◎数据库、数据挖掘、机器学习◎

基于事件语义距离的V1-V2述结式判别研究

马 腾,詹卫东
MA Teng, ZHAN Weidong

北京大学 中文系,中国语言学研究中心,计算语言学教育部重点实验室(北京大学),北京 100871
Department of Chinese Language & Literature, Center for Chinese Linguistics PKU, Key Lab of Computational Linguistics, MoE, Peking University, Beijing 100871, China

MA Teng, ZHAN Weidong. Identification of verb-resultative construction based on event distance. Computer Engineering and Applications, 2015, 51(17):107-112.

Abstract: The Dictionary of Verb-Resultative Construction in Contemporary Chinese is created by hand, to describe the knowledge about verb-resultative construction. Based on this dictionary and the big scale of corpus data, it proposes an algorithm to calculate the event distance in a V1-V2 verb-resultative construction. With this model, it can identify the verb-resultative construction from different V1-V2 combination. Using this method, it can extract the V1-V2 verb-resulative constructions from corpus automatically with high accuracy.

Key words: verb-resultative construction; compound event; event distance

摘 要:“现代汉语述补结构用法词典”是人工建立、用于描述述补结构相关信息的语言知识资源。经过人工对词条的收集、释义等编写工作,词典已形成一定规模。在此基础上,尝试借助计算机技术,依据事件语义学的理论,利用现有语言知识资源以及大规模语料数据,寻找述结式复合事件语义距离计算的方法,对述结式进行定量描写,以帮助扩大词典规模,同时有助于深化对特有语言现象——述补结构的认识。实验结果表明该方法具有较高的准确率和识别率。

关键词:述结式;复合事件;语义距离
文献标志码:A **中图分类号:**TP391 doi:10.3778/j.issn.1002-8331.1309-0414

1 引言

述补结构,如吃饱、哭肿、看懂、飞走等,是现代汉语比较常见、也是比较特殊的一种句法结构。其中尤以上面所举的表示结果的述结式使用最为频繁,也是汉语学界讨论较多的一种结构。从形式上来看,述结式由两个谓词性成分V1-V2组合而成。语义上,V1所表示的动作导致了V2所表示动作或状态的发生与出现。如述结式“吃饱”,形式上由动词“吃”和形容词“饱”组合而成;语义上表示“吃”这个动作发生后,导致了“饱”这一状态的出现。

由于述结式结构的特殊性,在面向计算机的自然语言处理中,如何对此类结构进行识别、分析以及生成,就成为了亟待解决的问题。观察可以发现,并不是任意两个谓词性成分组合都能构成在话语环境中出现的合法的述结式。如例1中列出的三组V1-V2组合。

例1 (1)吃-饱、洗-干净、哭-肿
(2)*吃-饿、*洗-饱、*笑-肿
(3)?吃-懂、?洗-脏

其中,(1)是比较常见且容易接受的合法述结式结构;(2)是在正常会话中不会使用的组合形式,V1-V2组合并不能构成合法的述结式结构;(3)的两个例子,在使用上介于(1)、(2)之间,并不把它们看作和(1)一样的典型

基金项目:教育部人文社会科学研究项目规划基金项目(No.12YJA740104);国家社科基金项目(No.12BYY061)。
作者简介:马腾(1989—),男,硕士,主要研究方向:计算语言学;詹卫东(1972—),通讯作者,男,副教授,主要研究方向:现代汉语、应用语言学及计算语言学。E-mail:matengneo@gmail.com
收稿日期:2013-09-27 **修回日期:**2013-11-14 **文章编号:**1002-8331(2015)17-0107-06
CNKI**网络优先出版:**2014-02-26, http://www.cnki.net/kcms/doi/10.3778/j.issn.1002-8331.1309-0414.html

述结式,但在实际使用中也并不会像(2)三个例子一样一定不能出现,如下面的两个例子是在网上找到的,对应于例1(3)的两个组合。

例2 吃懂法兰西

例3 别让公共场所的劣质洗手液“洗脏”了你的手

汉语语言学界对述补结构进行过广泛和深入的研究,但是以往的讨论主要集中在V1-V2整体的论元结构如何由V1和V2各自的论元结构导出、述结式与相关句法结构(如“把”字句、重动句等)的互动、述结式的认知研究等方面[1-13],却很少论及上面所列出的组成述结式的两个谓词性成分的组配约束限制的问题。

对于例1(1)的例子,可以将其作为典型述结式收入述补结构词典,计算机处理时可直接查询词典得到相关述结式信息。但是对于诸如例1(3)中不典型且大量存在的述结式组合,则需要寻找一种计量方式以判定其构成述结式的可能性。

基于此问题,为了对现代汉语中的述补结构相关信息进行详尽的描写,提供可供机器使用的语言知识资源,北京大学联合日本早稻田大学,创建了“现代汉语述补结构用法词典”。该词典对现代汉语述补结构的搭配、组合、语义及用法等信息进行了详尽的描写。

目前词典经过长时间的人工添加、编辑,已形成了一定的规模(述补条目共21 031,其中述结式7 942个),为做述补结构自动判定、生成的工作提供了有力的依据。但正如上文所指出的,对于实际使用中存在的不常见述结式组合(如例1(3)),以及没有收录的其他常见组合,都需要在已有词典的基础上做进一步的扩展。但随着词典规模的进一步扩大,人工进行添加的成本也会相应增加(在语料中搜索的成本加大)。为了辅助人工进行词典扩展,本文提出了一种基于复合事件语义距离计算的述结式自动识别算法。该算法可对V1-V2组合构成述结式的可能性进行定量描述,以便抽取潜在述结式结构进行词典的扩充。

2 现代汉语述补结构用法词典

词典(http://ccl.pku.edu.cn/vc)主要描述了现代汉语述补结构以下三个方面的信息。

(1)述语和补语能否搭配的信息。例如,词典中收录“吃-饱”组合,表示“吃”跟“饱”可以搭配成为述补结构,语义解释为“吃”的动作发出者在经过“吃”这个动作之后,会导致“饱”这个状态的出现;词典不收录“吃-醒”组合,表示“吃”跟“醒”不能搭配成为述补结构,因为不存在由“吃”到“醒”的致使-结果变化。

(2)述补结构跟相关结构在用法上的相同点和差异。例如,“补定对比/补状对比”属性描述了词条分别用作补语和定语(或状语)时的相同点和差异,借此可判定诸如“画直[补语]了”与“画直[定语]线”的用法差别。

(3)述补结构整体的用法特点。例如述语动词的语义角色相对于述补结构的前后位置;以及述补结构能否用于“把”、“被”句,重动句中等。

利于以上信息,词典对述补结构的各组成部分以及整体结构的用法都做了详尽的描写,而且词典信息存储于关系型数据库,有利于计算机对数据进行查询。

词典对述补结构条目进行描述的基本框架是以述语为纲列出词典的条目。对每个述语(动词或形容词),按补语的不同语义类型来分项描写所能搭配的补语。补语按照不同的语义类型主要分为五类,分别是结果补语、趋向补语、可能补语、程度补语、介词补语。

具体的,一个述补结构条目的内容主要由以下四部分组成:

(1)基本信息。包括[述语词条][拼音][词性][义项][释义]等。

(2)语义角色。描写参与该述语事件的各种名词性成分。

(3)补语分项举例。描写各个具体的述补结构的用法特点和语义性质。

(4)对述补结构整体的用法特点和语义性质进行概括说明。如比较述补结构跟相关的状中结构在表达功能上的差异等。

表1、表2分别为词典整体数据统计以及按补语类型分类的述补条目数据统计(数据统计截止日期:2013年1月1日)。

表1 词典整体数据统计表

类型		数据
述语	词	1 639
	义项	2 014
补语	词	494
	义项	580
述补结构		21 031

表2 词典中各类型补语数据统计表

补语类型	数目	百分比/%
结果补语	7 942	37.76
趋向补语	7 267	34.55
可能补语	3 390	16.12
程度补语	1 336	6.35
介词补语	1 096	5.21
总数	21 031	100.00

3 述结式复合事件语义距离计算

3.1 述结式中的复合事件

詹卫东[14]提出,从事件语义学的角度来看,现代汉语的述结式实际上是一个复合事件的压缩编码形式。如下面所举出的两个例子。

	事件1	事件2	复合事件(压缩编码形式)
例4	妈妈喂女儿。	女儿饱了。	妈妈喂饱了女儿。
例5	张三洗衣服。	衣服干净了。	张三把衣服洗干净了。

其中，事件1和事件2在语义上存在关联，事件1的发生导致了事件2的发生，并且这两个事件存在着共有事件角色。对于这种语义上存在致使关联的两个子事件，可以将其压缩编码为一个复合事件输出。复合事件的语义结构图可以用图1进行表示（以例4中所举事件为例）。

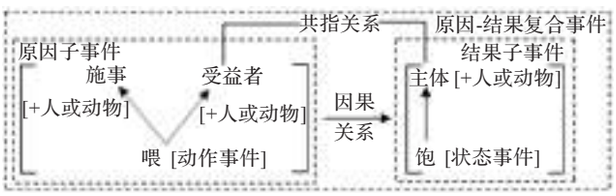

图1　复合事件“喂饱”的事件语义结构图

从事件语义学的框架来看，V1-V2能否组合成为一个合格的述结式，实则是这两个谓词性成分所激活的事件能否被压缩编码为一个复合事件。而构成复合事件与子事件之间的语义关系有关，可以用子事件间的语义距离来定量描述。据此，对述结式中V1-V2的组配约束的考察，可以转化为对这两个谓词性成分所激活的两个事件的语义距离的计算。若两个事件之间语义距离较大，则语义相关度低，压缩编码为一个复合事件的可能性较小，这两个事件所对应的谓词性成分能够构成一个述结式的可能性也相应较小；反之亦然。

认为两个子事件构成述结式中的复合事件需要满足以下两个条件。

条件1：事件A和事件B存在共有事件角色。

条件2：事件A和事件B之间存在“致使-结果”的语义关系。

例如下面的两个不能构成述结式中复合事件的例子。

	事件1	事件2	复合事件（压缩编码形式）
例6	妈妈生病了。	女儿失眠了。	*妈妈生病失眠了女儿。
例7	张三吃饭。	张三饿了。	*张三吃饿了。

其中，例6中的两个子事件不存在共有事件角色。事件1的事件角色集合为{妈妈}，事件2的事件角色集合为{女儿}，二者不存在交集，即不存在共有事件角色，缺少构成复合事件的必要关联。例7中的两个子事件虽然存在共有事件角色“张三”，但是事件1“吃”发生后，很难引起事件2“饿”作为一个状态出现，不存在构成述结式复合事件的必要语义关系。

从上面的两个例子来看，无论是共有事件角色，还是“致使-结果”的语义关系，归根到底是事件语义距离制约着两个事件是否能够复合成为一个述结式事件。因此，下文重点对事件语义距离的计算进行讨论。

3.2　复合事件“致使-结果”语义相关度计算

复合事件的“致使-结果”语义相关度计算，利用两种资源，分为两个部分进行计算。一是依赖大规模语料库的基于概率统计的计算，二是依赖现有语言知识资源的基于事件相似度的计算。基于知识资源的计算方法准确率高，但覆盖率较低；基于概率统计的计算覆盖率较高，但准确率往往较低。本文使用的两部计算方法可以有效地融合这两种计算方法的优点，弥补二者的缺点。

据此，事件A、B的“致使-结果”语义相关度可以用公式（1）表示。

$$ER(A,B)=\gamma\cdot ER1(A,B)+\delta\cdot ER2(A,B),\gamma+\delta=1 \quad (1)$$

其中，γ,δ 是调整两种计算方法贡献度的权值，可根据实验效果进行调整。本次实验设置 $\gamma=\delta=0.5$ 。

3.2.1　基于概率统计的“致使-结果”语义相关度计算

若事件A、B具有“致使-结果”语义关系，则在时间顺序和文本顺序上事件B在事件A之后出现。据此通过计算语料中事件B在事件A之后出现的条件概率近似模拟两个事件的“致使-结果”语义关联度，计算步骤如下。

（1）统计语料中事件B在事件A之后出现的加权次数，以及事件A出现的总次数。

（2）计算事件B在事件A之后出现的条件概率，公式如下。

$$ER1(A,B)=P(B|A)=\frac{P(A\cdot B)}{P(A)}\cong \frac{\text{事件B在事件A之后出现的加权次数}}{\text{事件B出现的总次数}} \quad (2)$$

其中加权次数是根据事件B在事件A之后出现的位置而调整后的出现次数。在实际计算中，以一个完整的句子为窗口来统计事件B在事件A之后出现的次数。其中，事件B在事件A之后出现的加权次数等于事件B在事件A之后出现的序数的倒数。例如，若在一个句子中存在事件序列A E1 E2…En B，事件B作为第 $(n+1)$ 个事件出现在事件A之后，则此时的加权次数为 $1/(n+1)$ 。

3.2.2　基于知识资源的“致使-结果”语义相关度计算

假设存在于述补词典中的典型述结式的“致使-结果”语义相关度为1。以此为标准，对于任意两个V1-V2组合，计算其与典型述结式的最大事件相似度，亦可以用来表征两个谓词性成分之间的语义相关度。

例如，计算未在词典中出现的“吃-懂”组合，可以计算事件“吃”与能够导致事件“懂”作为结果发生的事件（即补语“懂”的述语事件，如“看、读”等）间的事件相似度，以及事件“懂”与事件“吃”所导致结果事件（即述语“吃”的补语事件，如“饱、光”等）间的事件相似度，用来表征“吃-懂”与典型述结式事件的事件相似度。从而可以得到相对于“致使-结果”语义相关度为1的典型述结式的语义相关度结果，即若事件相似度高，则该V1-V2组合的“致使-结果”语义相关度高，反之则语义相关度较低。

这里用“现代汉语述补结构用法词典”作为典型述结式词典，凡在该词典中出现的述结式皆认为是典型述

结式。事件相似度计算转化为谓词层面的词语相似度，利用刘群、李素建[15]提出的基于《知网》的词汇语义相似度计算方法。

对于子事件A，其基于知识资源的“致使-结果”语义相关度计算步骤如下：

(1)查询词典得到所有由B作补语的述结式的述语集合，记为$\|A\|$，对于$\|A\|$中的每一个词语$w1$计算其与A的词语相似度$Sim(w1,\|A\|)$。

(2)查询数据库得到所有由A作述语的述结式的补语集合，记为$\|B\|$，对于$\|B\|$中的每一个词语$w2$计算其与B的词语相似度$Sim(w2,\|B\|)$。

(3)最终输出结果为：

$$ER2(A,B)=m\cdot\mathrm{Max}(Sim(w1,A))+ n\cdot\mathrm{Max}(Sim(w2,B)),w1\in\|A\|,w2\in\|B\| \quad (3)$$

$Sim(x,y)$的计算参考文献[15]。m、n是分别作用于述语事件相似度和补语相似度的权值，可根据实验进行相应调整。这里假设述语与补语对整体事件相似度的贡献一致，因此，在下面的计算中取$m=n=0.5$。

4　实验及结果分析

为了验证计算方法的效果，分别在小规模实例测试、大规模随机测试以及分类定向测试等三个方面进行述结式复合事件语义关联度的计算实验。

计算中用到的语料库是“北京大学中国语言学研究中心现代汉语语料库(CCL语料库)”，规模为4.77亿字(1.06 GB)。分词以及词性标注工具为中国科学院计算技术研究所开发的ICTCLAS分词及词性标注系统。

4.1　小规模实例测试实验

利用上面给出的公式，首先对本文中提到的V1-V2组合进行定量计算。

表3列出了上文提到的三组述结式实例。其中A组是接受程度较高的典型述结式；B组是接受程度较低，但在实际语料中出现的述结式；C组是不被接受且极少在语料中出现的V1-V2组合。

从表3的计算结果来看，整体上还是比较符合三组的分类的。如“洗-干净”和“吃-饱”作为接受程度较高的典型述结式，计算得分也是最高的；而作为接受程度低且在语料中极少用作述结式出现的“笑-肿”和“吃-饿”则得分较低。

有问题的计算结果主要有以下两点。

(1)A组中的“哭-肿”：作为接受程度较高且在词典中出现的述结式，“哭-肿”的得分并不像“洗-干净”和“吃-饱”的那么高。从表中数据可知，问题主要出现在基于概率的计算得分上，仅为0.002 41，得分不仅低于B组的“洗-脏”，还远低于C组的“吃-饿”。

(2)B组中的“吃-懂”：得分较低，基本上和C组的“洗-饱”处于同一级别。但是作为在实际语料中作为述结式出现的“吃-懂”，在认知上作为述结式的可接受程度远远高于“洗-饱”。分析数据发现，得分较低的原因主要是因为补语相似度的得分较低。虽然“吃”和“懂”的述语“看”的相似度为1，但是“懂”和“吃”的补语的相似度最高也只是0.242 42。由于实验将述语相似度和补语相似度的权重设为一样，即使述语相似度较高，也被很低的补语相似度给拉低了。

上面这两个问题也反映出了本实验所用到的公式的不足之处。

一是基于概率的“致使-结果”语义关联度的计算方法。这里用基于位置的条件概率来近似描述两个事件之间的“致使-结果”语义关联度，方法比较简单，同时误差也就会比较大，还需要在此基础上寻找更加精确的计算方法。

二是基于知识库资源的复合事件相似度两部分的权值设置。本次实验分别计算了述语的事件相似度和补语的事件相似度，并将两个结果按照相同的权重进行相加以得到最终的复合事件相似度结果。但是从上面的实验结果也可以看出来，这种权重一致的计算方法存在问题。在两个子事件的相似度一致(同高或同低)的情况下，不会出现问题，但如果出现诸如“吃-懂”这种两个子事件的事件相似度相差较大的情况，就会另最终结果出现严重的误差。

4.2　大规模随机测试实验

本节测试计算公方法大规模V1-V2组合上的效果。

从词典中随机抽取了100个能够出现在述结式中的述语以及100个能够出现在述结式中的补语，计算两两组合构成的述结式的复合事件语义距离。

表3　三类V1-V2组合实例分析

分组	排序	V1	V2	*Rd*	*ER*1	*ER*2	述语相似度(最大值)	补语相似度(最大值)
A	1	洗	干净	0.585 5	0.171 1	1.000 0	1.000 0	1.000 0
	2	吃	饱	0.582 3	0.164 6	1.000 0	1.000 0	1.000 0
	4	哭	肿	0.501 2	0.002 4	1.000 0	1.000 0	1.000 0
B	3	洗	脏	0.520 6	0.041 3	1.000 0	1.000 0	1.000 0
	5	吃	懂	0.312 4	0.003 7	0.621 2	1.000 0	0.242 4
C	6	洗	饱	0.311 3	0.001 5	0.621 2	0.242 4	1.000 0
	7	笑	肿	0.198 0	0.000 0	0.396 1	0.347 8	0.444 4
	8	吃	饿	0.126 8	0.031 4	0.222 2	0	0.444 4

表4　计算结果排名前20的V1-V2组合实例及相关计算信息

排序	V1	V2	*Rd*	*ER*1	*ER*2	述语相似度(最大值)	补语相似度(最大值)
1	删	掉	0.884 5	0.769 1	1.000 0	1.000 0	1.000 0
2	晒	黑	0.742 5	0.485 0	1.000 0	1.000 0	1.000 0
3	熏	黑	0.710 5	0.421 0	1.000 0	1.000 0	1.000 0
4	掐	死	0.703 6	0.407 3	1.000 0	1.000 0	1.000 0
5	钉	死	0.689 6	0.379 2	1.000 0	1.000 0	1.000 0
6	砍	掉	0.680 7	0.361 5	1.000 0	1.000 0	1.000 0
7	杀	掉	0.660 3	0.320 6	1.000 0	1.000 0	1.000 0
8	憋	死	0.655 7	0.311 4	1.000 0	1.000 0	1.000 0
9	扫	清	0.654 2	0.308 4	1.000 0	1.000 0	1.000 0
10	念	完	0.649 9	0.299 9	1.000 0	1.000 0	1.000 0
11	剪	断	0.644 7	0.289 4	1.000 0	1.000 0	1.000 0
12	扒	掉	0.639 2	0.278 5	1.000 0	1.000 0	1.000 0
13	剪	掉	0.638 9	0.277 9	1.000 0	1.000 0	1.000 0
14	浇	灭	0.635 4	0.270 8	1.000 0	1.000 0	1.000 0
15	憋	足	0.634 4	0.268 9	1.000 0	1.000 0	1.000 0
16	蜇	死	0.627 1	0.254 3	1.000 0	1.000 0	1.000 0
17	修	通	0.626 7	0.253 4	1.000 0	1.000 0	1.000 0
18	绞	尽	0.657 3	0.384 1	0.930 5	1.000 0	0.861 1
19	应用	成功	0.617 7	0.235 4	1.000 0	1.000 0	1.000 0
20	望	见	0.613 7	0.227 4	1.000 0	1.000 0	1.000 0

表4是计算得到的排名前20的实例以及相关计算信息。

从结果来看,计算结果排名前20的组合大多数在本文的述补结构用法数据库中出现,且都是比较常用且接受度较高的典型述结式。除了能够使得真实述结式的计算结果比较靠前,还能获取到尚未被数据库收录的典型述结式,如表4标出的“砍掉、减掉、修通、绞尽”等,除了“绞尽”以外,其他三个毫无疑问都是比较常用且典型的述结式结构。

为了验证计算方法对未在“数据库”中出现的述结式的预测能力,抽取了排名前500且未在数据库中出现的结果,共176条数据。对这176条数据人工进行验证,共有85个能在实际语料中出现的述结式,这85条述结式在总体排名中的分布情况如图2所示。

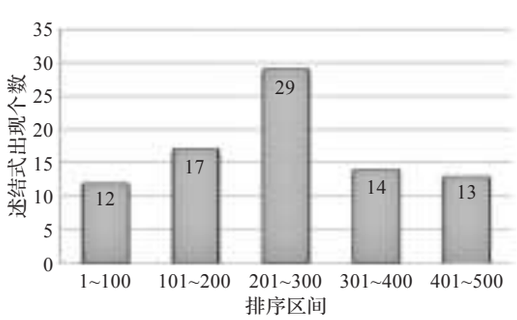

图2　排名前500且未被数据库收录的述结式的分布情况

从图2可以看出来,未被数据库收录的述结式主要出现在200到300这一排名区间。排名靠前和靠后的都相对较少。因为本文的“述补数据库”优先收录接受程度较高典型述结式,因此排名靠前的多是数据库中已收录的组合;而随着排名降低,作为述结式的可能性也就相应降低。以上两种因素结合,就出现了图2所示的“尖峰”现象,计算结果复合理论分析。

进一步地,统计排名前300的每个排序区间的未在数据库中出现的V1-V2组合作为述结式的准确率,结果如表5所示。

表5　不同排序区间下的述结式预测准确率

排序区间	V1-V2	述结式	准确率
1~100	20	12	0.600 000
101~200	31	17	0.548 387
201~300	34	29	0.852 941
总计	85	58	0.682 352

表5的数据显示,200~300排序区间内的预测准确率最高,达到了85.29%,进一步证明了本文计算方法的有效性。

4.3　分类测试实验

此部分考察述语或补语为实义动词和虚义动词时的计算效果,以便观察本文的计算公式是否对语义类别敏感,是否具有普适性。

实验按照语义类别共分为四组进行,分别为:

(1)实义动词+实义动词:洗-干净、洗-累、吃-干净、吃-累;

(2)实义动词+虚义动词:洗-好、洗-完、吃-好、吃-完;

(3)虚义动词+实义动词:搞-干净、搞-累、弄-干净、弄-累;

(4)虚义动词+虚义动词:搞-好、搞-完、弄-好、弄-完。

表6是上面的四组实例具体的计算结果。

表6 按“虚以/实义”划分的不同类别实例的计算结果

排序	类别	V1	V2	*Rd*	*ER*1	*ER*2	述语相似度(最大值)	补语相似度(最大值)
1	D	搞	好	0.519 9	0.039 8	1.000 0	1.000 0	1.000 0
2	B	吃	好	0.510 0	0.020 1	1.000 0	1.000 0	1.000 0
3	B	洗	好	0.509 7	0.019 4	1.000 0	1.000 0	1.000 0
4	D	弄	好	0.509 5	0.019 1	1.000 0	1.000 0	1.000 0
5	A	洗	干净	0.509 4	0.018 9	1.000 0	1.000 0	1.000 0
6	B	吃	完	0.509 4	0.018 8	1.000 0	1.000 0	1.000 0
7	B	洗	完	0.507 8	0.015 6	1.000 0	1.000 0	1.000 0
8	A	吃	干净	0.500 2	0.000 5	1.000 0	1.000 0	1.000 0
9	A	洗	累	0.500 2	0.000 4	1.000 0	1.000 0	1.000 0
10	C	搞	干净	0.456 9	0.000 2	0.913 6	1.000 0	0.827 2
11	D	弄	完	0.454 8	0.002 3	0.907 4	1.000 0	0.814 8
12	D	搞	完	0.444 7	0.001 6	0.887 9	1.000 0	0.775 8
13	C	搞	累	0.443 6	0.000 3	0.887 0	1.000 0	0.774 0
14	C	弄	干净	0.381 5	0.001 7	0.761 3	0.615 3	0.907 4
15	A	吃	累	0.359 5	0.000 7	0.718 4	0.615 3	0.821 4
16	C	弄	累	0.347 6	0.000 6	0.694 7	0.615 3	0.774 0

从表6的结果来看,最终计算结果并没有因为“虚义/实义”的不同而出现明显的差异,并没有出现某一类组合集聚出现的情况。这说明本文的计算方式并不会因为谓词性成分语义类别的不同而出现差异,方法的普适性较高。

5 小结

为了对现有人工建立的词典进行扩展,提出了一种对V1-V2构成述结式的可能性的定量描述,以实现对V1-V2构成述结式的自动判定。该方法既有基于大规模语料的概率统计方法,也用到了现有语言知识资源,既保证了结果的准确性,又使得方法的使用范围较广。利用这套定量计算方法对100×100个V1-V2组合进行考察,得到了符合实验预期的结果。

本文给出了述结式复合事件语义距离计算的公式,但是公式还有很多细节部分需要改进。比如利用条件概率对复合事件“致使-结果”语义关系的近似描写,以及计算V1-V2组合与典型述结式事件相似度计算中用到的词语相似度计算方法,都需要对算法进行进一步的优化改进。这也是下一步的工作所要重点解决的问题。

参考文献:

[1] Li Yafei.On V-V compounds in Chinese[J].Natural Language and Linguistic Theory,1990(8):177-207.

[2] 黄锦章.行为类可能式V-R谓语句的逻辑结构与表层句法现象[J].语文研究,1993(2):57-62.

[3] 王红旗.动结式述补结构配价研究[C]//现代汉语配价语法研究.北京:北京大学出版社,1995:144-167.

[4] 郭锐.述结式的配价结构和成分的整合[M]//现代汉语配价语法研究.北京:北京大学出版社,1995:168-191.

[5] 郭锐.述结式的论元结构[C]//汉语语法研究的新拓展(一)——21世纪首届现代汉语国际研讨会论文集.杭州:浙江教育出版社,2002:169-186.

[6] 袁毓林.述结式配价的控制——还原分析[J].中国语文,2001(5):399-479.

[7] 施春宏.动结式论元结构的整合过程及相关问题[J].世界汉语教学,2005(1):5-21.

[8] 施春宏.动结式的配价层级及其歧价现象[J].语言教学与研究,2006(4):45-57.

[9] 施春宏.动结式致事的类型、语义性质及其句法表现[J].世界汉语教学,2007(2):21-39.

[10] 施春宏.汉语动结式的句法语义研究[M].北京:北京语言大学出版社,2008.

[11] 宋文辉.补语的语义指向为动词的动结式的配价[J].河北师范大学学报,2004,27(3):94-99.

[12] 宋文辉.再论现代汉语动结式的句法核心[J].现代外语,2004,27(2):163-172.

[13] 宋文辉.现代汉语动结式的认知研究[M].北京:北京大学出版社,2007.

[14] 詹卫东.复合事件的语义结构与现代汉语述结式的成立条件分析[J].对外汉语研究,2013(1).

[15] 刘群,李素建.基于《知网》的词汇语义相似度计算[J].中文计算语言学期刊,2002,7(2):59-76.