

文章编号: 1003-0077(2015)03-0016-05

词汇计量研究与常用词知识库建设

俞士汶^{1,2}, 朱学锋¹

- (1. 北京大学 计算语言学研究所 计算语言学教育部重点实验室, 北京 100871;
2. 语言能力协同创新中心, 江苏 徐州, 221009)

摘 要: 面向自然语言处理的词汇语义研究应该以词汇的计量研究为基础。该文在评述汉语词汇计量研究的主要成果以后, 提出一个汉语常用词知识库的建设任务, 并给出常用词表的构造性定义、词表常用性的定量评价方法以及“部件词”的概念, 最后介绍现代汉语常用词知识库的总体设计和已经做的工作。期望常用词知识库的建设能为汉语词汇语义学研究、为中文信息处理事业的发展做出贡献。

关键词: 汉语常用词知识库;《中国语言生活状况报告》;综合型语言知识库;《现代汉语语法信息词典》;部件词

中图分类号: TP391 **文献标识码:** A

Quantitative Lexicon Study and Knowledge Base Construction for Commonly Used Words

YU Shiwen^{1,2}, ZHU Xuefeng¹

- (1. Key Laboratory of Computational Linguistics (Peking University), Ministry of Education
Institute of Computational Linguistics, Peking University, Beijing 100871;
2. Jiangsu Collaborative Innovation Center for Language Ability, Xuzhou, Jiangsu 221009, China)

Abstract: Natural language processing oriented lexical semantics researches should be based on quantitative study of the lexicon. After a brief survey on the main achievements of the quantitative Chinese lexicon, this paper proposes a project to build a knowledge base of commonly used words, for which we describe 1) a constructive definition of commonly used words list, 2) a quantitative method to measure the coverage of a given word list over an annotated corpus, and 3) the concept of “component word”. We also introduce the overall designs of the knowledge base and the current progress of this project. It is expected that the construction of such a knowledge base can contribute to the Chinese lexical semantics researches and the development of Chinese information processing.

Key words: knowledge base of Chinese commonly used words; *Language Situation in China*; comprehensive language knowledge base; *Grammatical Knowledge base of Contemporary Chinese*; component word

1 引言

面向自然语言处理的词汇语义研究应该以词汇的计量研究为基础, 汉语最大规模的计量研究成果当属中国国家语言资源监测与研究中心(本文简称其为 LRMR)每年发布的《中国语言生活状况报告》^[1]。据 2010 年的光盘数据, 仅 25 546 个常用词语即可覆盖全部语料的 95%。可见常用词语的语言知识库建设在通用型语言知识库建设中具有举足

轻重的地位。

北京大学计算语言学研究所(ICL/PKU)的“综合型语言知识库”(简称 CLKB)^[2]是词汇计量研究领域的另一项成果^[3]。笔者于 2007 年提出建设现代汉语常用词库的构想^[4], 只是立足于 CLKB 的基础。CLKB 的计量研究的规模远逊于 LRMR。不过, CLKB 的词汇计量研究也有更为深入之处, 可供常用词库建设借鉴。

本文在评述汉语词汇计量研究以后, 提出词表常用性的计量表示和“部件词”的概念, 并给出以“部件

词”为主体的现代汉语常用词知识库的概要设计。

2 词汇计量研究述评

2.1 关于中国国家语言资源监测与研究中的研究

词汇的计量研究必须立足于大规模的语料库。LRMR 自 2006 年以来每年收集各类媒体语料约十亿汉字,对媒体的用字用语一般情况、年度新词语、年度流行语等等进行调查、统计、分析,在汉语计量研究领域,这项研究的规模和广度应该是最大的。

LRMR 采用软件技术对语料进行了词语切分和词性标注,得到丰富的统计数据,并有选择地在《中国语言生活状况报告》上发表了部分成果。LRMR 明确说明调查对象是由加工软件得到的切分单位,意在保留语言生活的实态,同时也保留了语言技术的实态。公开发表的内容(包括光盘数据),特别是“高频词语表”,经过多种形式的人工校订,质量是上乘的,在诸多领域都是有参考价值的,更值得常用词库参照。

LRMR 的语言数据资源是逐年积累的,已持续七年,可以从共时和历时两个角度进行研究。LRMR 已对 2005~2009 的五年语料进行了分析。“覆盖整个语料 80% 的词种个数稳定在 4 500 个左右,覆盖率达到 90% 的词种个数稳定在 12 000 左右。可见,高频常用词语的数量相对稳定。”这个结论显然对常用词知识库建设具有指导意义。

LRMR 在词汇计量研究领域也有理论上的创新,提出了一些新的统计量及其计算公式,如分布率、使用率、频率比值、频序比值等等。

2.2 关于《常用词表(草案)》等数据成果

《中国语言生活状况报告》属于《中国语言绿皮书》之 B 系列,其 A 系列主要刊载引导语言生活的“软性”规范。到 2011 年止,A 系列出了两册。第一册是《现代汉语常用词表(草案)》^[6](以下简称《常用词表(草案)》),简介如下:

(1)《常用词表(草案)》有 56 008 个使用频率高、适用范围广的词语;

(2)《常用词表(草案)》正文中的“词语”按“频序号”排列。另外提供了音序索引;

(3)《常用词表(草案)》“优先收录带有词根性质的词语”,“原则上双音节者适当从宽,三音节及其

以上者适度从严”,这与笔者对“部件词”^[4]的认知可谓殊途同归;

(4)对同形异义词,《常用词表(草案)》实际上只区分了汉字相同而读音不同的情况。

较早公开出版的基于计量研究的词表还有《现代汉语频率词典》(包含 16 593 个词语^[6])和《信息处理用现代汉语常用词表》(将近 4 万词语)^[7]。

2.3 CLKB 相关的计量研究

综合型语言知识库中的“现代汉语多级加工语料库”^[8]与词汇计量研究直接相关,实现多级加工的数据资源基础是 CLKB 中的《现代汉语语法信息词典》(GKB)^[9]和以及 ICL/PKU 研制的“现代汉语语义词典”(CSD)^[10]。

看现代汉语多级加工语料库的一个实例:

① 19980101-05-001-018/m 为鼓励中学生多听多讲普通话

② 19980101-05-001-018/m 为/鼓励/中学生/多/听/多/讲/普通话/

③ 19980101-05-001-018/m 为/p 鼓励/v 中学生/n 多/a 听/v 多/a 讲/v 普通话/n

④ 19980101-05-001-018/m 为/p!B 鼓励/v 中学生/n 多/a 听/v 多/a 讲/v!1 普通话/n

⑤ 19980101-05-001-018/m 为/p!B 鼓励/v 中学生/n 多/a 听/v 多/a 讲/v!1-1 普通话/n

①原始语料的语句前的一串数字指示本语句在语料库中的位置,即《人民日报》1998 年 1 月 1 日第五版第一篇文章第 18 句。原始语料包括 1998 年和 2000 年两年完整的《人民日报》,共 5 200 万字,都完成了②和③的基本加工:词语切分和词性标注,如 p 是介词,v 是动词。④在③的基础上进行粗粒度义项标注,即标注《现代汉语语法信息词典》(GKB)的“同形”信息。在 GKB 中,动词“讲”的“同形”字段区分为“1”和“2”,“1”是“说,解释,商量”义,“2”是“讲求,讲究”义,本语句中“讲”是“1”的意思,标注为“!1”;完成这一步的语料有 2 800 万字(1998 年 1 月和 2000 年全年的)。⑤在④的基础上并依据 CSD 的“义项编码”字段进行细粒度义项标注;CSD 基于 GKB 的“同形”,增设了“义项编码”字段,如将 GKB 中“同形”为“1”的动词“讲”的“义项编码”区分为“1”,“2”,“3”,分别代表“说,解释”义,“就某方面而言”义,“商量”义,这个语句中“讲”的意义与“义项编码”“1”相符,动词“讲”的细粒度义项标注为“讲/v!1-1”;完成这一步的语料有 700 万字(2000 年

1 月至 3 月的)。

可以把上面的实例改造成数据库文件形式(见表 1,简称为 D)。利用 D 和 CSD 的共有字段“词语”、“词类”、“同形”、“义项编码”可以把 D 和 CSD 集成到一起;如果语料库只标注到同形信息,D 中没有“义项编码”字段,可以和 GKB 集成。

表 1 多级加工语句转换成数据库文件的示例(D)

词语	词类	同形	义项编码	频次	年	月	日	版	篇	句	位
为	p	A		1	1998	1	1	5	1	018	1
鼓励	v			1	1998	1	1	5	1	018	2
中学生	n			1	1998	1	1	5	1	018	3
多	a			1	1998	1	1	5	1	018	4
听	v			1	1998	1	1	5	1	018	5
多	a			1	1998	1	1	5	1	018	6
讲	v	1	1	1	1998	1	1	5	1	018	7
普通话	n			1	1998	1	1	5	1	018	8
...

如果在表 1 中删去“位”字段,“多”有了两个相同的记录,删去一个,将留下的一个的“频次”改为“2”,便得到了这个语句的词频,实际上是“多”这个词的细粒度义项的频次。按同样操作过程,逐次删去“句”、“篇”、“版”、“日”、“月”、“年”,就能得到一篇文章、一个版面、一天、一个月、一年乃至整个语料库的词频(细粒度义项的频次)。

如果只标注到“同形”,可按照同样的办法得到每个词语的粗粒度义项(即标注 GKB 的“同形”)的频次。如果再删去“同形”字段,得到的将是区分词性的词频;如果连词类代码也删去,得到的只是词形的频次。

更深入的,还有关于均根匀度的研究^[11]以及关于词语的语法属性的计量研究^[12]。

3 词表常用性的计量表示

3.1 常用词表的构造性定义

“常用”通常是个模糊的概念。本文给出基于计量数据的常用词表的构造性定义:将给定的文本语料简化为二元组 $C=\{u_j, p_j \mid 1 \leqslant j \leqslant m\}$, u, p 都是 m 维向量: $u=(u_1, u_2, \cdots, u_j, \cdots, u_m)$, $p=(p_1, p_2, \cdots, p_j, \cdots, p_m)$ 。 u_j 代表语料中互不相同的构成单元,即当 $i < > j$, 则 $u_i < > u_j$; p_j 为 u_j 在语

料库中出现的频率 $p(u_j) = p_j$ 。 p_j 满足归一化条件,即式(1)。

$$\sum_{j=1}^m p_j = 1$$

(1)

将 p_i 按降序排列,于是当 $s < t$ 时, $p_s \geqslant p_t (1 \leqslant s \leqslant m, 1 \leqslant t \leqslant m)$ 。 C 实际上就是语料库的全部构成单元的列表形式。

基于给定的常数 $\delta (0 \leqslant \delta \leqslant 1, \text{称之为覆盖系数})$ 确定 C 的一个子集 L ,

$$L = \{u_j, p_j \mid 1 \leqslant j \leqslant k\}, 1 \leqslant k \leqslant m, \text{使其满足}$$

$$\sum_{j=1}^k p_j \geqslant \delta \quad \text{和} \quad \sum_{j=1}^{k-1} p_j < \delta$$

则 L 为语料库 C 的关于覆盖系数 δ 的常用构成单元列表。对于不同深度的加工语料,构成单元不同:词语,带词性的词语,区分同形(粗粒度义项)的词语,区分细粒度义项的词语,不妨概称为“词语”。对于原始语料,构成单元就是字。还可以把构成单元看成是句子。

本定义的一个重要理念是常用词表是语料库和覆盖系数的函数,即 $L=L(C, \delta)$,不存在对任何语料库都适用的常用词表。

3.2 已有词表的常用性检测

对一个已经存在的词表 W ,可用以下三个指标检测它关于给定语料库 C 的“常用性”。

一次覆盖率 R_1 : 在语料库 C 中出现的词表 W 中的词语数 n 与 C 中所有不同词语的总数 m 之比,即式(2)。

$$R_1 = \frac{n}{m}$$

(2)

多次覆盖率 R_t : 在语料库 C 中出现的词表 W 中的 n 个词语的频次 $g_k (k=1, 2, \cdots, n)$ 之和 与 C 中所有词(总数为 m)的频次 $f_j (j=1, 2, \cdots, m)$ 之和的比,即式(3)。

$$R_t = \frac{\sum_{k=1}^n g_k}{\sum_{j=1}^m f_j}$$

(3)

词典的有效率 V : 在语料库 C 中出现的词表 W 中的词语数 n 与 W 中的词语总数 N 之比,即式(4)。

$$V = \frac{n}{N}$$

(4)

R_1, R_t, V 的值都在区间 $[0, 1]$ 上。 R_1, R_t 的值越大则覆盖率越高,理想值是 1。当 R_1 的值不大,而

R_i 的值却较大, 表示该词表覆盖了语料库 C 的较多的常用词。若 V 取理想值 1, 表示词表 W 中的词在语料库 C 中都用到了。

4 常用词表和“部件词”

4.1 “部件词”的概念

无论 LRMR 还是 CLKB 已做的有关词语的计量研究实际上都是基于“切分单位”的, 词语的频次就是切分单位的频次。“切分单位”与通常的“词”显然有差别。另外, 从通常认可的“词”中还可以析出更基本的有构词能力的词, 笔者将其称为“部件词”^[4], 大体相当于《常用词表(草案)》所指的“带有词根性质的词语”。“部件词”的实例有:

(1) 像“一九九八年”、“一九九七年”等都是切分单位, LRMR 又称其为“时间表达式”, CLKB 认为它就是时间词。其中, 数词“一”、“九”、“七”、“八”和名词“年”是“部件词”。

(2) 在“积极”、“积极分子”和“积极性”这三个词中, “积极”、“分子”、“性”是“部件词”。

常用词表应该以“部件词”为主体。

把一部词典收录的所有词语(登录项或词条)或语料中的所有切分单位看作词的全集, 可以从中区分出“部件词”和“非部件词”, “非部件词”是由“部件词”构成的。“部件词”和“非部件词”之间、它们的频次以及同总频次之间有以下 ①~④ 的关系。

① 部件词集 \cup 非部件词集 = 词的全集;

② 部件词集 \cap 非部件词集 = \emptyset (空集);

③ 在语料库中, 部件词 e 的累计频次 = 部件词 e 本身的频次 + 所有包含 e 的非部件词 w 的频次;

④ 一个非部件词拆分后, 部件词的频次和总频次要重新计算:

设 某个非部件词 w 可拆分为 n 个部件词 $e_j (j=1, 2, \dots, n)$,

w 拆分前的频次为 f , 每个 e_j 的频次为 f_j , 所有词的总频次为 F ,

w 拆分后, 每个 e_j 的频次为 f_j' , 总频次为 F' ,

则 $f_j' = f_j + f, (j=1, 2, \dots, n)$,

$F' = F + (n-1) \times f$

4.2 基于“部件词”的常用词表的确定

常用词表的确立要基于频率等计量数据, 还应

该以“部件词”为主体。

第一步: 选择足够大的有限的现代汉语语料库 C 。自然优先考虑 LRMR 8 年积累的约 80 亿字的基本加工语料库和 CLKB 中的多级加工语料库。

第二步: 对 C 进行多级加工。尽可能继承、利用 LRMR 和 CLKB 的既有成果。计算 C 的构成单元 $u_j (j=1, 2, \dots, m)$ 的频次 f_j 和所有构成单元的总频次 F , 按 f_j 降序排列 u_j , 得到 u_j 的列表。

第三步: 对每一个 u_j 进行辨析, 分出部件词和非部件词, 将非部件词拆分为部件词(由于列表很大, 需要对频次给出下限, 频次少于下限的 u_j 不予处理)。

第四步: 重新计算部件词的累计频次和由部件词构成的所有词语的总频次 F' 。

第五步: 根据每一个部件词 e_j 的累计频次和总频次 F' 计算 e_j 的频率 p_j , 按 p_j 降序排列 e_j , 得到部件词列表。

第六步: 考虑计算部件词的均根匀度, 按均根匀度和频率的加权平均值调整部件词列表。

第七步: 给定覆盖系数 δ , 构造出基于部件词的常用词表。

第八步: 适当选择若干典型的常用非部件词, 加进常用词表。

将非部件词拆分为部件词也比较复杂。在多级加工语料中, 非部件词可能带有词性、同形、义项编码等信息, 拆分后如何确定部件词的相应信息, 这一步虽有共性规律可循, 但很多情况需要个别处理。笔者拟另撰文阐述其共性规律和处理个性情况的实践经验。

5 常用词知识库的总体设计与工程实践

在主要由部件词构成的常用词表的基础上建设常用词知识库, 可以大大提高常用词知识库的有效性和常用词知识库的建设效率。

常用词表的构造和常用词知识库的设计与实现可以并行进行, 相互促进, 因为最常用的一些词(几百个乃至几千个)总会在常用词表中。

常用词知识库的框架结构不妨仍继承《现代汉语语法信息词典》(GKB) 和现代汉语语义词典(CSD)的数据库文件格式。

常用词知识库可划分为词汇知识库、句法知识库、语义知识库、构词规则库、例句库等五部分。常用词知识库的所有数据库文件的第一个字段都是

“词语”。

词汇知识库相当于 GKB 的总库,可有选择地继承 GKB 总库的内容,将各类词库中的“释义”字段移到这里,另增加一些字段,如“异形”、“变体”、“异读”等。

句法知识库大体继承 GKB 的各类词的数据库,字段也要有所拆分、删节,使其更精炼,同时吸收 GKB 以外虚词知识库、成语知识库等的部分内容。

语义知识库有选择地继承 CSD 和 CLKB 中的中文概念词典 CCD 的内容。

构词规则库反向利用部件词拆分规则,构造非部件词的生成规则。

例句库汇集每个词的若干典型例句,建立全息语料库(每一个词的各种词法、句法、语义信息都参照其所在的上下文显性地标注出来),为词语属性的计量研究做好准备。

常用词知识库是一个浩大的语言工程。不过,基础是坚实的,已有诸多成果可以集成。对于新增的任务,也有了相当多的工程实践经验,如已拆分了数以万计的非部件词,为三万多高频词语的每一个都选取了 3~5 个例句^[13]。也做过全息语料库的小规模试验。

6 结语与谢辞

CLKB 是 ICL/PKU 师生多年努力的成果,也得到中文信息处理学界同仁的扶植,不无侥幸地获得 2011 年度国家科技进步奖二等奖。笔者衷心期望这项成果能继续发展。常用词语知识库是值得关注的-一个发展方向。限于精力和条件,笔者已经不可能挑起这副重担。幸运的是,鲁东大学汉语辞书研究中心主任亢世勇教授和邱立坤博士已表示对合

作研制常用词知识库有兴趣。衷心期望常用词知识库的研制能成为有相同志趣者共同努力的目标,并为汉语词汇语义学研究、为中文信息处理事业的发展做出贡献。

参考文献

- [1] 国家语言资源监测与研究中心.《中国语言生活状况报告》[M],北京:商务印书馆,2011.
- [2] 俞士汶,穗志方,朱学锋.综合型语言知识库及其前景[J],中文信息学报,2011,25(6):12-20.
- [3] 俞士汶,朱学锋,段慧明,等.汉语词汇语义研究及词汇知识库建设[J],语言暨语言学,2008,9(2):359-380.
- [4] 俞士汶,朱学锋,支流.基于计量研究的现代汉语常用词库的构建[C],张普、王铁琨主编《中国语言资源论丛》,北京:商务印书馆,2009:289-301.
- [5] 《现代汉语常用词表》课题组.《现代汉语常用词表(草案)》[M],北京:商务印书馆,2008.
- [6] 北京语言学院语言教学研究所编.《现代汉语频率词典》[M],北京:北京语言学院出版社,1986.
- [7] 刘源,谭强,沈旭昆.《信息处理用现代汉语分词规范及自动分词方法》[M],北京:清华大学出版社,1994.
- [8] 俞士汶,段慧明,朱学锋,等.北京大学现代汉语语料库基本加工规范[J],中文信息学报,2002,16(5):49-64.
- [9] 俞士汶,朱学锋,王惠,等.现代汉语语法信息词典详解[M],第二版,北京:清华大学出版社,2003.
- [10] 王惠,詹卫东,俞士汶.现代汉语语义词典规范[J],汉语语言与计算学报,2003,13(2):159-176.
- [11] 张化瑞.以均根勾度为中心的语言信息计量研究[D],北京大学博士学位论文,2010.
- [12] 王萌.面向概率型词汇知识库建设的名词语言知识获取[D],北京大学博士学位论文,2010.
- [13] 朱学锋,张化瑞,段慧明,等.《汉语高频词语法信息词典》的研制[J],语言文字应用,2004,3:98-104.



俞士汶(1938—),教授,主要研究领域为计算语言学,语言知识库。
E-mail: yusw@pku.edu.cn



朱学锋(1937—),副教授,主要研究领域为计算语言学,语言知识库。
E-mail: yusw@pku.edu.cn