基于《知网》的中文语块抽取器

董 强　郝长伶 董振东
中国科学院计算机语言信息工程研究中心 北京 100083
E-mail: support@keenage.com

摘　要：我们根据“中文信息结构”的理论，以《知网》和《知网-中文信息结构库》为主要资源，开发了中文语块抽取器。本文简要的介绍了中文信息结构的理论，重点说明了中文语块抽取器的工作原理、过程、实现方法及其独有的特征。重点包括以下几个方面：切分、组词、消歧和中文语块抽取以及本系统重要的组成部分 — 信息结构解析器。中文语块抽取器将可用于中文文本的部分分析，计算机辅助的中文语块库的建设，结构和语义消歧，以及将可成为信息抽取（如实体、事件等）的工具。

关键字：语块；语块库；中文语块抽取；知网；知网-中文信息结构库

HowNet-based Chinese Chunk Extractor

Qiang Dong　Changling Hao　Zhendong Dong
Research Center of Computer &Language Engineering, Chinese Academy of Sciences, Beijing, 100083
E-mail: support@keenage.com

Abstract: On the basis of the theory of Chinese message structure, we have developed HowNet-based Chinese Chunk Extractor, using HowNet and HowNet-Chinese Message Structure Database as the main resources. This paper touches upon the theory on Chinese message structure. It describes the process of HowNet-based Chinese chunk extractor, mainly including: character segmentation, word/expression grouping, message structure Parsing, disambiguation, chunking, and its main component – message structure analyser. Chinese chunk extractor will surely be applied to text partial parsing, computer-aided chunk bank construction, WSD, and information extraction.

Keywords: chunk; chunk bank; Chinese chunk extraction; HowNet; HowNet-message structure database;

1. 引言

语块的辨识、分析、捆绑（chunking）是当前语言技术研究的热点之一。事实证明，不同的语言，有着不同的特点，应该采取不同的技术策略。与印欧语言相比较，中文没有那么丰富的形态变化，中文的词类与句法功能不是一一对应的，中文的词、短语、句子之间的界线是模糊的。鉴于这样三个特点，董振东在 1996 年提出了中文信息结构理论，并于 1999 年开始研究和建设了基于《知网》的《知网-中文信息结构库》。这也被认为是知网向中文研究的延

伸。根据这一理论我们研发了基于知网的中文语块抽取器（Chinese Chunk Extractor）。中文语块抽取器的理论关键是：第一，它对于中文的词、短语进行一体化的处理；第二，它主要是基于语义的。它从文本中抽取的不只是一种句法结构，更主要的是一种语义结构。这将是它与流行的语块分析或树库加工的主要区别。有人因此担心是不是会大大增加难度。根据我们的经验，事实并不如此。例如，不管用什么方法，都会面对诸如“打击力度”和“加大力度”的辨识问题，关键是要有较有力的资源。

中文语块抽取器将有很广阔的应用前景。它可以被用于中文文本的部分分析，计算机辅助的中文语块库的建设工具（因为它可以人机交互方式从大规模语料中抽取中文语块），结构和语义消歧的工具以及成为信息抽取（如实体、事件等）的工具。

2. 中文语块抽取器

2.1 中文信息结构

如前所述，中文语块抽取器是从中文文本中识别我们所认定的语块的一个软件包。我们所认定的语块即是中文信息结构。中文信息结构是中文中句法和语义合理的一个语言片段，它可能是传统被认定的词语，也可能是一个比词语更大的语言片段。它是利用《知网》中文信息描述语言描述中文词语的各个组成部分之间的、由《知网》所规定的动态角色关系和属性的表达式。通过对信息结构的揭示，我们可以认识到中文是如何描述诸如事件、实体、属性、属性值等等概念的，或如何由简及繁的表达意义的。例如，“高跟儿鞋”通常是被作为一个词而收入词典的，但“圆领衫”、“长统袜”一般不作为词收入词典。对于我们的中文信息结构而言，它们有着完全相同的句法和语义结构，它们都是我们所认定的语块，也就是中文语块抽取器要辨识和抽取的目标。在《知网-中文信息结构库》（2000 版）中，我们曾总结了 271 种中文信息结构。

2.2 知网－中文信息结构库

《知网-中文信息结构库》是中文信息结构的集合，它主要有以下几个特点：

1） 反映中文的真实特征。《知网-中文信息结构库》真实体现了中文语块的构词特征，以及这些短语中词语之间的语法语义等的关系。

2） 内容真实性。这是因为他的素材来源于实际语料库，而另一方面又是经过反复的真实语料的验证和人工进行筛选整理。

3） 分类使用。信息结构按照一定的标准进行了分类，这样就可以根据不同的应用调用不同的信息结构模块。

4）它可以被认为是袖珍型的经典语料库，它的覆盖面广但又能避免统计价值不高的重复。

2.3 中文语块抽取器

中文语块抽取器使用了两个知识资源：《知网知识库》和《知网-中文信息结构库》。中文语块抽取器的功能主要包括两个方面：第一个方面是对《知网-中文信息结构库》的数据进行管理和维护；第二个方面是运用这些知识分析抽取中文文本中的中文语块。本文以第二个方面为主要论述对象。在抽取中文语块的过程中，根据《知网-中文信息结构库》中信息结构模块优先级的不同制定相应的抽取策略，这是中文语块抽取器重要的设计思想之一。

2.4 信息结构解析器

信息结构本身是静止的，这就需要信息结构解析器对信息结构进行解释，判断中文语块符合信息结构的条件，然后分析出中文语块的语法语义等信息，通过这样的机制，中文信息结构就能够活跃起来，从而能够达到抽取中文语块、进行文本部分分析的最终目的。

信息结构解析器的基本工作原理：根据要求，调用《知网-中文信息结构库》的信息结构，解析信息结构，逐个分析中文语块中各个词语在《知网知识库》中的概念含义，判断推理中文语块的合法性，最后得到中文语块的各种信息，例如词语语义信息和语块结构信息等。

信息结构解析器通过以往经验的积累（信息结构库），来认知一个新的语块。例如：抽取“红-葡萄”短语特征建立的信息结构，信息结构解析器使用这个信息结构就可以分析“红-苹果”的各种信息。

需要强调的是，信息结构解析器判断推理的过程实际上是基于语义判断推理的过程，这是本系统的最大的特征。通过以上的介绍，我们可以认识到，信息结构解析器是中文语块抽取器的最关键的组成部分，是本系统的灵魂。

信息结构解析器的另一个功能是在维护《知网-中文信息结构库》的时候，按照信息结构的语法规范分析信息结构的合法性，给出分析报告。

2.5 中文语块抽取器的工作流程

图 1 是中文语块抽取器的工作流程图，在这个图中简单展示了一个例子的处理过程。通过这样的处理，句子的各个部分从形式上就得到了简化，但是得到的信息却很丰富。

2.5.1 切分

这是中文语块抽取器工作流程的第一步工作，主要是对输入文本中的一个完整的句子（内容 A）以字为单位进行切分，最后得到内容 B。以字为单位对文本进行切分，是本系统和其他系统的不同之处，这是由中文语块抽取器所使用的资源以及本系统的设计思想所决定的。

2.5.2 组词

利用《知网知识系统》的词表对切分的结果进行任意的组词。图 1 中的内容 C 为组词的结果，其中标有 * 的部分“云南边境”为歧义字段。值得注意的是：在组词的同时，我们会保留切分的结果，而不是将切分的结果丢掉，这样做是为了保证中文语块抽取器能够进行可逆性的处理。

在一般的系统中，歧义字段是在分词的时候被发现，并且进行消歧的。而在语块抽取器中，

歧义是在组词的过程中产生的，这种歧义我们称之为组词歧义。而对于歧义字段的消歧是在后续的功能模块中，通过信息结构库提供的消歧模块或者通过其他的策略逐步消歧的。同时随着对上下文的深入分析，消歧的结果也会由于不合理而重新被打散并且再次消歧。

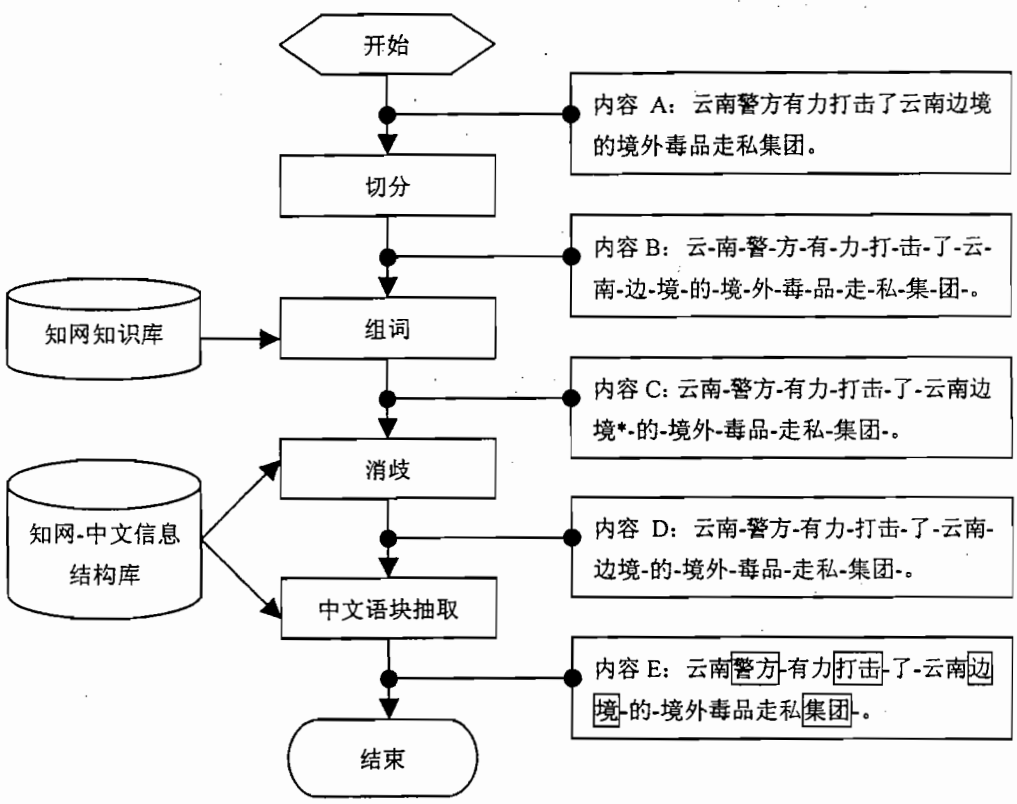

图 1：中文语块抽取器工作流程图

2.5.3 消歧

在消歧之前要做如下的准备工作：

1) 找出歧义字段。在我们的例子中，只有一个歧义字段是“云南边境”，而在真实的文本分析中，一个句子中的歧义字段可能不止一个。
2) 取得歧义字段的组合。所谓歧义字段的组合是这样的：歧义字段中所有可能的词语，通过首尾相连的排列，可以得到的所有可能的排列组合。例如内容 C 中歧义字段的组合有 5 种，分别是：云南-边境、云南-边-境、云-南边-境、云-南-边境、云-南-边-境。我们根据这些组合的部分数对它们进行排序，将部分数少的组合放在前面。这是因为大量的实验表明，歧义切分词语越少，正确性就越高，经过这样的排列，在程序实现上可以提高消歧功能模块的处理效率。

歧义字段的歧义往往有一定的规律可寻，可以根据这些规律建立用于消歧的信息结构，这些信息结构就构成了《知网-中文信息结构库》中的消歧模块，中文语块抽取器的消歧功能主要依据的就是这个消歧模块。

中文语块抽取器的消歧功能模块是在信息结构解析器的驱动下工作的，它的基本实现原理是：依次输入歧义字段的组合，用信息结构解析器分析这些组合，如果某一个组合成功匹配一条信息结构的话，就认为这个组合是正确的。在《知网知识系统》中，词语的意义可能有多个，但是通过中文语块抽取器的推理判断，中文语块中词语意义是唯一的。在图 1 中，内容 D 为消歧的结果，歧义字段“云南边境”通过消歧模块的处理得到它的正确组合，即“云南-边境”。尽管消歧模块中可以处理掉大量的歧义，但是仍然会存在一些歧义字段无法通过消歧模块被处理的情况，其中的原因是多方面的。第一，尽管我们希望尽可能多的抽象歧义字段的组合特征，但是这样的特征不可能穷举；第二，还有一些歧义不是观察歧义字段的特征就能够进行消歧的，需要通过考察它的上下文的情况才能够进行消歧。对于这样的歧义字段，我们将在中文语块抽取的过程中对它们进一步消歧。

2.5.4 中文语块抽取

中文语块抽取指的是：通过信息结构解析器的分析判断，将一段文字中依次的一组词语与信息结构进行匹配，匹配上的一组词语将被看作是一个完整意义的中文语块，这个过程被称作中文语块抽取。例如“境外-毒品-走私-集团”，它是一个包含了一定的语法和语义的中文语块，在信息结构库中能够找到它对应的信息结构。

和前面说的消歧功能一样，中文语块抽取也是在信息结构解析器的驱动下工作的。这个模块是中文语块抽取器中最重要的功能模块。通过中文语块抽取，达到以下的几个目的：

1) 提取中文语块的中心词语。
2) 得到中文语块中每一个词语含义，实现中文语义分析和语义消歧的功能。
3) 计算出中文语块的意义，并且这个意义是唯一的。
4) 得到中文语块中各个词语之间的关系，例如“红-苹果”中“红”修饰“苹果”，这个中文语块是定中结构。
5) 简化句子。我们可以使用中文语块的中心词语替代中文语块，并且这个短语有自己的含义。也就是说抽取出来的中文语块在这里被看作是一个整体，以中心词语来表现它自己。

在图 1 的内容 E 中，我们使用符号“–”来分割不同的语块，在不同的语块中方框中的词语为语块的中心词语。可以用中心词语代替中文语块，但是除中心词语外的其它信息仍然需要保留。例如“云南警方”的中心词语是“警方”，可以用它来代替中文语块“云南警方”，而其中“云南”这个信息不能删除，它是对“警方”的一种限定。这样在需要这个信息时，可以把它提供给用户。

中文语块抽取功能的实现设计思想是：反复扫描要处理的内容 D，每次扫描抽取内容 D 的一个能够匹配上一条信息结构的中文语块，抽取出来的中文语块在系统内部被看作一个完整的概念，在下次扫描抽取的时候可以利用上次处理的结果，直到没有可以抽取的中文语块为止。例如对图 1 中的例句进行语块抽取的过程如下：

内容 D：云南-警方-有力-打击-了-云南-边境-的-境外-毒品-走私-集团-。
==> 云南-警方-有力-打击-了-云南-边境-的-境外毒品走私[集团]-。
==> 云南-警方-有力-打击-了-云南[边境]-的-境外毒品走私[集团]-。
……

==> 云南[警方]-有力[打击]-了-云南[边境]-的-境外毒品走私[集团]-。

为了更加具体的体现中文语块作为一个整体进行处理的特征，以“一个红苹果”这个语块的抽取过程为例：

一-个-红-苹果 ==> 一-个-红[苹果] ==> 一个红[苹果]

由此可见，一个语块被看作一个整体以后，句子就会在反复的扫描过程中逐步得到简化。由于在中文中存在这样的句子，如：“很红的苹果”，其中“很”修饰“红”，“红”修饰“苹果”。如果把“红”和“苹果”作为一个语块的被抽取后，“很”和“红” 就不能作为一个语块被抽取出来。因此，中文语块抽取器将根据《知网-中文信息结构库》中信息结构模块优先级进行抽取，例如我们会优先抽取类似“副词-形容词”的语块，然后抽取“形容词-名词”的语块，这样不仅提高了处理的效率，而且降低了不合理的语块被抽取的几率。

很-红-的-苹果。 ==> 很[红]-的-苹果。 ==> 很红的[苹果]。

通过中文语块抽取，我们不仅要找出中文文本中的语块，同时还要最大限度的获取这些语块中的各种信息。以“香港特区政务司司长陈方安生”这一语块为例，它的中心词语为“陈方安生”，从这个语块中，我们可以获得如下的信息：“陈方安生”是女性，姓“方”，名“安生”，已婚，她的丈夫姓“陈”，“陈方安生”有职位，她是“香港特区政务司司长”等。这些信息都是隐含在这个语块当中，而没有明确的表示的，但是这些信息对于中文信息处理系统而言又是很重要的，而利用《知网－中文信息结构库》就可以获得这些信息。

3. 结束语

中文语块抽取器从功能实现方面充分体现了《知网》和《知网-中文信息结构》所阐述的理论，通过信息结构解析器的驱动，实现了中文文本自动切分、组词、消歧、中文语块抽取等这几方面的工作的有机结合，最终达到了抽取中文语块、获得语块的内在信息的目的。

参 考 文 献

[1] Christer Johansson. 2000. A context sensitive maximum likelihood approach to chunking. In Proceedings of CoNLL-2000 and LLL-2000. Lisbon, Portugal

[2] 董振东，董强（1999），“知网”，http://www.keenage.com

[3] 董振东，董强《关于知网-中文信息结构库》

[4] 刘开瑛《歧义切分与专有名词识别软件》，山西大学计算机科学系，《语言文字应用》2001.3

[5] 国家技术监督局，中华人民共和国国家标准 GB/T13715-92《信息处理用现代分词规范》，中国标准出版社，1993

[6] 梁南元，书面汉语自动分词系统-CDWS，《中文信息学报》，第 1 卷，第 2 期，1987

[7] 周强，张伟，俞士汶，汉语树库的构建，中文信息学报，1997，11