文章编号：1003-0077(2007)01-0009-08

面向机器辅助翻译的汉语语块自动抽取研究

姜柄圭，张秦龙，谌贻荣，常宝宝

（北京大学 计算语言学研究所，北京 100871）

摘 要：本文提出了一种统计和规则相结合的语块抽取方法。本文使用 Nagao 串频统计算法进行基于词语的串频统计，进一步分别利用统计方法、语块边界过滤规则对 2-gram 到 10-gram 语块进行过滤，得到候选语块，取得了令人满意的结果。通过实验发现，在统计方法中互信息和信息熵相结合的方法较单一的互信息方法好；在语块边界规则过滤方法中语块左右边界规则和停用词对语块抽取的结果有较大影响。实验结果表明统计和过滤规则相结合的方法要优于纯粹的统计方法。应用本文方法，再辅以人工校对，可以方便地获取重复出现的多词语块。在机器辅助翻译系统中，使用现有的语块抽取方法抽取重复的语言单位，就可以方便地建设翻译记忆库，提高翻译的工作效率。

关键词：人工智能；机器翻译；语块抽取；串频统计；内部结合紧密度；信息熵；语块组合规则

中图分类号：TP391 **文献标识码**：A

Chinese Multi-word Chunks Extraction for Computer Aided Translation

Kang Byeong-Kwu, ZHANG Qin-long, CHEN Yi-rong, CHANG Bao-bao

(The Institute of Computational Linguistics, Peking University, Beijing 100871, China)

Abstract: This paper suggests a methodology which is aimed to extract multi word chunks for translation purposes. Our basic idea is to use a hybrid method which combines the statistical method and linguistic rules. The extraction system used in our work operated at four steps: (1) Tokenization of Chinese corpus; (2) Extraction of multi-word chunks(2-gram to 10-gram) using Nagao's Algorithm and Substring Reduction Algorithm; (3)Statistical Filtering which combines Mutual Information (or Log-likelihood Ratio) and Left/Right Entropy; (4) Linguistic filtering by chunk formation rules and stop-word list. As a result, hybrid method proved to be a suitable method for selecting multi-word chunks, it has considerably improved the precision of the extraction which is much higher than that of purely statistical method. We believe that multi-word chunks extracted in this way could be used effectively to supplement existing translation memory database.

Key words: artificial intellgence; machine translation; chunk; Nagao's algorithm; M. I; log-likelihood; entropy; chunk formation rules

1 引言

机器辅助翻译的重要思想是处理重复性的语言现象。为了处理重复的语言现象，首先要找出翻译文本中重复出现的语言单位。根据语言单位的大小，重复出现的语言单位分为句子层、短语层、词语层。其中，本文主要关心的是重复出现的短语，包括名词短语、动词短语、形容词短语、数量短语等等。我们不妨把这些短语称作多词组合的语块(multi-word chunk)。这些多词语块不仅是一个比较完整的翻译单位，而且重复出现的频率也相当高，可以弥补句子一级匹配技术的不足。

在翻译过程中，单词可能有很多义项，有时候很难找到合适的对译词。但是，多词语块作为一个整体来进行翻译，歧义现象比较少，找出相应的译文也

收稿日期：2006-01-14 **定稿日期**：2006-04-05

基金项目：国家 973 资助项目(2004CB318102)；国家 863 计划资助项目(2001AA114210；2002AA117010)

作者简介：姜柄圭(1973—)，男，博士生，主要研究方向为汉韩机器辅助翻译。

容易多了。例如,汉语词“打”,查词典,有很多不同的意思。取出语块“打乒乓球”、“打个电话”、“打一件毛衣”,那它们都只有一个意思十分明确。又如,汉语语块“酝酿着新的突破”、“取得突破性进展”,比较好的韩文翻译是:

“酝酿着新的突破”—“새로운 돌파구를 모색하다”(do something with a promising breakthrough)

“取得突破性进展”—“획기적인 진전을 이루다”(have made a significant breakthrough)

如果按词翻译,那就不可能有正确的翻译结果。所以,基于语块的翻译方法有利于解决机器翻译中的歧义现象。而汉语语块的提取是基于语块方法的基本环节。

关于作为研究对象的多词语块,学者们有很多论述[9~11]。我们认为,以多少的语块作为自然语言处理系统的对象至少要应用目标、技术与理论的发展水平以及语言类型。本文中,语块抽取有比较明确的应用目标,那就是为机器辅助翻译系统为服务。机器辅助翻译的基本思想不是计算机自动翻译,而是把重复出现的语块保存下来,给出参考译文。如果按照语块进行自动翻译,所有的句子都应该分析为不同类型的语块。但目前的技术还不容易驾驭数不清的语块。甚至连确定语块的边界都有困难。针对机器辅助翻译的特点,我们并不彻底分析所有的语块,而是抽取文本中比较固定的、重复出现的语块。这些语块不局限于词语的长度,有的语块可以由两个词组合(2-gram),有的语块甚至是由10个词(10-gram)以上组合的。抽取语块时,只要在句子中稳定共现、语义比较完整,就把它当作一个合法的语块。在我们的语块抽取中,没有涉及到语块的类型识别问题。我们更关心的是,语块的重复性。以机器辅助翻译的眼光看,出现十次的语块比只出现一次的语块还有用。因为重复的语言现象频繁出现,翻译系统的优势会更加明显。

2 语块抽取的基本方法

如果以“抽取重复而固定的语言单位”这样的模式来审视现有的研究格局,中文语块抽取研究已经有了一定基础。相比之下,二元(bi-gram)或三元(tri-gram)语言单位的研究成果较为丰富,4-gram、5-gram、10-gram等多词单位的研究则显得相对薄弱一些。不过,有关多词单位的抽取工作实际上一直都有学者在尝试。Nagao & Mori[1]曾经提出过一个N-gram串频统计算法。他们的算法不仅降低了时间复杂度(其复杂度为$O(n\mathrm{Log}n)$),而且提高了抽取多元(N-gram)组合的效率。这种串频统计方法已经在自然语言处理领域中广泛应用,包括新词抽取、专业术语的抽取、固定搭配的抽取等等。在面向中文信息处理的研究工作中,吕学强和张乐[2]利用Nagao的N-gram统计算法,在大规模汉语语料中进行抽取语块的实验。他们在论文中还提出一个删除同频子串的算法,提高了语块抽取的准确率。谌贻荣[8]在术语抽取工作中用到串频统计的算法,抽取候选术语,然后利用内部结合紧密度和最大熵模型结合的统计值,对这些候选术语进行过滤和排序[8]。本文也在Nagao的串频统计方法的基础上开展语块抽取的工作。有别于以往的抽取主要是基于字的方法(Character Based Method),本文的研究工作是基于词语的方法(Word Based Method)。

在抽取多词语块时,串频统计的方法虽然可以作为简单而有效的方法,但是可能引起准确率不高的问题。因为过去的串频统计方法只用字符串的频度信息,在抽取结果中包含很多不合法的字符串。为了弥补字符串统计方法的缺点,很多人提出过不同的统计关联度度量方法。这些统计关联度计算方法大致分为两种。第一种方法跟语块内部的结合紧密度有关。我们可以把这些方法称作“内部方法(Internal Measures)”[6~8],常用的统计值包括:互信息(MI)、x^2统计值、对数可能性分值(Log-likelihood)等。第二种方法跟语块的独立性有关,也就是候选语块到底有多大的独立性,跟其他词语有没有清晰的界限等等。我们不妨把它们称作“外部边界方法(External Measures)”。从统计模型的眼光看,目前比较常用的方法是基于最大熵的方法(Maximum Entropy Model)[7,8]。鉴于前人的研究成果,本文研究工作将根据几个统计关联度公式和最大熵方法,判别候选语块的结合紧密度和独立性。

最后,我们按照语言学规则和停用词(Stop Word)来进行候选语块的过滤(filtering)。如果候选语块的出现频度很低,统计方法不容易判断语块的合法性。比如,“避免了发达国家早期探索所走过的一些弯路”、“双宾语前一个是体词性成分”都是在测试语料中只出现两次的词语串,有数据稀疏的问题,统计方法不容易判定这两个语块的合法性。这时,根据汉语语块的构成原则,可以判定前一个语块是合适的,而后者却不符合汉语的语法规则。本文工作中语言学规则的作用主要在于判定语块的边界。

综上所述，本文工作走的是一条多种统计信息和语言规则相结合的路线。图 1 给出中文语块抽取的流程：

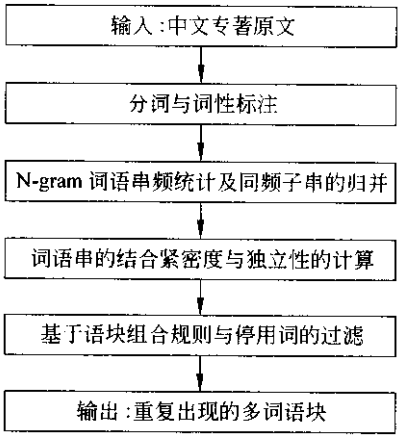

图 1　中文语块抽取的流程

3　候选词语串的抽取

（1）基于词语的串频统计方法

词语串频统计的基本思路是对原始文本进行切分，然后对切分后的文本使用 Nagao 算法来进行频率统计。Nagao 的 N-gram 统计算法，最小的处理单位是字，比如："中"、"国"、"民"等汉字都看作是一个统计单位。基于字的串频统计方法可能有利于新词和专业术语的抽取。但是，在语块抽取的过程中，基于字的串频统计方法不一定是一个有效的方法。语块不是多个字组合的单位，而是多个单词组合的单位。因此，针对多词语块的抽取，更合适的计算单位是中文单词，比如："中国"、"国家"、"现代"等单词都看作是一个统计单位。

从多词语块抽取的实际需求出发，我们首先对中文科技专著进行分词和词性标注。以《现代汉语语法信息词典详解》（以后简称《详解》）作为测试语料，本文使用北京大学计算语言学研究所的词性标注系统，进行原文的分词和词性标注。例如：

随着/p　社会/n　生活/n　的/u　日益/d　信息化/v　，/w　人们/n　越来越/d　强烈/a　地/u　希望/v　用/p　自然/n　语言/n　同/p　计算机/n　交流/v　信息/n　。/w

串频统计过程中，Nagao 的 N-gram 算法主要有两个好处。第一，算法的空间复杂度比较低，可以减少内存消耗。假如一个测试语料有 5000 个单词，计算 10-gram 的串频统计时，通常需要 5000^{10} 个存储空间，数据量达到几百万 GB 的量级。但是，目前的计算机不能做出这么大的计算。与此相比，Nagao 的 N-gram 算法对 m 个字只需要 $7m(2m+4m+1m)$byte 的存储空间。比如，在处理 10Mbyte 的汉语语料时，该算法所需要的存储空间只有 70Mbyte。目前的计算机对这样的内存消耗一点问题都没有。第二，该算法的时间复杂度也比较低（复杂度为"$O(n\log n)$"），短时间内可以进行 N-gram（如，10-gram）的计算。比如，处理一本普通的中文科技专著时，10-gram 计算时间大概几十秒左右，通常不超过一分钟。另外 N-gram 的长度可以扩展到 255-gram（$1\leqslant n\leqslant 255$）。

作为语块抽取的第一个步骤，我们利用 Nagao 的 N-gram 算法抽取了 2 次以上出现的所有词语的组合。通过抽取结果的分析，我们可以发现大部分的词语串是在 10-gram 范围之内。在测试语料中，最长的语块是 14-gram 词语串，如，"汉语/句子/的/构造/原则/与/短语/的/构造/原则/基本上/是/一致/的/"一共 14 个词构成一个语块。但是，这些超长的语块很少，其所占的比例还达不到 0.1%。大部分的语块都在 10 个词以内。

（2）删除同频子串

N-gram 串频算法所抽取的词语串，除了合法的词语串以外，有很多垃圾信息。其中最显著的现象是大量重复的子串（Substring）问题。例如：

表 1　同频子串的情况

N-gram	左　子　串	右　子　串	频度
8-gram	不同的量词间用逗号隔开	不同的量词间用逗号隔开	8 次
7-gram	不同的量词间用逗号隔	的量词间用逗号隔开	8 次
6-gram	不同的量词间用逗号	量词间用逗号隔开	8 次
5-gram	不同的量词间用	间用逗号隔开	8 次
4-gram	不同的量词间	用逗号隔开	8 次

上面的词语串虽然其长度不一样，但实际上都是一个词语串的集合。其中，8-gram 是最长的词语串，而 7-gram 以下的词语串都是它的子串。从翻译的角度看，最好留下其中最长的一个词语串，而其他都被删除。同一频度的子串都可以看作是不必要的部分。吕学强等[2]在论文中，提出了一个删除同频子串的算法。该算法改进了已有的子串归并的算法。早期的算法有 $O(n^2)$ 的时间复杂度，而他们的算法只有 $O(n)$ 的复杂度。我们根据吕学强的算法，删除了同一频度的子串。通过子串归并以后，词语串的数量就大大减少了。表 2 给出子串归并以后的词语串数量：

表 2 大致反映了 N-gram 串频统计中同频子串的重复程度。子串归并之前，N-gram 词语串最多能产生“$(n-1)*2$”个子串。经过子串归并，除了最长的词语串外，其他的同频子串都被删除了。通过这种方法，我们从测试语料中得到将近 10 067 个候选语块。但是，最长的语块不一定是最合适的语块，有时候，这些语块是不符合语法的单位。比如，“的主要参考标准”等词语串都是不完整的语言单位。因此，我们还需要采取方法来过滤不合法的语块。

表 2 《详解》的切分结果

	2-gram	3-gram	4-gram	5-gram	6-gram	7-gram	8-gram	9-gram	>=10-gram
子串归并之前	7745	4649	2093	942	452	232	125	73	72
子串归并之后	5777	2702	975	353	125	63	25	10	18

4 基于统计的过滤方法

关于候选语块的过滤，我们可以考虑两种方法。一种方法是度量语块内部的结合紧密度。语块通常是一个完整的语言单位，词语之间一定有比较紧密的结合关系，经常同时出现。另一种方法是观察语块和语块周围语境的关系。一般来说，一个语块跟其他词汇之间有比较清晰的界限。本节从“内部结合紧密度计算方法(Internal Measures)”和“外部边界的判定方法(External Measures)”两个方面阐述基于统计的过滤方法。

(1) 语块内部结合紧密度计算(Internal Measures)

以统计学的眼光看，语块内部词语之间的结合紧密度主要取决于它们的共现频度。这种方法的基本假设是“某一个词语串的共现频度越高，词语串的结合紧密性越强”。根据这个假设，高频的词语串可能是一个完整的语块。这些方法，除了简单的共现频度方法以外，还有几种常用的关联度度量方法。其中本文使用了两种度量方法，即互信息(MI)和对数可能性分值(Log-likelihood)方法。表 3 中列举了这四种方法的计算公式。在表 3 中 XY 表示某两个词语或词语串；P_X 是词语 X 的出现频率和出现概率；P_{XY} 表示词语串 XY 的出现概率；$\overline{X}$ 表示词语 X 不出现的次数。计算方式如下所示：

表 3 内部结合紧密度计算方法

Method	Formula	Method	Formula
Mutual Information (MI)	$\log_2 \frac{P_{XY}}{P_X P_Y}$	Log-likelihood(Log-L)	$-2\log \frac{(P_X P_Y P_{\overline{X}} P_{\overline{Y}})^{f_Y}}{(P_{XY} P_{\overline{X}\overline{Y}})^{f_{XY}} (P_{X\overline{Y}} P_{\overline{X}Y})^{f_{XY}}}$

(2) 外部边界的判定方法(External Measures)

为了提取合法的语块，这里采用另外一种统计方法，即外部边界的判定方法。本文将使用最大熵的方法(Maximum Entropy Model)来判别候选语块的独立性和边界。从最大熵的角度看，语块的独立性计算可以分为左边界的信息熵和右边界的信息熵：

$$Le(W) = -\sum_{\forall a \in A} P(aW \mid W) \cdot \log_2 P(aW \mid W)$$

$$Re(W) = -\sum_{\forall b \in B} P(Wb \mid W) \cdot \log_2 P(Wb \mid W)$$

在上面的公式中，Le 和 Re 分别表示左边信息熵(Left Entropy)和右边信息熵(Right Entropy)；W 表示 N-gram 的词语串，$W=\{w_1, w_2 \cdots w_n\}$；$A$ 表示候选语块左边出现的所有词语的集合，a 表示左边出现的某一个词语；B 表示候选语块右边出现的所有词语的集合，b 表示右边出现的某一个词语。根据信息熵的理论，信息熵 Le 和 Re 的数值越大，

词语串 W 左右出现的词语越多，W 就更有可能是一个完整的语块。

(3) 内部结合度和外部边界信息熵相结合的方法

不管是内部结合紧密度方法，还是外部边界的信息熵方法，都有各自的可取之处。因此，如果把这两种方法结合起来，就能取长补短，更有效地抽取合法的语块。

内部结合紧密度的方法一般只用来度量两个成分之间的结合概率。但是，为了确定多词语块的结合度，不能局限于 bi-gram 内部的计算，需要考虑 3-gram 以上的情况。也就是说，二元的关联度度量方法需要扩展为 N-gram 的计算。关于二元关联度计算的扩展方法，比较典型的是互信息(MI)公式。Magerman & Marcus[6] 曾经提出过广义互信息(Generalized Mutual Information)的方法[6,8]。他们利用广义互信息的方法来确定 N-gram 的英文短语的边界。谌贻荣[8]把广义互信息的方法应用于中文术语识别工作。他的工作不仅简化了已有的广义互信息公式，而且改进了 N-gram 互信息的计算公式(称作 Connect Rate)。本文在这些研究成果的基础上进行二元统计方法的扩展。为了比较，本文除了互信息方法外，还要采用 Log-likelihood 方法来度量 N-gram 词语串。

关于三个词以上的词语串，我们先把它们分成两个成分来计算其概率。比如，3-gram 的词语串可以分为[ab]c(如"[语言/ 信息/] [处理/]")和 a[bc](如"[中国/] [科技/ 人员/]"这两种 2-gram 词语串。同样，N-gram 的词语串可以切分成两个部分。在切分一个词语串为两个部分时，一个 N-gram $(w_1..w_n)$内部可能有 $N-1$ 个切分点。这里 k 作为词语串中第 k 个词；第 k 个词的左边为 W_k；第 k 个词的右边为 W_{k+1}：

$$W(W_k,W_{k+1}) = \{W_k(w_1\cdots w_k),W_{k+1}(w_{k+1}\cdots w_n)\}$$

如果以互信息(MI)和 Log-likelihood 公式为例，对 $W(W_k,W_k)$的计算方法如下所示：

$$MI_n^k(W_k,W_{k+1}) = \log_2 \frac{P_{(W_k)(W_{k+1})}}{P_{(W_k)}P_{(W_{k+1})}}$$

$$\mathrm{Log}L_n^k(W_k,W_{k+1}) = -2\log \frac{(P_{(W_k)}P_{(W_{k+1})}P_{\overline{(W_k)}}P_{\overline{(W_{k+1})}})^{f_{(W_{k+1})}}}{(P_{(W)}P_{(\overline{W})})^{f_{(W)}}(P_{(W_k)}P_{(\overline{W_k})})^{f_{\overline{(W_k)}(W_{k+1})}}}$$

根据上面的计算公式，我们可以求每一个二元统计值，为了保证最坏情况不对我们的计算产生影响，我们选取其中分值最小的组合作为整个词语串的分值。这就成为内部结合紧密度的度量方法(Internal Measure)：

$$\mathrm{InternalMeasure}(W) = \min_{1\le k\le n-1}(MI_n^k(W_k,W_{k+1}))$$

$$\mathrm{InternalMeasure}(W) = \min_{1\le k\le n-1}(\mathrm{Log}L_n^k(W_k,W_{k+1}))$$

另外，我们还要考虑 N-gram 词语串的外部边界问题。在(2)中，我们采用最大熵模型，制定左边界的信息熵(Le)公式和右边界的信息熵(Re)公式。根据这个信息熵公式的基础上，把左边信息熵和右边信息熵结合起来，就可以推导一个外部边界的信息熵公式。在这个公式中，我们还考虑词语串的频度信息($F(W)$)。具体的计算方法如下所示：

$$ExternalMeasure(W) = \left(1-\frac{1}{F(W)}\right)\cdot\sqrt{(1-Le(W)\cdot(1-Re(W))}$$

最后，我们把内部公式和外部边界公式结合起来，形成一个综合的统计过滤公式(Unified Measure)：

$$UnifiedMeasure(W) = InternalMeasure(W)\cdot ExternalMeasure(W)$$

这种综合的统计过滤方法，针对 N-gram 词语串，就能抽取出很多合法的语块，并且效果比单纯的统计方法还好一些(关于抽取的准确率，请参照实验结果)。

5　基于语言规则的过滤方法

通过 N-gram 词语串的观察，我们可以发现高频的词语串不一定都是合法的。有时候高频的词语串并不成为一个合法的语块。比如："上并没有"是一个高频的词语串，从统计值的大小来看，这可能是一个有意义的组合。但是，从语言学的眼光看，这个词语串却不是一个完整的语块。

大部分的情况下，不合法的词语串跟语块边界问题密切相关。为了叙述的方便，我们可以把语块的边界分为左边界和右边界。在汉语语块中，有些词语不能位于语块的最左边，而有些词语不能位于语块的最右边。比如，如果汉语的方位词位于词语串的左边，很可能是一个不合法的词语串(如"中分化出来的")。又如，词语串的右边若以副词为结束，这可能是不完整的语块("这种花儿很")。对此，统计方法不能彻底排除这些不合法的词语。如果制定一个语块组合的规则，就能弥补统计方法的不足，排除不合法的词语串。

(1) 汉语语块的边界特点

汉语的语块通常是一种语法结构，是符合一定语法功能的短语。大部分的语块都有一个中心词，以中心词作为语块的开始或结束。在汉语的语块中，有些中心词位于语块的左边，而有些中心词位于右边。我们根据北大计算语言学研究所的词性标注规范，对25个词类进行考察它对语块边界的关系。

通过观察汉语词类跟语块边界的关系，语块左边能出现的词语为“动词、名词、形容词、副词、介词”等等。但是，“方位词、助词、语气词、量词”等词类一般不能位于语块的最左边。如果在词语串中有这些词类，那就可能是一个不合法的语块。另外，语块的最右边能出现的词语有“不及物动词、名词、形容词、方位词、助词”等等。但是，语块的最右边一般不能是“形式动词、能源动词、副词、区别词、数词、连词、前接成分”等词类。如果语块的最右边出现这些词类，就成为不完整的语块。需要指出的是，这些语块边界的特点是相对的，而不是一个绝对标准。我们可以根据文本类型和实际需要增补或改变这些规律。

(2) 基于语言学规则的过滤算法

鉴于汉语语块边界的特点，我们制定一个词语串过滤算法。具体的算法大致如下所示：

BEGIN

1. 读入一个N-gram词语串；
2. 如果当前词语串的最左边以方位词(f)、助词(u)、语气词(y)、量词(q)、后接成分(k)、连词(c)等词类开始，就删除该词语，输出其右边的词语串；
3. 如果当前词语串的最右边以形式动词(v)、能源动词(v)、副词(d)、区别词(b)、数词(m)、连词(c)、前接成分(h)等词类结束，就删除该词语，输出其左边的词语串；
4. 如果词语串左右没有所制定的词类，读入下一个N-gram词语串；
5. 反复执行2—4的命令；
6. 如果读完最后一个N-gram词语串，就退出；

END

6 实验以及结果分析

(1) 实验过程以及结果分析

语块抽取的实验数据选用《详解》中的语句。从中我们利用Nagao的N-gram统计方法抽取出16 383个词语串。然后对这些词语串进行同频子串的归并工作，删除同频的子串。通过子串归并以后，词语串的个数减到10 067个。根据词语的个数，词语串分为9组，即2-gram到10-gram词语串。最后，我们看统计方法和语言学规则方法对这1万个词语串过滤效果如何。

为了比较统计方法和语言学规则方法的作用，我们进行了4个不同的实验：

- 统计方法1(Internal Measure)：互信息(Mutual Information)
- 统计方法2(Unified Measure1)：Mutual Information + Left/Right Entropy
- 统计方法3(Unified Measure2)：Log-likelihood + Left/Right Entropy
- 统计方法和规则结合的方法：Unified Measure1+Rule(Left/Right Stop word)

我们逐个检查了词语串的处理结果，表4显示了处理结果的总体情况：

表4 语块抽取的准确率

N-gram	Mutual Information	UnifiedMeasure-1	UnifiedMeasure-2	Unified Measure-1+Rule	频度	阈值	语块数
2-gram	77.80%	86.80%	75.20%	94.20%	≥3	≥0.05	<3 000
3-gram	81.00%	81.60%	71.60%	93.40%	≥2	≥0.01	<2 000
4-gram	77.00%	82.40%	80.40%	92.60%	≥2	≥0.01	<1 000
5-gram	77.10%	79.60%	79.50%	90.50%	≥2	≥0.00	<500
6-gram	77.40%	81.30%	82.90%	89.10%	≥2	≥0.00	<300
7-gram	72.10%	80.30%	78.60%	90.10%	≥2	≥0.00	<150

续表

N-gram	Mutual Information	UnifiedMeasure-1	UnifiedMeasure-2	Unified Measure-1＋Rule	频度	阈值	语块数
8-gram	75.00%	80.00%	70.80%	88.90%	≥2	≥0.00	<100
9-gram	70.00%	70.00%	75.00%	77.80%	≥2	≥0.00	<50
≥10gram	66.70%	66.70%	66.70%	77.80%	≥2	≥0.00	<50

表 5　语块抽取的样例

N-gram	例　子	频度
2-gram	信息/ 处理	180
	直接/ 修饰	59
3-gram	词语/ 分类/ 体系	28
	完成/ 了/ 任务	15
4-gram	本/ 字段/ 就/ 填	98
	北京大学/ 计算/ 语言学/ 研究所	32
5-gram	现代/ 汉语/ 语法/ 信息/ 词典	146
	中国/ 标准/ 技术/ 开发/ 公司	5
6-gram	中文/ 信息/ 处理/ 应用/ 平台/ 工程	6
7-gram	理论/ 体系/ 和/ 计算/ 模型/ 的/ 探索	5
8-gram	不同/ 的/ 量词/ 间/ 用/ 逗号/ 隔/ 开	8
9-gram	汉语/ 的/ 词/ 分为/ 以下/ 18/ 个/ 基本/ 词类	2
≥10-gram	词/ 的/ 意义/ 不/ 能/ 作为/ 划分/ 词类/ 的/ 依据	2
	避免/ 了/ 发达/ 国家/ 早期/ 探索/ 所/ 走过/ 的/ 一些/ 弯路	2

实验结果表明，选取不同策略的过滤方法对语块抽取的影响是相当大的，表 4 中最好的实验结果要比最差的结果高出 15 个百分点。总体上看，混合方法的准确率比其他三个统计方法高。在统计过滤方法中，单用互信息的方法，其准确率比较低（约 75%）。如果互信息方法和信息熵方法结合起来，就能提高 5 个百分点。但是，Log-likelihood 分值和信息熵相结合的方法，其准确率没有像互信息方法那么高。通过这三种统计方法的比较，我们可以看出，互信息和信息熵相结合的方法得到比较好的效果（约 80%）。如果把统计方法和语言规则结合起来，就能得到更好的效果。在实验结果中，“Unified Measure1”和“Rule(Left/Right Stop word)”相结合的方法就更为突出，提高 10 个百分点的准确率（90%）。按照准确率的高低，这些实验结果可以作如下排序：

Unified Measure-1＋Rule ＞ Unified Measure-1＞ Unified Measure-2＞ Mutual Information

表 5 给出一些合法的多词语块样例。在 2-gram 到 5-gram 的语块中，有的语块在文本中出现几十次，甚至也有 100 次以上出现的语块。这些重复的语块在翻译过程中，作为一个整体来进行翻译，就会提高其效率。

（2）错误结果以及难点分析

大致说来，用统计方法和过滤规则来抽取多词组合的语块，效果还是比较好的。表 5 中的样例，语块长度大都在 10 个词（大约 20 个字）以内，基本都是合适的短语结构。但是，除了这些合适的语块以外，抽取结果中也有一些不合适的词语串。这些语块之所以造成结构不合适的问题，主要原因是现有的过滤规则缺乏有效的约束条件描述，即描述的语言知识是不充分的。现有的过滤规则虽然有助于解决语块边界上的一些常见的错误，但是不能判断出语块内部结构的合法性。例如：

＊a. 划分 词类 的 目的 是 把 语法

＊b. 继承 了 朱 德熙 先生 的

在人看来，上面的词语串的确有不合适的地方。a 中，“划分 词类 的 目的”是一个合适的名词短语，而后面的动词短语却不合适，其结构不完整。b 中，“继承 了”和“朱 德熙 先生 的”之间没有搭配关系，没有动词“继承”的宾语。a-c 中都是较长的词语串（5-gram 以上），这类的错误情形比较突出的。需要说明的是，在很多情况下，这些较长的词语串若切分成小语块，小语块本身就成为一个合适的短语，如“[体现 了] [词 与 词 之间 的]”。

在错误结果中，有的词语串虽然不是一个完整的语法单位，但是音节上经常读在一块。比如：“作者认为”、“需要指出的是”等词语串，从语法上看是不合适的，而从语音停顿上看，可以看作是一个单位。像“是”、“认为”这样的动词，其后面能搭配的成分性质非常多，在进行分析时很难确定边界。但在翻译过程中，这些词语串也可以作为一个有用的成分。这些不完整的主谓短语结构可以变换成一个翻译模板，用固定的翻译方式来进行翻译。

7 结语

本文提出了一种统计和规则相结合的语块抽取方法。本文在 Nagao 的 N-gram 串频统计算法的基础上，实现了基于词语的串频统计的方法。并且利用统计方法和语块边界规则来过滤 2-gram 到 10-gram 之间的候选语块，取得了比较令人满意的结果。在统计方法中，互信息和信息熵相结合的方法比互信息方法好一些。此外，语块左右边界规则和停用词对语块抽取的影响相当大。最好的方法还是统计和规则相结合的方法。实验结果表明混合的方法比单纯的统计方法好。

应用该方法，再辅以人工校对，可以方便地获取重复的多词语块。语块抽取的方法也可以应用于机器辅助翻译系统。在机器辅助翻译系统中，使用现有的语块抽取方法抽取重复的语言单位，就可以方便地建设翻译记忆库，提高翻译的工作效率。在下一步的研究中，除重复的语块外，还要抽取重复的语言结构，包括词性序列的组合、特殊句式等等。这些工作会有助于建造翻译的模板，提高语句匹配的效率。

参考文献：

[1] Makoto Nagao, Shinsuke Mori. A new method of N-gram statistics for large number of n and automatic extraction of words and phrases from large text data of Japanese [A]. In: Proceedings of ACL-1994 [C], 1994.

[2] Xueqiang Lv, Le Zhang and Junfeng Hu. Statistical Substring Reduction in Linear Time [A]. In: Proceedings of IJCNLP-2004 [C], 2004.

[3] Haodi Feng, Kang Chen, Xiaotie Deng, Weimin Zheng. Accessor Variety Criteria for Chinese Word Extraction. Computational Linguistics [J]. Vol. 30, 2004.

[4] Dias G., Guillor S. & Lopes J. G. P. 2000b, Combining Linguistics with Statistics for Multiword Term Extraction: A Fruitful Association?[A]. In: Proceedings of RIAO-2000 [C], France. 2000.

[5] Schone, P., Jurafsky D. Is knowledge-free induction of multiword unit dictionary headwords a solved problem?[A]. In: Proceedings of EMNLP[C] 2001.

[6] Magerman, D. & Marcus, M. Parsing a natural language using mutual information statistics [A]. In: Proceedings of AAAI '90 [C]. 984-989, 1990.

[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[J]. 中文信息学报，2003，17(3)：9-14.

[8] 谌贻荣. 中文术语自动提取技术研究[D]. 北京大学计算机系，硕士学位论文，2005.

[9] 李素建，刘群，杨志峰. 基于最大熵模型的组块分析[J]. 计算机学报，2003，(12).

[10] 周强. 汉语短语的自动划分和标注[J]. 中文信息学报，1997，11(1)：1-10.

[11] 张昱琪，周强. 汉语基本短语的自动识别[J]. 中文信息学报，2002，16(6)：1-8.

[12] 俞士汶. 现代汉语语法信息词典详解[M]. 北京：清华大学出版社. 2003.