基于语料库的汉语短语边界划分的研究

张国煊　郁梅　王小华

（杭州电子工业学院计算机系）

摘　要： 短语作为句子的一个层次，其结构的研究在自然语言理解，尤其是汉语理解中占有重要位置。有效的短语分析对降低其后句法分析的难度，缩小句法分析器的搜索空间是很有帮助的. 本文介绍的是一种利用汉语句子词性标记串信息，分析其短语结构、边界的方法。

A Approach of Parsing Chinese Phrase's Boundaries Based on Corpus

Zhang Guoxuan　Yu Mei　Wang Xiaohua

Hangzhou Institute of Electronics Engineering

Abstract: Differing from the western languages, the constitution of Chinese phrases is most like that of the sentences. Therefore, the structure of Chinese phrases reflects the structure of the sentences. For this reason, the research of Chinese phrases is more important comparing with other languages. This paper introduces a corpus based approach of parsing Chinese phrase's boundaries. It just considers the tag string of Chinese sentences as the investigating objects.

1　汉语短语研究的意义及难点

由于自然语言的复杂性，要让计算机理解自然语言，就必须用层次结构的观点来分析语言现象。短语作为语言的一个层次，占有十分重要的位置。语言学家朱德熙先生说过：与印欧语言不同，汉语句子的构造原则跟短语的构造原则基本一致，如果我们把各类短语的结构和功能都足够详细地描写清楚了，那么句子的结构实际上也就描写清楚了。可见，短语结构的研究在自然语言理解，尤其是汉语理解中占有重要位置，它对于句子结构的分析、句型的确定和区别以及复杂短语结构的分析都有着十分重要的意义。

汉语相对于印欧语言，有着许多不同于印欧语言的特殊语言现象。对短语分析来说，主要有以下两点：

(1) 汉语缺乏形态标记

在现代汉语中，一个句子可以有多个动词，而英语中一般只有一个定式动词(finite verb)。

例如汉语句子:“我倒了一杯茶给他喝”,该句有三个动词,这就使确定中心谓语动词的工作比较麻烦。而如果把“倒”字理解为中心谓语动词,那么译成英文就只有一个动词,即:

I offered him a cup of tea.

此外,汉语中的许多词类在不同的上下文中可以有多种不同的用法,比如动词既可以作句子的述语,也可以构成主谓、述宾、述补或小句等去修饰其它中心语。而句子当中并没有标记来表明其具体的用法。例如:

她很爱干净

爱干净的她

(2) 汉语短语的构造原则与句子的构造原则基本一致

汉语的这一特点,导致了许多句法上的歧义结构。例如下面的几组短语,仅从其构成形态上说是一样的,而实际上它们的内部结构并不相同。

成为 V ((人类 N 进入 V 太空 N)的前奏 N)　　研究 V (人 N 的心脏 N)

(((爱 V 国 N)(献 V 身 N))的精神 N)　　((研究 V 科学 N)的人 N)

又如,汉语里存在着主谓谓语句这种句型,如:“故宫建筑群规模宏大”。该句的大主语应为“故宫建筑群”,“规模”是小主语,“规模宏大”这个主谓短语作整个句子的谓语。而类似的情况也可能是前面的体词性成分构成一复合名词,其后的形容词作整个句子的谓语,例如“科技情报事业很重要”。

印欧语言中,由于其短语的构造原则有别于句子的构造原则,因而这种歧义结构相对较少。

汉语的这些特点,都无疑给汉语的分析与处理增加了不小的困难。

2　汉语短语划分的策略及实验系统

汉语短语边界划分所需的信息很多,我们尝试了一种只利用句子词性标记串信息的统计信息与规则相结合的分析方法,而未用到其它句法、语义信息。所采用的标记集(Teg Set)共计有 74 个词性标记,是以“八·五”的《关于汉语语料库词性标记的规定》(1990 年 6 月第二稿)和山西大学的《汉语语料库语法标记》(1991 年 6 月第三稿)为基础,并经过部分修改而得到的。

我们选取的语料主要来自清华大学进行汉语词性自动标注所用语料中的一部分(约十万字),包括军事、新闻、科技、科普等方面题材。语料质量高,词性标注一致性好。

2.1　统计信息在汉语短语边界划分中的应用

在汉语短语边界划分的过程中,我们主要使用了两类统计信息:互信息和动词统计信息。

2.1.1 互信息

互信息(Mutual Information)的概念较早见于 Fano 1961 年发表的有关信息论的论文

《Transmission of Information》，其中，互信息被作为一种衡量两个信号关联程度的尺度。这种二元互信息可表示为这两个信号概率的函数，如式(1)所示。在自然语言处理中，则是把句子的词或词性序列作为一串随机事件，然后用互信息来对它们进行分析与研究。

$$MI(x,y) = log \frac{P_{XY}(x,y)}{P_X(x)P_Y(y)} \qquad (1)$$

在过去的十年间，互信息法在语音识别、名词分类及其它一些领域内得到了较为成功的应用，由此可把互信息法进一步推广应用于短语边界的划分。

考察式(1)，如果我们把 x、y 分别看作两个词或词性标记，那么这个二元互信息即为：

$$MI(x,y) \approx log \frac{\dfrac{xy\ 二元对在语料中出现的次数}{语料中二元对的总次数}}{\dfrac{一元项\ x\ 在语料中出现的次数}{语料的总词次} \times \dfrac{一元项\ y\ 在语料中出现的次数}{语料的总词次}} \qquad (2)$$

由式(2)可见，互信息值的大小反映了 xy 之间结合的紧密程度。而自然语言中的短语都是按一定的结构方式构成的，一般说来短语内部的词之间的结合较之于短语与短语之间的词的结合往往更为紧密。因此考察句子的词或词性序列的互信息值可以为确定短语的边界提供一定的帮助。可以证明，互信息值最小的地方很可能就是短语的一个边界。

基于互信息的短语边界的分析过程实际上就是考察由句子的各个可能分界的互信息所构成的一条函数曲线，并找出该曲线的各个极小值点的过程。若相邻若干点值的大小极为接近(以其周围点的值的大小作为参照)，则应认为它们是等同的。换句话说，可能出现几个相邻的点对其周围点来说都是极小值，也就是都为短语边界的情况。这与数学上极小值的概念有所不同。

互信息法是一种建立在语料库基础上的统计方法。国外语料库研究起步较早，语料的自动词性标注已在较大规模语料库(如 Brown 语料库和 LOB 语料库)上实现，从而为进一步的句子结构研究提供了良好条件。由于国内对汉语语料研究尚处初级阶段，无论在语料库的规模还是在语料加工深度方面都还远不能满足要求。因而目前的基于统计的汉语短语的研究所面临的数据稀疏问题要比英语的研究严重得多，这无疑会降低系统的性能，限制研究的深度。

由于所选取的语料约为 6 万词次(Tokens)，是一个小规模语料库，故采用二元互信息模型较为合适。因为高元模型将会面临严重的数据稀疏问题，降低统计信息的可靠性。实践证明，在目前语料规模下，不可能明显提高短语划分的精度。词性标记集的大小对数据稀疏程度也有较大影响，这是由于系统的参数空间与词性标记集的大小呈指数关系，若词性标记集过于庞大，则必然会导致统计信息过于分散，产生大量的低频度信息，影响统计信息的可信度。但若词性标记集过小，又会掩盖语料中一些细微有用的语言现象。我们采用标记集为 74 类标记，较为适中。

实验结果表明：

(1) 纯互信息方法对汉语语料中短句的短语划分效果较佳。

例如：　(目的)(是)(消灭　敌人)。

(学员)(先　在　地面)(接受)(教员　指导)。

(2) 对于结构较为复杂的汉语语料,纯互信息方法划分的结果不能很好体现句子中短语的层次结构。

这主要由于汉语缺乏形态特征及短语的构造原则与句子的构造原则基本一致,因而也就缺乏词在句子中的结构层次信息。

例如: (物理学)(是)(研究　物理　现象)(的　科学)。

句中动词"是"和"研究"分属句子结构的不同层次。

此外,上例中,由于"的"与"科学"之间的互信息值较大,而被归并为一个短语,但实际上"的"归并在前面一个短语中更为合适,即"研究物理现象的"。这就说明单纯用互信息值的大小决定短语分界点还是不够完善的,这涉及对语料中所获取的互信息进一步加工处理的深层次问题。

由上可见,纯互信息的方法对结构较为简单的汉语短句比较适用,为了进一步提高互信息方法对汉语短语边界划分的处理能力,要在下列几方面予以改善:

(1) 建立更大规模的语料库,使互信息的获取具有更高的可信度。

(2) 适应工程应用需要,建立一套更为完善的词性标记集。

(3) 对所获取的互信息进行深层次加工。

(4) 采用其它适当方法作为互信息方法的补充,以弥补互信息方法的不足。

2.1.2 动词统计信息

一般说来,汉语的句子大都主语在前,述语居中,然后是宾语。而用作述语最多的是动词类词语。因此,动词在汉语句子分析中有着十分重要的作用。此外由于汉语型态标记的缺乏,一个汉语句子中常常有多个动词成份,这就对主动词的确认带来了一定的困难。

为此,我们事先对句子中可能出现的动词对及其中主动词的情况进行了统计,以便利用这些信息对句子中动词的层次情况作出推测,并确定句子的主动词。

2.2　规则在汉语短语划分中的应用

2.2.1 短语构成规则

为提高二元互信息的准确性,互信息的统计是在已经部分归并过的汉语句子上进行的,这个归并就是利用短语构成规则实现的。主要是把一些较为简单的修饰成份与它们所修饰的中心词语归并起来,另外对部分连用的同类词也进行了归并。这样,互信息可以更好地反映句子中心词和中心词之间的联系,摈弃某些低频度信息,使统计信息相对集中,改善数据稀疏的状况。

2.2.2 短语分界规则

短语分界规则主要是针对一些较为明显的短语边界而设计的。事实上,基于互信息的短语边界划分方法的理论依据是短语结构稳定,其内部词与词之间的联系较短语与短语之间的联系更为紧密。然而这一依据本身就有例外的情况。例如在汉语中,如果名词后跟副词,则显然这里是短语的一个分界(这是对内层短语结构而言),而通过对语料的统计可以发现名词与副词的共现频度相当高。由此可见,分界语法规则对基于互信息的短语边界分析器来说是必不可少的,它可以避免由于这种例外情况所带来的错误。但分界规则本身有一个适用层次的问题,

它并不对分析的所有层次(指短语的不同层次结构)有效。

2.2.3 介词短语规则

介词短语在汉语句子中出现的频度也相当高,并且介词短语往往跟谓词性词语配合起来使用。因此它跟谓词有一定的语义联系。由于各种介词的意义不同,或者由于介词后边实词(或短语)意义的差异,介词短语跟谓词之间发生关系后所表示的语义也是多种多样的。作状语时,介词短语大都放在主语和谓语之间;作补语时则放在谓语动词之后。可见介词短语与谓词之间有着十分密切的联系,对它的确认也可以为句子结构的分析提供有益的帮助。为此,也专门为介词短语的分析提供了少量规则。

3 综合系统实验结果举例

(文献 中)(论证 了)((利用)(双曲模型)(来 解决)(这 一 问题)(的)(可能性))。

((物理学家)(研究)(物理 问题)(的)(方法))(是)(多种多样 的)。

(这 位 老人)(叫)(平措)。

((用户)(使用)(这 一 系统))(就 可 得到)((这些 命令)(的)(中文 接口))。

(北洋 政府)((在 北京 南苑)(建立 了))((中国)(第 一 所 航空 学校))。

((使)(大规模 地 训练)(飞行员))(成为)(可能)。

(Yu 和 Rotertson)((也 对 参数 估计)(作 了))((深入)(的)(探讨))。

((一些 国家)(的)(军队))(认为)((核 火力)(将 成为)(防御 火力配系)(的)(组成 部分))。

4 对基于语料库的汉语短语研究的思考

目前,我们已完成对上述6万词次语料的分析。从分析的结果来看,这种统计信息加规则的分析方法对不太复杂的句子来说还是令人满意的。当然,我们的分析对较为复杂的句子来说还只是部分分析,而非分析句子的所有层次。另外并不对所分析出的短语进行标注。由于目前我们的语料规模较小,因而不可能用互信息法对汉语短语进行更深层次的研究。但不管怎样,我们选用的是真实语料,这是处理大规模真实文本的基础。自然语言是十分复杂的,汉语尤其如此。尽管目前我们的分析还只是部分的,但只要达到一定的程度,就能够对以后的句法分析提供有益的帮助,使其站在一个比较高的层次上来对汉语的句子进行分析。此时,句法分析所需考察的不再是各个词之间可能有的关系,而只需考察短语与短语之间以及短语内部各词之间的关系,因而可以大大降低句法分析的工作强度。

对使用规则的分析方法来讲,规则的正确与否及它的适用范围决定了分析正确率的高低。然而,由于自然语言并不象形式化语言那样规范,它可以有许多的变化甚至本身就存在语法上的错误,以至于绝大多数的规则都不是说绝对正确适用的。而且,由于设计者对语言千变万化认识的不足,以及语料中语言现象的不完整,也使得规则本身存在一定的偏差和空缺,这必然要影响到分析的正确率。另外,在使用规则的方法中,规则之间的相容性也是必须注意的问题。

当规则增加到一定数量后，人们往往很难保证它们之间完全相容。此外，规则还有一个适用层次范围的问题。

汉语相对于印欧语系而言，有许多自己所独有的特点。对短语研究来讲，其短语的构成原则与句子构成原则基本一致就是十分重要的一点。它导致了许多句法上的歧义结构，如果仅用词性标记的同现信息很难确定其具体结构。另外，汉语的定心短语可以拥有很长的修饰语，且往往放在中心语之前，这就导致了句子的谓语与宾语中心语之间相隔很远，其中又夹杂了其它小句的主、谓、宾语，从而使分析器很难仅凭词性标记来确定一个句子的谓语及其所带宾语。正是由于汉语短语构成的这种独特方式，使单纯依靠词性标记同现信息的互信息法在处理复杂汉语句子时可能就显得力不从心。为此，可在目前的基础上再增加有关语义、语用等方面的知识。

就计算语言学来讲，知识的获取与表示是解决问题的根本。目前为止，人们还是更多地继承了传统语法的一套理论体系，这是前人在语言学研究方面的成果，应该加以借鉴。但这种单纯从语言学角度考虑而得出的一套体系，究竟适不适合工程应用的需要还是值得思索的。目前的汉语词性标记分类主要来自于语言学上汉语词的分类方法，依照这种分类，一类词可以担当太多的语法功能，而且标记本身难以体现句子的层次结构。能否在词性标记的设立上更多地考虑句法功能和层次性，增加词性标记所携带的信息量，是今后研究中可以探讨的。

此外，自然语言处理是一项十分庞大而繁复的工程，人们不可能一步就达到大规模真实文本处理的目标，而必须逐层逐步地加以分析和解决。这就涉及到了各个层次研究之间的衔接问题。总的说来，自然语言学各层次的研究既相对独立，又一环紧扣一环，有着十分密切的联系。对每一层次的研究，都应该充分考虑更高层次研究的需要，并以此为目标来开展研究，紧紧围绕处理非受限大规模真实文本这一最终目标来组织与实施自然语言处理的研究。

参 考 文 献

[1] Magerman D. M. and Marcus M. P., Parsing a Natural Language Using Mutual Information, Statistics Proceedings of AAAI'90, 1990. PP. 984——989.

[2] 范晓，汉语的短语，商务印书馆，1991. 11.

[3] 白栓虎，基于统计的汉语语料库词性自动标注的研究与实现，清华大学硕士论文，1992.

[4] 史有为、黄昌宁、刘开瑛，关于汉语语料库词性标记的规定(第二稿)，1990. 6.

[5] 孟琮、郑怀德、孟庆海、蔡文兰，动词用法词典，上海辞书出版社，1987.

[6] Keh—Jiann Chen, Design Concepts for Chinese Parsers，第三届中文信息处理国际会议，1992. 10.