面向特定领域的多字词表达式的提取

刘荣[1,2] 王丽娟[3] 张志平[1] 刘健文[4] 胡竟伟[5]

[1]北京语言大学应用语言学研究所 北京 100083

[2]太原理工大学文法学院外语系 太原 030012

[3]太原理工大学计算机与软件学院计算机科学技术系 太原 030012

[4]北京市玉渊潭中学 北京 100038

[5]内蒙古河套大学数学与计算机科学系 临河 015000

E-mail: liurong@blcu.edu.cn

摘要： 本文在阐述了多字词表达式的定义的基础上，对面向特定领域的多字词表达式提取的技术路线进行了详细说明，并在方法、面向对象规模、效率等方面有所创新。利用领域高频词的新技术提取了多字词表达式，其面向的对象是数量为 G 级大规模的真实文本，且计算时间复杂度是线性的。对所提取的结果进行人工判断后表明，效果较为理想。

关键词： 多字词表达式 特定领域 高频词

The extraction of multiword expression in special field

Liu Rong, Wang lijuan, Zhang zhiping, Liu jianwen, Hu jingwei

Applied Linguistics Institute, Beijing Language University Beijing 100083

Foreign Languages Department, Art and Law College, Taiyuan University of Technology, Taiyuan 030012

Computer Science Department, Computer and Software College, Taiyuan University of Technology Taiyuan 030012

Yu yuantan Middle school Beijing 100038

Mathematics and Computer Science Departement, Hetao University of Inner Mongolia, Linhe, 015000

E-mail:Liurong@blcu.edu.cn

Abstract: The definition of multiword expression is fixed in the paper. The technology of extracting multiword expression in special field is discussed in detail, which appeals to high frequency word. The texts processed in the paper amount to G level. The computation of time for large corpora is linear. The result is preferable by manual judgment.

Keywords: multiword expression, specific field, high frequency words

1. 引言

对于大多数自然语言处理工程而言，词汇资源是非常重要的。在自然语言理解和自然语言处理的许多实际应用中，短语或语块有着很重要的作用。本文所提取的多字词表达式有广泛的应用前景，它对于词典编纂、中文词语的歧义消解、提高中文文本自动分类的准确率、提高搜索引擎的效率、中文信息处理的浅层句法分析、自动文摘、信息抽取、对外汉语教学的教材更新，机器翻译等方面都会有所帮助。尤其在词典编纂领域，由于大部分词汇资源需要手工编纂，这就需要耗费大量的人力物力财力，受限于编者所用资料的规模和取舍态度，一部词典总是难免出现这样那样的不可规避的问题， 另外随着现代信息社会的发展，很多新词新义项难以很快收录。因此，尽可能的面向大规模真实文本，利用机器自动生成词汇的方法是一个当前研究的热点和难点。

2. 关于多字词表达式（multiword expression）

2.1 多字词表达式的广义定义

多字词表达式的难点是它的定义和自动识别。这两个问题互相关联，就产生了一个循环问题。一旦我们不能完全体现和清楚描述出多字词表达式的特性，那么对多字词表达式的定义就不全面，进而言之，如果定义不全面，那么多字词表达式的提取就不充分，不能显现其所有的特性。在自然语言处理领域和语言学界，有相当多的研究都借鉴了 Sag et al(2002)和 Wray(2002)的广义的定义[1]：Sag et al(2002)把多字词表达式粗略的定义为“跨越词的边界或空格的特质的解释。” Wray(2002)则从心理语言学的角度给出了定义：一个连续或非连续的词语序列。它是预先构成的，即存储在记忆中或在使用时能从记忆中整体检索到，不受到语法分析限制或不同年龄人群限制的多字词序列。

2.2 特定领域的多字词表达式和术语的区别与联系

由于本文的多字词表达式是基于特定领域的，所以有必要对这种多字词表达式与术语的区别与联系做一个说明。

术语这个概念是相对于一般词语而言的。他们的具体区别如下：术语(Term)是在一个学科领域中使用，表示该学科领域内概念或关系的词语。术语可以是词，也可以是短语。术语可以只在一个学科领域中存在，也可并存于多个学科领域中。[2]

而一般词语（Common Words）是指在一个学科领域中除了术语之外的词语。

本文所定义的特定领域的多字词表达式不仅只有术语，也包括一般词语。例如教育领域中的“投档线”、“小语种”这样的术语，也有“爸爸妈妈”、“工作人员”这样的一般词语。

2.3 多字词表达式和短语或语块的联系

在语言学层面，短语是意义上和语法上能搭配而没有句调的一组词，所以又叫词组。它是大于词而又不成句的语法单位。[3]短语可被分为固定短语和自由短语。固定短语包括命名实体、成语、惯用语、歇后语、缩略语，除此之外就是自由短语。在信息处理层面，短语又可被称作组块（chunk）、语义块、短语结构等。短语的内部结构比较稳定，通常作为一个独立的整体和句子中的其他成分发生作用。

国外的 Abney 最早提出了完整的组块描述体系，他把组块定义为句内的一个非递归的核心成分。这种成分包含核心成分的前置修饰成分，而不包含后置附属结构。中科院自动化所的赵军[4]则根据汉语的特点突破了非嵌套的束缚，把汉语基本名词短语（BaseNP）定义为：

BaseNP→ BaseNP+ BaseNP

BaseNP→ BaseNP+名词| 名动词

BaseNP→限定性定语+ BaseNP

BaseNP→限定性定语+名词|名动词

限定性定语→形容词|区别词|动词|名词|处所词|西文字串|（数词+量词）

而我们认为，多字词表达式与短语或语块有相似的地方，但更关注内部的词汇语义组合关系。其句法形式主要包括：修饰、并列、述宾、述补关系；其语义形式主要包括：属性-主体、

实体-实体、动作-补充关系等。多字词表达式是多个词语按照特定结构聚合形成的信息单元。因此，多字词表达式内部词与词之间的关联性，或者说词与词之间的互信息非常重要，而自由短语内部词与词之间的关联性就不那么强。

2.4 多字词表达式的定义

本文所研究讨论的多字词表达式是指内部结合紧密、使用稳定，整体表示一个概念意义，基本可以当作一个固定短语来使用的信息单位。一般而言，固定短语包括熟语和专名。熟语包括成语、谚语、歇后语等，专名在信息抽取领域又叫命名实体。本文在“结合紧密，使用稳定”的原则上，通过“音节适长”（吕叔湘：词的长度是有音节数目限制的，词大不过三）的考虑，确定了提取的多字词表达式最大是五个分词单位。

3. 多字词表达式提取技术路线

3.1 基本思想：

按照“结合紧密，使用稳定”的标准，通过在不同领域内考察字词的同现和搭配，把领域内高频同现的字符串当作“结合紧密，使用稳定”的多字词表达式。结合紧密在信息处理领域的一个表现是字词之间的互信息大不大。通过互信息的度量，我们可以确定字词之间的结合紧密程度。

3.2 考虑因素

在确定技术路线时，我们主要考虑到了以下几个因素：面向大规模真实文本，基于空间和时间计算复杂度，音节适长。

3.2.1 面向大规模真实文本

本文所用语料来源于北京语言大学 DCC 语料库。DCC 语料库是从 2001 年开始建立的一个“动态流通语料库”。跟据媒体的流通度我们选择了 15 种主流报纸媒体资源进行建库。

目前，DCC 语料库已经入库了从 2001 至 2007 年的约二十多亿字的语料。动态收集了十五种报纸的网页、txt 文本，并进行了初步加工，既标注了文本的文本外信息（媒体名称、文本发表日期、文本大小等），又标注了文本内信息（文本的标题、段落、作者、日期等），而且将进一步以动态流通的标准加入新的报纸和新的媒体语料，并进行动态加工。

本文所用语料是 DCC 语料库 2007 年语料，规模是 5 亿多字。实验所提取出的候选字符串是 1.77G，总计 7083449 行记录。

3.2.2 基于空间和时间复杂度的考虑：

传统的 N-gram 方法（哈希表，稀疏矩阵）对 3 元以上的计算在空间和时间复杂性上要求很高，当 n 取 5 时，就很难完成。例如，在作 3 元计算时，假设计算对象是平均长度为 2（即 4 个字节）的 50000 个中文字符，我们用 16 位 2 字节存储，3 元计算过程所需空间是（5×10^4）$^3\times$（4+2）=7.5×10^{14}bytes≈700000Gb。即使实际语料库对存储空间要求没有这么大，对现有的硬件而言，时间和空间也是相当大的[5]。面对 DCC 语料库 2007 年这么大的一个语料量，显然 3 元以上的计算用普通 PC 机是难以完成的。基于这样的考虑，我们不能采用传统的 n 元算法而是

采用互信息的方法去考察字词间的同现和搭配。

3．2．3 音节适长的考虑：

吕叔湘说过：“词的长度是有音节数目限制的，词大不过三”。另外，黄昌宁在《基于字的分词方法的实验研究》[6]一文中做的是基于 5 个字的实验，周强等也是最大开到 5 个字词的窗口进行的短语实验。综合语言学规律及前人的实施办法，本文确定了提取最大五个分词单位的多字词表达式。

3．3 技术路线：

罗盛芬，孙茂松[7]分别考察了九种常用统计量在汉语自动抽词中的表现,对二字词的自动抽词实验结果表明,这九种常用统计量中,互信息的抽词能力最强,F-measure 可达 54.77%,而组合后的 F-measure 为 55.47%,仅比互信息提高了 0.70%,效果并不显著.他们的结论是:(1)各种统计量并不具备良好的互补性;(2)通常情况下,建议直接选用互信息进行自动抽词,简单有效.

基于他们的结论，本文采用互信息方法。

互信息是用来度量一个消息中两个信号之间的相互依赖程度。二元互信息是两个事件的概率函数：

$$MI(X,Y)=\log 2\ \frac{P(X,Y)}{P(X)P(Y)}$$

互信息值越高，X 和 Y 组成短语的可能性越大，反之，互信息值越低，X 和 Y 之间存在短语边界的可能性越大。

但是确定句子中多字词表达式的边界，不能只局限于两个字词组合的互信息，即 bigram。而是要考虑上下文信息，即要考虑多个字词间的互信息，把 bigram 扩展为 n-gram。Magerman&Marcus 提出的广义互信息（generalized mutual information）概念，根据两个相邻的词类标记的上下文来决定它们之间是否是一个短语边界。借鉴其思想，我们通过词性序列的左右界来确定多字词表达式的边界。本文所用的语料库中字、词的数量和种类很庞大，但是词性的种类和数量相对而言就少得多，在线形时间内是可计算的。我们对于每个词性序列计算互信息。（例如词性序列/v/c/n/d 的互信息值远小于词性序列/v/NUM/n/n 的互信息值。)然后对互信息值高的词性序列选取其中对应的高频字词组合。

4. 实验方案

4．1 特定领域文本的提取

本文目前考察的是 2007 年基于 DCC 语料库的四类文本。使用 DCC 张志平博士的文本分类软件提取。其中：

教育类文本大小：154M，文本数：46868 个

经济类文本：91577 个文件

体育类文本：84701 个文件

娱乐类文本：63820 个文件

4．2 教育类和其它类共有的词中高频词的确定

首先用分词软件（自动化所，赵军）进行分词、词频统计。使用领域相交的方法提取出教育类和其它类共有的词，并在教育类词表中按频次排序。具体做法是把已分好的教育类文本进行词种数和频次统计。同理，再把其它类文本的词种数和频次进行统计。然后再把教育类文本和经济类、体育类、娱乐类文本进行交集运算，得到教育类文本和其它类文本共有的词语，按频次排序。本文所取的高频词是频次排在前 5000 位的字词。

4．3 由领域高频词提取字符串

确定领域种子词：将教育类和其它类共有的前 5000 高频词作为领域种子词。以领域种子词为中心左右各开最大两个窗口，分别统计各个窗口的频次。例如种子词 A 可能产生的各个窗口是：A*，A**, *A，*A*, *A**, **A，**A*，**A** .开窗口运算完成后，合并相同的字符串，并统计频次。实验结果按词性序列放入数据库，本文只举出三个词性序列的实验结果。

4．4 实验结果

ID	word	cixing	数据	ID	word	cixing	数据	ID	word	cixing	数据
1	多/人	/a/n	2688	1	电子/信箱	/n/n	7080	1	告诉/记者	/v/n	9300
2	和谐/社会	/a/n	2648	2	责任/编辑	/n/n	6630	2	让/孩子	/v/n	7060
3	普通/高中	/a/n	2516	3	传统/文化	/n/n	3420	3	招生/计划	/v/n	5784
4	困难/学生	/a/n	1930	4	人力/资源	/n/n	3198	4	用人/单位	/v/n	5246
5	新/课	/a/n	1858	5	职业/学校	/n/n	3104	5	让/学生	/v/n	4388
6	新/学期	/a/n	1738	6	家庭/经济	/n/n	2878	6	让/人	/v/n	4374
7	优秀/学生	/a/n	1712	7	思想/政治	/n/n	2592	7	工作/人员	/v/n	4018
8	新/课程	/a/n	1672	8	大学/毕业生	/n/n	2464	8	填报/志愿	/v/n	3924
9	旧/版	/a/n	1598	9	特色/社会主义	/n/n	2406	9	求学/指南	/v/n	3886
10	高/水平	/a/n	1506	10	才/能	/n/n	2380	10	有关/负责人	/v/n	3746
11	全/社会	/a/n	1418	11	专业/技术	/n/n	2378	11	培训/机构	/v/n	3742
12	优秀/教师	/a/n	1416	12	职业/资格	/n/n	2338	12	咨询/电话	/v/n	3176
13	独立/学院	/a/n	1246	13	硕士/学位	/n/n	2308	13	综合/素质	/v/n	3056
14	热门/专业	/a/n	1240	14	彩图/版	/n/n	2300	14	考试/成绩	/v/n	2972
15	小/语种	/a/n	1230	15	硕士/研究生	/n/n	1990	15	录取/通知书	/v/n	2940
16	贫困/学生	/a/n	1164	16	博士/学位	/n/n	1964	16	工作/经验	/v/n	2920
17	长/时间	/a/n	1142	17	资格/证书	/n/n	1950	17	高考/成绩	/v/n	2756
18	和谐/文化	/a/n	1076	18	学生/家长	/n/n	1884	18	教育/部门	/v/n	2748
19	高/技能	/a/n	1022	19	校园/文化	/n/n	1808	19	有关/部门	/v/n	2578
20	大/程度	/a/n	974	20	高中/毕业生	/n/n	1774	20	学习/成绩	/v/n	2550

ID	word	cixing	数据	ID	word	cixing	数据	ID	word	cixing	数据
21	高/分	/a/n	910	21	价值/体系	/n/n	1740	21	教学/质量	/v/n	2534
22	实际/情况	/a/n	908	22	类/专业	/n/n	1698	22	有/机会	/v/n	2464
23	好/成绩	/a/n	868	23	核心/价值	/n/n	1682	23	教育/资源	/v/n	2304
24	贫困/家庭	/a/n	840	24	职业/技能	/n/n	1600	24	助学/贷款	/v/n	2116
25	优秀/人才	/a/n	830	25	基础/知识	/n/n	1594	25	教育/机构	/v/n	2070
26	热/招	/a/n	808	26	成语/故事	/n/n	1592	26	看/书	/v/n	1932
27	高/质量	/a/n	778	27	社会主义/核心	/n/n	1560	27	是/学生	/v/n	1918
28	高/科技	/a/n	770	28	电子/商务	/n/n	1530	28	采访/时	/v/n	1886
29	贫困/大学生	/a/n	760	29	学士/学位	/n/n	1482	29	培养/学生	/v/n	1882
30	高/学历	/a/n	728	30	专业/知识	/n/n	1480	30	是/学校	/v/n	1784
31	多/月	/a/n	718	31	高中/阶段	/n/n	1462	31	鉴赏/全书	/v/n	1764
32	高级/教师	/a/n	714	32	区/县	/n/n	1454	32	考试/院	/v/n	1730
33	科学/发展观	/a/n	712	33	科学/发展观	/n/n	1436	33	研究/成果	/v/n	1716
34	新/时期	/a/n	688	34	年级/学生	/n/n	1406	34	联系/电话	/v/n	1698
35	新/政策	/a/n	686	35	本科/院校	/n/n	1350	35	考试/时间	/v/n	1696
36	高/中学生	/a/n	682	36	技术/人员	/n/n	1348	36	教育/专家	/v/n	1670
37	高/考生	/a/n	682	37	人才/市场	/n/n	1304	37	免费/师范生	/v/n	1654
38	重要/原因	/a/n	674	38	教师/队伍	/n/n	1268	38	报名/时间	/v/n	1632
39	普通/本科	/a/n	670	39	初中/毕业生	/n/n	1256	39	是/孩子	/v/n	1606
40	好/学校	/a/n	668	40	本科/专业	/n/n	1244	40	娱乐/专栏	/v/n	1578
41	知名/企业	/a/n	656	41	重点/大学	/n/n	1220	41	劳动/合同	/v/n	1576
42	重要/思想	/a/n	622	42	名牌/大学	/n/n	1208	42	毕业/证书	/v/n	1574
43	好/朋友	/a/n	614	43	爸爸/妈妈	/n/n	1156	43	高考/作文	/v/n	1522
44	大/限度	/a/n	604	44	高职/院校	/n/n	1146	44	解决/问题	/v/n	1518
45	具体/情况	/a/n	596	45	专业/技能	/n/n	1136	45	录取/分数线	/v/n	1494

5. 数据处理和下一步工作：

5．1 多字词表达式的噪声剔除：

停用词表法[8]

停用词指的是不能出现在多字词表达式前部和尾部的字词。例如在词性序列“/v/n”中，让/孩子、让/学生、让/人、是/学生、是/孩子等字符串中，“让”、“是”就作为停用词被选出来。

停用词表完全通过人工筛选来完成。

经过总结，初步确定可放在多字词字符串表达式首部的词类有：介词、助动词（例如能、不能、将、将要）、副词、形容词、疑问代词、连词。

可放在多字词字符串表达式尾部的词类有：数词、副词。

基于停用词表和词性标注，过滤掉首部和尾部出现停用词的多字词表达式。其算法执行过程，

Step1：读入一条多字词表达式，如果多字词表达式中首部和尾部没有停用词，到 step3，否则到 step2

Setp2：停用词在可接受的规定之内，到 step3，否则到 step4

Step3：输出多字词表达式

Step4：如果到词表末尾，则算法结束，否则到 Step1，分析下一条多字词表达式。

5.2 下一步的工作

在利用停用词表进行初步过滤后，我们准备再引入董振东老师的知网（Hownet）信息结构 271 条、詹卫东老师的短语结构规则 89 条及其他规则的内容，从句法与语义上进一步提高多字词表达式的准确率。

6. 总结：

本文所提出的用高频词提取特定领域的多字词表达式，技术路线易于操作，很好得应用于大规模真实语料库，降低了计算的复杂性，经过停用词表过滤后，效果很好。

参考文献：

[1] Irina Dahlmann,Svenja Adolphs. Pauses as an Indicator of Psycholinguistically Valid Multi-Word Expressions(MWEs). Proceedings of the workshop on A Broader Perspective on Multiword Expressions. Pages 49-56 ACL

[2] 王强军. 信息技术领域术语提取的初步研究，术语标准化与信息技术，2003 年 01 期

[3] 黄伯荣, 廖序东主编　现代汉语.　高等教育出版社.　59 页

[4] 赵军, 黄昌宁 汉语基本名词短语识别研究[A]. 汉语计量与计算研究[C]. 1998

[5] zhang Le, YAO Tian-shun, et. A statistical Approach to extract Chinese chunk Candidates from Large Corpora In:Proc.of ICCPOL-2003.Sheng Yang:PP.109-117.2003

[6] 陈晓 靳光瑾 黄昌宁 基于字的分词方法的实验研究 内容计算的研究与应用前沿——第九届全国计算语言学学术会议论文集 2007

[7] 罗盛芬，孙茂松. 基于字串内部结合紧密度的汉语自动抽词实验研究. 中文信息学报 2003 03 期

[8] Shailaja Venkastsubramanyan, Jose Perez-Carballo. Multiword Expression Filtering for Building Knowledge Maps. Second ACL Workshop on Multiword Expressions, July 2002, PP 40-47