基于 MC-Value 的非句蜕广义对象语义块的边界识别

臧翰芬 [1,2]

中国科学院研究生院 [1]

中国科学院声学研究所 [2]

E-mail: zanghf@mail.ioa.ac.cn

摘 要：多词语单元识别问题是目前计算语言学的研究热点之一。多词语单元是一个描述相对完整的若干词语组合，包括固定或半固定搭配。在语义块分析系统中，本文采用修改后的 C-Value 方法自动识别汉语非句蜕广义对象语义块，以词语、词性为特征自动识别非句蜕广义对象语义块边界，探讨了识别的统一度（Unithood）原则。在计算机辅助下，得到了非句蜕广义对象语义块的参考答案集，并进行了小范围的封闭测试和开放测试，封闭测试的正确率和召回率分别为 75.31%和 70.5%，开放测试的准确率和召回率分别为 73.44%和 60.5%。

关键词：C-Value；语义块；句蜕；广义对象语义块

The Research of Automatic Non-ecdysis Chunks Boundary Recognition based-on the MC-Value

Zang Hanfen [1,2]

Graduate School of Chinese Academy of Sciences [1]

Institute of Acoustics, Chinese Academy of Sciences [2]

E-mail: zanghf@mail.ioa.ac.cn

Abstract: The problem of multi-word unit is one of the key researches in the area of computational linguistics, which can form a complete description of certain terms used in combination units, including fixed or semi-fixed combination. In the system of Semantic Chunks Analysis, this paper presents a modified C-Value method to recognize automatically non-Ecdysis GBKm, taking the words and their part-of-speech as the characteristics of recognizing a GBKm. Two concepts are discussed in this paper, one is the unithood, another is the termhood. Under the coordination with computer, we get a candidate sets for GBKm. We show close test and open test. And the precision rate and recall rate are 75.31%, 70.5% and 73.44%, 60.5% respectively.

Keywords: C-Value, Chunks, Sentences-Ecdysis, GBKm

1 引言

近些年来，语义块的自动识别和标注已经在自然语言处理领域中成为了一个热门话题。组块分析作为一种预处理手段，可以大大降低短语划分和短语分析处理的复杂性，为进一步对句子进行深层次分析提供了基础,使句法分析在某种程度上得到简化，对机器翻译、信息提取、信息检索、专有名词识别、邮件自动回复等都具有非常重要的意义。从2000年开始，与块抽取的几个相关工作，如文本中的块、子句识别、语义角色标注都被CoNLL会议确定为主要评估任务。许多学者也在“块”的获取和分析上做了卓有成效的工作，他们认为块识别是浅层的句法分析技术。

早在 1991 年 Abney[1]就从心理学的角度首先提出了块（Chunk）的概念（1991）。周强[2]等比较系统地研究了汉语的成分组块和边界的识别问题，并利用基本规则和扩展规则、4 种依存关系

和三种拓扑结构，通过词汇信息和语境信息实现对多词语块的自动识别，取得了比较好的效果。有的学者把中文组块的识别问题看成是一个分类问题，并提出了基于 SVM[3]的组块识别算法，实验结果表明 SVM 算法取得了比 HMM 更好的分类效果尤其是小样本的情况下。还有学者[4,5]基于条件随机场来抽取汉语块，在试验中取得了较好的效果。本文采用术语识别的技术研究语义块的边界识别问题，因为术语是一个概念整体，语义块也是一个概念整体。Damerau[6]在候选词语中用词语的 MI（互信息）作为权重去测量术语的结合性问题。日本学者[7]认为“术语度”应该包含 2 个方面，一个是术语的频次；另一个是术语空间。他使用简单的二元名词去评价术语空间，用四种公式（即 GM(几何均值)、FGM(Frequency and GM)、MC-Value（修改的 C 值）和 F（Frequency）进行评分，最后得出用 FGM 评分效果最好等结论。Frantzi[8]采用 C-Value 的方法从名词短语、形名短语和介词短语中识别术语，并加入了语境的信息，取得了较好的效果。

本文主要是基于 HNC 理论研究非句蜕广义对象语义块的自动识别问题。HNC 理论全称为概念层次网络（Hierarchical Network of Concepts）理论[9]。HNC 把句子中的语义块分为三类：广义对象语义块（GBKm）（m=0...m）、特征语义块(EK)和辅语义块（fK）。所谓非句蜕广义对象语义块（Non-ecdysis GBKm）是指不包含有动态概念，且不含有句蜕和块扩的广义对象语义块，即大多数情况下是广义的静态概念。所谓广义的静态概念是 HNC 定义的 g（静态）、r（效应）、z（值）、u（属性），加上 v（动态）则称为五元组。HNC 理论不同于传统的语言学，它没有词性的概念，但有可用于描述词性的概念类别。在本文中使用的概念类别除了五元组以外，还包含 HNC 定义的一些工程上的概念类别，具体的概念类别符号集，如表 1 所示。

表 1　概念类别一览表

序号	概念类别	序号	概念类别	序号	概念类别
1	g	12	19	23	pw
2	u	13	h	24	pj2
3	r	14	s	25	pe
4	ug	15	jw	26	fp
5	gu	16	f	27	l
6	z	17	zz	28	jx
7	uv	18	fpe	29	x
8	uu	19	gw	30	fpwj2
9	j3	20	p	31	pj01
10	wj3	21	w	32	rw
11	fg	22	j1	33	wj01

由于词语的多句类代码、多概念类别等原因，计算机并不能很好地自动识别出 fK、EK 与 GBKm 的边界。为了更好地进行语义块边界的识别，本文运用 HNC 理论中的概念类别（相当于语言学中的词性），使用基于 C-Value 的统计方法加上规则来处理非句蜕广义对象语义块的边界识别，并给出相应的实验结果，以验证术语识别的方法是否适用于语义块的边界识别。由于不同学者对于块的类别的划分有不同的划分标准以及语料使用的不同，因此实验结果还无法与别的块识别结果进行比较。

2、C-Value 及其算法实现

C-Value 原来多用于“术语”的抽取，主要是对于词语频率统计的改进，修正了嵌套块的识别方法，提高了准确率。下面首先介绍 C-Value 的理论，其次说明在计算机上的实现，最后通过一个例子来进行具体详细的说明。

2.1 C-Value 理论

对于多词语自动抽取问题，C-Value 是与领域无关的。它主要关注提高嵌套短语的抽取的准确率。这个方法以一个元语言数据库作为输入，产生一个多词语候选集。C-Value 方法既包含语言学知识也包含统计学的信息，而更强调后者。语言学信息包含语料库的概念类别（词性）的标注，语法的筛选器限制抽取的候选集合的类型以及一些停用词的使用。统计部分包含统计的特征的候选字符串。本文所用的 C-Value 方法，包含 2 个概念。一个是统一度（Unithood），所谓统一度是指短语句法搭配的粘合度或内聚力，多词语出现频次的方差越小，词语之间的搭配越稳定，即短语的内部框架比词语更加稳定，反之亦然。比如一个词语“安全”和另一个词语“网络”，两个词语有不同的含义，但是如果给定了短语或者 GBKm“网络安全”或者“安全（的）网络”，就固定了其一个含义，因为“安全”是“网络”的一个属性概念（u;g;ug），其意思就被固定住了。另一个是 Termhood（术语度），是指语言学单位与特定领域概念相关的程度。C-Value 更加强调统一度（Unithood），对于抽取复合名词短语较有效。对于一个非句蜕广义对象语义块 GBkm，定义（GBKm）的统一度（Unithood）Z 如下：

$$Z = \left(\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \frac{1}{N}\left[\sum x_i^2 - \frac{1}{N}(\sum x_i)^2\right]\right) \qquad (1\text{-}1)$$

x_i 是非句蜕广义对象语义块中第 i 个词语出现的频次，

$\bar{x}$ 是一个非句蜕广义对象语义块中词语出现的平均频率，

N 是一个非句蜕广义对象语义块中所含词语的数目

C-Value 统计方法由如下变量所决定：

1、候选 GBKm 在语料库中出现的频率；

2、作为其他更长候选 GBKm 的频次；

3、更长候选 GBKm 的个数；

4、候选 GBKm 的中词语的个数。

对于一个短语（GBKm），本文只探讨一个语义块含 2 个或 2 个以上词语的构成情况。并且 GBKm 的构成要同时考虑两个数值，一个是 C-Value 值，另一个是 Z 值，Z 值取（0<Z<1）。只有它们同时被满足了，我们才说 GBKm 识别具有了统一度（Unithood）。任何一句话只要出现，就有其存在的合理性，因此句子中的 GBKm 有“存在即合理”的原则，即使是只出现 1 次。对于没有出现的词语（由于语料库中词语共现稀疏的现象），采用了简单平滑的处理方法。

C-Value 的计算区分两种情况。

1、如果“a”串是一个最大长度的字符串或者还没有被发现是嵌套的情况，那么它的 GBKm 的统一度是在语料库中全部的频次和其长度的结果。

2、如果 a 是其它任意一个更短长度的字符串，那么必须考虑它是否是其它更长候选 GBKm 的一部分；如果它出现作为更长候选 GBKm 的一部分，那么它的统一度将考虑作为

一个内嵌的 GBKm 的频次，以及这些更长候选 GBKm 的个数。

修改后的 C-Value 的计算公式如下：

$$MC-value(a)=\begin{cases}\log_2|a|*f(a)*Z & a\text{是不嵌套的} \\ \log_2|a|(f(a)*Z-\frac{1}{P(T_a)}\sum_{b\in T_a} f(b)) & otherwise\end{cases} \quad (1\text{-}2)$$

在公式（1-2）中，“a”是待判断是否为 GBKm 的字符串；f（.）是其在语料库中出现的频次，Ta 是包含“a”的候选 GBKm 的集合，P(Ta)是这些候选 GBKm 的全部数量，b 是包含 a 的更长的字符串，f(b)是它出现的频次。

在程序实现中，C-value 的阈值取 0。即如果候选字符串 a 的 C-Value 值大于 0，那么就确认它是 GBKm（非句蜕广义对象语义块）。C-Value 值为 0 的候选 a 只能作为嵌入另一个更长候选集中的子字符串。

2.2 算法

这一部分介绍通过 C-Value 方法识别 GBKm 的处理步骤。

首先，标注语料库，需要自动分词和标注概念类别，使用人工筛选的概念类别来限制多词语的抽取类型。由于只研究非句蜕广义对象语义块，因此，标注为动词的全部被排除或进行了体词化处理（即人工校正），这样就得到了 GBKm 候选集。

其次，对于用户输入的句子进行分词和概念类别的自动标注，形成以“+”连接概念类别来表示的句子，其中可能有多个字符串候选集合 a 形成 GBKm。

最后，以 GBKm 候选集为标准在句子中抽取相应概念类别的候选字符串 a，并获得其在候选集中出现的频次，然后计算嵌套和非嵌套的 C-Value 值，进行排序，找出大于阈值的候选集合。最后找到词性对应的词语，用“[”和“]标注出左右边界，这些左右边界就是非句蜕广义对象语义块的边界。

2.3 一个简单的例子

下面给出一个来自语料库真实的例子来具体说明一下 C-Value 是如何工作的。语料库是人工标注过 GBKm 的语料库（206360 个词语）。假设用户在对话框中输入一句话，如“*通过 6 年多的大规模建设，我国的基础设施已有很大改观，能够为民间资本进入生产领域提供相对良好的硬件环境。*”之后，系统将进行自动分词，并自动标注概念类别符号；然后人工进行校正，保证分词无误和概念类别的正确。把整句的概念类别以“+”串接起来，结果为“ 113+j3+wj10+u+141+ug+g+143-1++pj2+141+g+pw+u+uu+q+v+143-1+luv+117+pj01+z+v+ug+wj2+v+uu+gu+141+w9 +g +143-5”。然后，到人工已经标注过的“参考答案”中进行匹配，如果部分匹配成功，则作为候选存入一个数组中。最后，对数组排序，最长的串放于数组最上面，最短的放最下面。计算数组中候选的 C-Value 值，如果大于阈值 0，就以“[”和“]”标出候选的左右边界；否则，候选就不是 GBKm。如果两个串都可以作为 GBKm，一个串包含另一个串时取长度大的串作为 GBKm；两个串互相交叉时，取 C-Value 值大的那个作为 GBKm。

这个处理过程是以最长的字符串开始。从下面的表 2 可以看出，最长的字符串是“我国 的 基础 设施”，因为它是最长的，没有被嵌套的情况，因此用公式（2-1）计算 C-Value 值即可。“我国 的 基础 设施”就是候选字符串 a。在这个公式中，f(a)是 a 在语料库中出现的频次，所以 a 的 C-Value(我国 的 基础 设施)=$\log_2 4 * 1$ =2。Z 值根据公式（1-1）计算而得到的，只有满足 Z 值很小，且 C-Value 值大于 0 时，才能得出词语组合的边界的信息。

$$C-value\ (a)=\log_2|a|*f(a)*Z \qquad (2\text{-}1)$$

表 2 匹配到的候选集

序号	词语组合	候选组合	下标	频次
1	我国 的 基础 设施	pj2+l41+g+pw	8 9 10 11	1
2	我国 的 基础	pj2+l41+g	8 9 10	1
3	大规模 建设	ug+g	5 6 //22 23	16
4	民间 资本	pj01+z	19 20	1
6	6 年	j3+wj10	1 2	1
7	多 的	u+l41	3 4	2
8	的 基础	l41+g	9 10	1

下面计算更短字符串的C-Value值，即“我国 的 基础”，这3个词语只在一个更长的候选短语中出现，即“我国 的 基础 设施”，因此，其C-Value值按照公式（2-2）计算。其它的两个词语组成的 GBKm，如“大规模 建设”、“民间 资本”等，根据其在语料库中被嵌套的情况，按照（2-1）或（2-2）计算。

$$C-Value\ (a)=\log_2|a|*Z*(f(a)-\frac{1}{p(T_a)}\sum_{b\in T_a}f(b)) \qquad (2\text{-}2)$$

表 3 从输入的句子中抽取出来的候选集

序号	频次	词语组合	词语索引
1	1	我国 的 基础 设施	8 9 10 11
2	1	我国 的 基础	8 9 10
4	1	的 基础	9 10

表3 列出了一个GBKm候选出现嵌套的情况。公式（2-2）就是用于计算嵌套情形的，“a”是嵌套在另外一个的候选 GBKm 的候选字符串，即“我国 的 基础”，f(a)是它在语料库中出现的频次，本例中其出现“1”次。Ta 是候选集中包含它的更长的词语组合的集合，本例中只有“我国 的 基础 设施”。p(Ta)是集合中词语组合的数目，在本例中是1。b 是包含 a 的在 Ta 中的某个词语组合。f(b) 是 b 在语料库中出现频次，在本例中是1。$\sum_{b\in T_a}f(b)$ 是 Ta 中所有 b 的出现频次的总和，在这个例子中也是 1。用得到的参数值带入公式（2-2）中，得到 C-Value(我国 的 基础)=$\log_2 3*(1-1)=0$。

另外1个更短词语组合（“的 基础”）的C-Value值，也用公式（2-2）计算如下：C-Value(的9基础10)=$\log_2 2*(1-1)=0$。

从本例可知，词语组合的 C-Value 大于阈值 0 的只有“我国 的 基础 设施”一个，即所对应的词语在句子中的索引是“8 9 10 11”。最后，在句子中就用“[”和“]”标注出来。

3、实验数据

从人工标注过GBKm的7,000多个句子中选取了492个句子，把句子中的808个GBKm（非句蜕广义对象语义块）作为封闭测试的试验数据。机器对句子进行分词和概念类别的自动标注，然后人工校对词语概念类别的标注结果，以人标注过的GBKm中的词语概念类别组合为标准，获得

句子中的GBKm候选，分别计算这些候选的C-Value值，对C-Value值大于阈值的候选进行去重、排除交叉等处理后，就得到句子中GBKm的自动识别结果。最后，由人对GBKm识别结果进行正确与否的判断，得到封闭测试GBKm自动识别的正确率、召回率和F值如表4所示。此外，从封闭测试集以外随机选取183个GBKm对应的105句作为开放测试集，用同样的方法识别句子中的GBKm，正确率、召回率和F值如表5所示。其中，准确率、召回率和F值的定义如下：

GBKm边界识别的准确率：P=GBKm边界识别正确的数目/识别出的GBKm总数

GBKm边界识别的召回率：R=GBKm边界识别正确的数目/人工标注的GBKm总数

GBKm边界识别的F值：$F = \frac{2PR}{P+R}$

表 4 封闭测试结果

句子数		正确率	召回率	F 值
492	加规则前	60.35%	58.22%	59.263%
	加规则后	75.31%	70.5%	72.83%

表 5 开放测试结果

句子数		正确率	召回率	F 值
105	加规则前	49.4%	43.25%	46.12%
	加规则后	73.44%	60.5%	66.35%

在识别出的 GBKm 中，发现不少错误可以使用一些规则进行改正，得到更加准确的识别结果，这些规则如下：

规则 1：如果识别出的两个块之间出现单个“的”字，则把这两个块和“的”字合并为一个块。

规则 2：如果标识出的 GBKm 的块首或块尾是“的”字，那么该 GBKm 与前面或后面的块合并为一个块；前面或后面不是块而是词语时，可以往前或往后合并成一个块，直到动态概念“v”、特定词语（如“了”、“着”、“过”、“于”、“下”）、或标点符号为止。

例 1：“往往不愿经营亏损[的业务]r”，在这句话中，只部分识别对了语义块的边界，应该把“的”前面的词语加进来，这样[亏损 的 业务]识别就正确完成了；

*例 2：“这样，既保证了普遍[服务目标的]实现 r，同时也保证了有效竞争的开展”。*在例 2 中，出现了“了”字，“普遍”的概念类别是“u,ug”。因此，前边界应该到“普遍”，而后边界用到规则 1 中的“的”字规则，划到标点符号处，前后边界识别完毕。

规则 3：如果在标识出的 GBKm 中的数量概念、方位概念或指代动量概念的后面没有接应的 g 类或 r 类等概念，那么把 GBKm 和后面的概念合并为一个块；

*例 3：“相对而言，[后一项]改革比较缓慢，至今未取得[实质性进展]r，其结果是在[行业管理]r 上出现‘越位’”。*在例 3 中，正确标注了[后一项]，但是应该继续标注到“改革”，才标注完毕。根据语法，这里的数量短语“后/j002 一/j3 项/zz”后面应该还有一个词语，因此把“改革”合并，语义块边界识别完毕。

规则 4：如果标识出语义块前面是 u、ug 等概念类别，那么把标识出来的块和 u 类和 ug 类概念合并，直到动态概念为止。

*例 4：对于我国而言，大力发展[循环经济]r，是走新型[工业化道路的题]r 中应有之义。*在例 4 中，“新型”的概念类别是 ug，“走”是动态概念，应该把“[“划到“走”后面，语义块的边界识别完毕。

规则5：如果标识出语义块前面或后面有“和”字，那么把识别出来的块与“和”字前面或后面的词语合并，直到动态概念或标点符号为止。

加入这些规则后，正确率大大提高了，在开放测试中从不到50%增长到了73.44%。这充分说明：统计方法(C-Value)加规则方法对于识别非句蜕广义对象语义块还是十分奏效的。上述规则也可以综合运用（如例2），从而达到更好的识别效果。

4、结论

本文用修改后的 C-Value 统计方法对句子中的非句蜕广义对象语义块（GBKm）进行了自动识别，虽然 C-Value 方法主要用于术语的识别，而术语可能只是非句蜕广义对象语义块 GBKm 的一部分，但封闭测试和开放测试的结果表明，用 C-Value 作为非句蜕广义对象语义块的识别方法还是有意义的。通过对识别错误的 GBKm 进行分析，发现错误的原因主要由于“的”、“和”、“、”这些词语或标点符号引起的。因此，加入了5条规则修正这些错误，使 GBKm 的识别正确率大幅提高。开放测试的结果加规则前不太理想，准确率只有不到 50%，加入规则后则提高到了73.44%。还有许多错误加入的5条规则还无法改正，例如“*老百姓编起顺口溜夸赞道：“陈司令，陈司令，他和人民心连心*”，规则处理后识别出的 GBKm 为“[他和人民心]”，识别结果还是错误的。在这个句子中，“人民”的概念类别是 p，“心”的概念类别是 g，由于“候选集”中有此组合模式，因此进行了组合作为候选。其实正确的 GBK 是“他和人民”，“心连心”是 EK。因此，GBKm 的自动不仅要考虑内部的概念组合，同时还要考虑上下文的概念类别。

还有一些错误是辅语义块的识别问题，如时间短语等。辅语义块的识别问题可以参见另外一篇文献[10]。如果加上辅语义块的识别规则和算法的话，其准确率和召回率应该还会有所提高。下一步的工作是继续扩大语料的规模，加入更多的特征和找寻更好的计算词语粘合度的权重的计算公式，丰富和完善规则等，使非句蜕广义对象语义块的边界识别更准确。

参考文献

[1] Steven Abney. Partial Parsing via Finite-State Cscades. 1996.

[2] 周强等,孙茂松,黄昌宁.汉语句子的组块分析体系.计算机学报. 1999, 22(11):1158-1165.

[3] 李珩,朱靖波,姚天顺.基于 SVM 的中文组块分析.中文信息学报. 2004, 18(2):1-7.

[4] 秦颖,王小捷,钟义信. 级联中文组块识别. 北京邮电大学学报. 2008, 31(1):14-17.

[5] Dongfeng CAI, Xin LIU,et al. Chinese Maximal Noun Phrase Parsing Based on Cascaded Conditional Random Fields. Journal of Chinese Information Processing. 2008.

[6] 马艳军, 刘颖. 基于隐马尔科夫模型和候选排序的汉语基本名词短语识别. 全国第八届计算语言学联合学术会议(JSCL-2005). 2005.

[7] Nakagawa H.; Mori T. Automatic term recognition based on statistics of compound nouns and their components. Terminology, 2003, 9(2): 201-219.

[8] Katerina Frantzi, Sophia Ananiadou and Hideki Mima. Automatic Recognition of Multi-Word Terms: the C-Value/NC-Value Method. International Journal on Digital Libraries, 2000, 3(2):115-130.

[9] 黄曾阳. 概念层次网络（HNC）理论. 北京:清华大学出版社. 1998.

[10] 臧翰芬,韦向峰等.基于 HNC 理论的汉语辅语义块自动辨识研究.微计算机应用.2009, 30(11):48-54.