信息检索系统返回给用户的结果过少时,有些系统会对查询式进行解析,重新进行检索。例如,用户将"基本体操"作为检索词进行文档检索时返回结果很少,若用户希望得到更多相关结果,此时,一般的全文检索系统会对检索词进行字面折分,如将"基本体操"拆分为:"基本、本体、体操"作为新的查询式重新进行查询,返回包含有"基本"、"本体"或"体操"的文档。实际上,含有"本体"的文档和用户的实际需求是不符合的。在排除词词典的控制下,则可避免这种语义不相符的情况发生。

3 排除词识别方法

3.1 最大准交集型歧义切分字段识别

前面提到,只有当 S 被认定为复合词时,交集型歧义切分字段 S 与 w_i 才有可能构成排除关系。在没有上下文约束的情况下,复合词 S 的正确切分形式不包括切分为"$c_1/\cdots/ w_i/\cdots/c_n$"的情形,例如,"发展中国家"可以切分为"发展中/国家","发展中"对"国家"起限定作用,但不可切分为"发展/中国/家"。在有上下文约束的情况下,复合词 S 有可能与周围的字符再次发生新的交叉关系。例如,句子"这反映了我国的 R&D 活动在发展中国家居中等水平",其中,"发展中国家"与后续字符"居"、"中"、"等"再次交叉,构成交集型歧义,即"发展中国家居中等"也是交集型歧义切分字段,其正确切分形式为"发展中国家/居/中等"。而在句子"致力于把个性化高品质的设计和一流的工业体系配套有效结合来发展中国家居产业。"中,"发展中国家居"为交集型歧义切分字段,其正确切分形式为"发展/中国/家居"。

由此可以看出,在有上下文约束下,即真实文本中,只有当包含准交集型歧义切分字段 S 的句子能切分出复合词 S 时,才能进一步进行排除词的识别。若该准交集型歧义切分字段与周围字符发生交叉,即成为交集型歧义切分字段,则需要进行歧义消解处理。若该准交集型歧义切分字段 S 不再与任何字发生新的交叉关系,则称该准交集型歧义切分字段为最大准交集型歧义切分字段,笔者借用最大交集型歧义切分字段的定义(Maximal overlapping ambiguity string,MOAS)[1],给出准交集型歧义切分字段的具体定义如下。

定义 5:最大准交集型歧义切分字段。设 $S=c_1c_2\wedge c_n$ 为

准交集型歧义切分字段,满足:

(1)$S_{max}=c_i \wedge c_j (1\le i<j\le n)$,且 $S_{max}\in S$;

(2)S_{max}为交集型歧义切分字段;

(3)S 中不存在包含 S_{max}的更大的交集型歧义切分字段;

则称 S_{max}为 S 的最大准交集型歧义切分字段(Maximal Quasi Overlapping Ambiguity Atring,MQOAS)。

例如,在句子"论法国会计模式对中国会计制度改革的借鉴意义"中,"法国会计"为准交集型歧义切分字段,"法国会"为交集型歧义切分字段,但"法国会计"涵盖了"法国会",同时不为任何交集型歧义切分字段所包含,因此,"法国会计"是最大准交集型歧义切分字段。

识别最大准交集型歧义切分字段的意义在于:由于最大准交集型歧义切分字段不再与周围任何字符发生新的交叉关系,具有一定的独立性,因此,可以将它们从上下文环境中分离出来[1],直接作为排除规则,加入到排除词词典中,从而实现前面所提到的功能。

3.2 排除词识别方法

根据 MQOAS 的来源不同,将 MEW 识别分为两类,即基于关键词词典的 MEW 识别方法与基于语料库的 MEW 识别方法,下面详细描述这两种识别方法。

(1) 基于关键词词典的 MEW 识别方法

该方法以关键词词典为基础,具体描述见例 1。首先,依据语料库从关键词词典中识别出所有 MQOAS,即识别出的 MQOAS 全部为关键词词典中的词语,在这里,笔者利用全切分的方法进行 MQOAS 的识别[2];然后,对每个 MQOAS 在分词训练语料 T 上进行检索,若 T 上存在该 MQOAS,则进行切分形式统计,对于其中某一切分形式"$c_1/\cdots/w_i/\cdots/c_n$",其在 T 上的出现概率,即 $p(c_1/\cdots/w_i/\cdots/c_n|T)$,低于给定阈值 θ,则认为 w_i 与该 MQOAS 构成排除词关系,并保存该结果;若在 T 上不存在该 MQOAS,则进行手工切分,人工判别得到排除词,并保存结果。

(2) 基于语料库的 MEW 识别方法

该方法以语料库为基础,具体描述见例 2。首先依据分词词典,在中文语料上利用全切分方法识别出所有最大交集形歧义 MOAS;然后判别每个 MOAS 当前是否为复合词,在这里,复合词的判别可以人工判别或采用统计方法进行辅助判别;若为复合词,则在分词训练语料 T 上检索该 MOAS,若 T 上存在该 MOAS,则进行切分形式统计,对于其中某一切分形式"$c_1/\cdots/w_i/\cdots/c_n$",其在 T 上的出现概率,即 $p(c_1/\cdots/w_i/\cdots/c_n|T)$,低于给定阈值 θ(本文称改阈值为伪歧义判别阈值),则认为 w_i 与该 MOAS 构成排除词关系,并保存该结果;若在 T 上不存在该 MOAS,则进行手工切分,人工判别得到排除词,并保存结果。

例 1:基于关键词词典的 MEW 识别方法

输入:关键词词典,分词词典 Ψ,分词训练语料 T,伪歧义判别阈值 θ

输出:MEW

步骤:

初始化:关键词词典,Ψ,θ

利用全切分方法识别出关键词词典中的所有 MQOAS

For each MQOAS

If MQOAS $\in$ T then

For each i

在 T 上进行切分形式统计,对于某种切分形式"$c_1/\cdots/wi/\cdots/c_n$"

If $p(c_1/\cdots/w_i/\cdots/c_n|T)<\theta$ then

保存 w_i 与当前 MQOAS

If MQOAS $\notin$ T then

对 MQOAS 进行手工切分,人工判别得到排除词,并保存结果

返回 MEW 识别结果

例 2:基于关键词词典的 NEW 识别方法

输入:分词词典 Ψ,中文语料,分词训练语料 T,伪歧义判别阈值 θ

输出:MEW

步骤:

初始化:Ψ,θ

利用全切分方法识别出中文语料上的所有 MOAS

For each MOAS

If MOAS 为复合词 and MOAS $\in$ T then

For each i

在 T 上进行切分形式统计,对于某种切分形式"$c_1/\cdots/w_i/\cdots/c_n$"

If $p(c_1/\cdots/w_i/\cdots/c_n|T)<\theta$ then

保存 w_i 与当前 MOAS

If MOAS 为复合词 and MOAS $\notin$ T then

对 MOAS 进行手工切分,人工判别得到排除词,并保存结果

返回 MEW 识别结果

4 排除词识别结果与应用

4.1 排除词识别结果

本文所用的关键词词典为 CSSCI 关键词词库,共包括 489 144 条关键词。分词词典有两种,一个是人民日报分词词典[8],约 10 万词,另一个是 Nju 分词词典,约 14 万词。本文根据分词训练语料的规模,设置伪歧义判别阈值的经验值为 $\theta=0.05$。

(1) 基于关键词词典的 MEW 识别结果

基于词典的方法是采用 CSSCI 关键词词典,分词训练语料为人民日报 1998 年 1~6 月标注语料。将识别的结果分为三字词、四字词、五字词及其他等 4 种情形进行统计分析,结果如表 1 所示。从 CSSCI 关键词词典中识别出 MQOAS 的数量为 58228 条,识别出 MEW 共 16964 条,MEW 占 MQOAS 的比率为 29.13%。随着词长的变化,MEW 占 MQOAS 的比率

阈在30%附近上下浮动。MQOAS、MEW 的识别结果分布情况如图 1 所示。可以看出,MEW 主要集中在四字词和五字词中出现。

表 1 基于 CSSCI 关键词词典的 MEW 识别结果

项目 \ 词类别	三字词	四字词	五字词	其他
关键词数量(个)	36 064	191 125	71 635	190 320
MQOAS 数量(个)	1 581	13 507	31 288	11 852
MEW 数量(个)	402	5 582	7 724	3 256
MEW/MQOAS(%)	25.42	41.33	24.69	27.47
例子	都柏林:柏林 地理学:理学 峨眉山:眉山 美国学:国学 蒙古文:古文	法伦理学:理学 科学学习:科学学 企业主页:企业主 移动机制:动机 移民生活:民生	差异化教育:异化 城市区域化:市区 创造性能力:性能 高原生态学:原生态 个性心理学:理学	道德教化学说:化学 人民法制建设:民法 西方进化学说:化学 中华人民共和国:华人 最高人民法院:民法

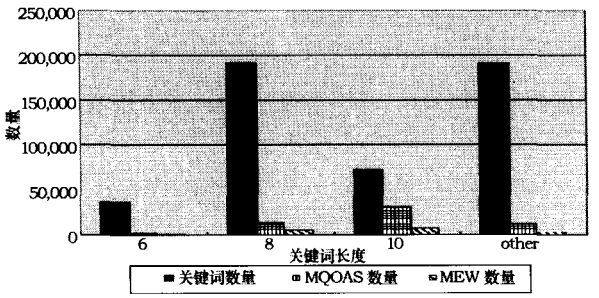

图 1 MEW 识别结果分布图

(2) 基于语料库的 MEW 识别结果

基于语料库的方法中,分别对人民日报 2000 年数据(约 1200 万汉字)与 1995 - 2001 年数据(约 7.2 亿汉字)做了测试,在识别 MEW 过程中,识别出的 MAOS 情况如表 2 所示。

表 2 人民日报 MAOS 识别结果

分词词典类别 \ 语料类别	人民日报 2000 年数据		人民日报 1995 - 2001 年数据	
	MAOS 总次数	MAOS 去重后	MAOS 总次数	MAOS 去重后
人民日报切分词典	622 656	110 507	3 453 946	341 987
Nju 切分词典	266 246	57 997	1 419 921	180 582

由表 1 可以看出,词典规模、语料库规模对识别 MAOS 都产生影响。Nju 切分词典规模大于人民日报切分词典规模,识别出来的 MAOS 是前者少于后者。随着语料库规模的增大,MAOS 的数量也在增加。由于识别出的 MAOS 规模较大,因此笔者目前只对 MAOS 的高频部分,即高频最大交集型歧义字段(High Frequent Maximal Overlapping Ambiguity String, HF - MOAS)进行识别。本文设置的频率阈值为 50,即当 HF - MOAS 在中文语料中的频率超过 50 时,才进行后续的 MEW 识别。表 3 给出了排除词识别结果样例。

表 3 排除词识别结果样例

排除词	对应排除词
本体	版本体系、话本体制、基本体操、基本体制、剧本体裁、刊本体制、日本体育、资本体系
动机	电动机、发动机、拉动机制、联动机制、流动机制、驱动机制、运动机制
法人	司法人才、司法人员、执法人才、执法人员
化学	动物演化学、个性化学习、气化学说、强化学习、数字化学习、文化学习
民法	公民法、难民法、人民法庭、人民法院、人民法治、移民法
日文	假日文化、节日文化、抗日文学、中日文化、宗日文化
文选	古文选本、古文选评、课文选材、论文选题、作文选材、征文选题

4.2 排除词词库的应用

笔者针对经济日报 1983 - 2003 年语料进行了缩检测试。测试方法为:从排除词词库中随机选取 100 条排除词记录,将排除词作为检索词,如“本体”,进行全文检索,考察缩检效果。测试结果为:缩检前,这 100 个查询词返回的文档总数为 425 350 条,平均返回文档数为 425,但缩建后,返回文档总数变为 28 753,平均返回文档数变为 29,参见图 2。很明显,排除词起到了明显的缩检作用,提高了检准率。

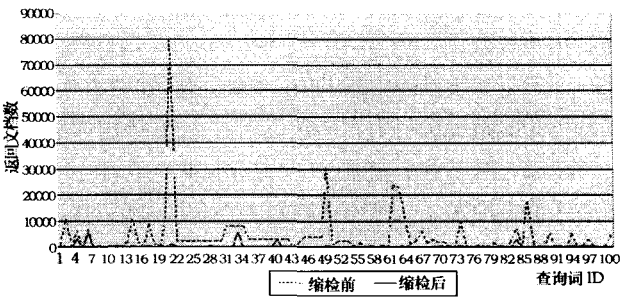

图 2 查询词缩检前后返回文档数

5 结 语

排除词实质上是最大准交集型歧义切分字段的伪歧义切分所导致的,文本给出排除词识别的两种方法,即基于关键词词典的识别方法与基于语料库的识别方法。利用这两种方法构建了排除词词库,并在实际的信息检索平台上进行了排除词词库的应用测评,结果表明,排除词起到了明显的缩检作用,提高了检准率。

进一步的工作主要包括:

(1)进一步提高排除词识别的自动化程度,例如,当训练语料中不包含 MQOAS 或 MOAS 时,如何自动获取 MQOAS 或 MOAS 的伪歧义切分形式,如何完全自动化地判断 MOAS 是否为复合词等。

(2)排除词词典的进一步推广应用,例如,在信息检索中

利用排除词词典,对查询式进行记忆式的解析。

参考文献:

1 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分词中的作用. 中文信息学报,1999,13(1):27 - 34

2 李斌,陈小荷,方芳等. 基于语料库的高频最大交集型歧义字段考察. 中文信息学报,2006,20(1):1 - 6

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

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

5 Mu Li, Jianfeng Gao, Changning Huang et al. Unsupervised Training for Overlapping Ambiguity Resolution in Chinese Word Segmentation. In: Proceedings of the Second SIGHAN Workshop on Chinese Language Processing. Sapporo, Japan, 2003. 1 - 7

6 赵铁军,吕雅娟,于浩等. 提高汉语自动分词精度的多步处理策略. 中文信息学报,2001,15(1):13 - 18

7 孙茂松,邹嘉彦. 汉语自动分词研究评述. 当代语言学,2001,3(1):22 - 32

8 Lexicon_tull_2000. mdb. http://ccl. pku. edu. cn/doubtfire/Course/Chinese% 20Information% 20Processing/Source_Code/Chapter_8/Lexicon_full_2000. zip(Accessed Apr. 20,2006)

(作者 E - mail:zcz51@citiz. net)

动态

弗吉尼亚大学图书馆成为 Google 馆藏图书数字化项目新成员

弗吉尼亚大学图书馆与 Google 合作,将该图书馆馆藏中的一部分进行数字化,并加入到"Google 图书搜索"(Google Book Search)这一产品中。这是继哈佛大学图书馆、牛津大学图书馆、纽约公共图书馆、密歇根大学图书馆等图书馆之后的第 9 个被邀请加入 Google 这一计划的研究型图书馆,通过与 Google 及这些机构合作,使图书馆的海量资源更容易地被用户获取。

弗吉尼亚大学图书馆共有 13 个分馆,拥有 500 万卷、共计 1700 万件手稿、珍善本以及数量迅速增加的数字馆藏。其特色馆藏是美国历史和文化方面的文献,该馆在数字技术研发和创新方面的能力也较为突出。

Google 图书数字化项目(http://books. google. com)将全球主要的图书馆的馆藏图书进行数字化,使其通过 Google 图书搜索被检索。从参与该项目的大学图书馆的馆藏中选择出数十万种图书,由 Google 进行数字化之后,使这些图书能够被用户检索得到。弗吉尼亚大学图书馆进行数字化的图书主要是馆藏中精选的关于历史、人文方面的著作,并加入到图书检索项目中。

图书检索项目的设计遵照版权法的要求,任何人将能够自由地查看、浏览、阅读公共领域的图书。对于那些受版权保护的不能够进行数字化的图书,读者可以通过关键检索找到相关的图书,并得到关于这本书的基本信息(如书名、作者等)、相关的检索信息、可以购买或借阅的信息等。如果出版商或作者不想使其图书数字化,可以提出来,在进行数字化项目的时候,将其排除在外。

"自 1992 年起,弗吉尼亚大学图书馆已将公共领域的著作在线自由获取",该馆馆员 Karin Wittenborg 说,"研究人员说图书馆的工作使得他们发现并提出新问题成为可能,因为这些著作可以以数字形式联机获取。通过 Google,我们将能够提供更多文献的获取。例如:以前很难被多数读者获取的 18 - 19 世纪的著作能够被新的发现和获得。"

(编译自:Google Books Library Project: University of Virginia joins leading research libraries in partnership with Google to increase discovery of knowledge and to offer library books to global audience. [2006 - 11 - 14]. http://www. lib. virginia. edu/pressroom/uvagoogle)

(本刊讯)