第 33 卷 第 4 期 计 算 机 工 程 2007 年 2 月
Vol.33 No.4 Computer Engineering February 2007
·人工智能及识别技术· 文章编号：1000—3428(2007)04—0157—03 文献标识码：A 中图分类号：TP39

自然语言处理中句群划分及其判定规则研究

吴 晨[1,2]，张 全[2]
(1. 中国科学院研究生院，北京 100039；2. 中国科学院声学研究所，北京 100080)

摘 要：在自然语言处理，尤其是在基于语法和语义规则的信息检索、机器翻译系统中，对于句群的处理显得格外重要。它是计算机从理解孤立的词义和句义上升到理解篇章整体中心内容的一个重要的跃变步骤。作为句群理解的关键一步，句群的识别显得尤为重要。该文从句群本身的构成特点出发，对句群进行了内部语义组合方式的划分，这一划分适宜计算机进行处理。根据已经取得的“HNC 语言概念空间表示”的研究成果，制定了识别具有以上构成特点句群的相关规则。实验表明，划分方法具有很高的句群覆盖率，同时切分规则具有很高的准确度。

关键词：句群；切分策略；计算语言学

Research on Rules for Detecting Chinese Sentence Groups in Nature Language Processing

WU Chen[1,2], ZHANG Quan[2]
(1. Graduate School, Chinese Academy of Sciences, Beijing 100039; 2. Institute of Acoustics, Chinese Academy of Sciences, Beijing 100080)

【Abstract】 In nature language processing, the capacity of processing the sentence group becomes more and more important. It is the key to obtain the meaning of a paragraph, a chapter, and even the full text. This paper proposes a method for marking off the Chinese sentence groups based on the semantic relationship between the sentences within the sentence group. The paper also presents some formalized rules for detecting the Chinese sentence group. They are based on the division method and take advantage of the symbolic system of language concept space which is used to express the meaning of a word or a sentence aforetime. The experiment indicates that the method and the rules are good at detecting the Chinese sentence group.

【Key words】 Chinese sentences group; Strategies for detecting; Computational linguistics

自然语言处理技术经过了几十年的发展，涌现了一大批行之有效的方法。为行之有效地解决自然语言处理问题，越来越多的科学工作者开始把研究转移到数学模型与语法学相结合的道路上来，从根本上解决语义的模糊问题，提升模型的性能。为了能够使计算机从词语的语义入手来解决实际问题，语义的形式化表示显得颇为重要。

WordNet[1]，HNC[2,3]，HowNet[4]都提出并实现了各自的语义表示方法。HNC利用语言的概念表示式给出了一套表示词语、句子语义及结构的形式化表示规则，并将其定义为语言概念空间的主要内容。为形成篇章中心语义的形式化表示，从而实现篇章级的自然语言处理奠定了坚实的基础。

然而，句子本身所能承载的上下文信息太小，从单个句子到篇章的语义过渡太大，为了能够更好地做到对篇章的处理，顺应传统语言学的思想，句群这一概念被引入计算语言学中。

传统语言学中对于句群的一些抽象定义无法满足计算机处理的需要。本文的研究就是在这一背景下提出来的。制定能够满足计算机处理需要的句群划分及判定规则，准确切分句群，为后续的句群处理奠定基础是本文研究的出发点。本文所描述的规则都是基于 HNC 自然语言处理框架之下的。本文将通过句子间的语义联系(内容上)和句子间的符号特征(形式上)两个部分来讨论切分问题。

1 相关工作

HNC制定了一系列的符号体系来形式化地表示词语的语义及句子的语义结构。同时提出并实现了一套计算机确定词语语义、句子语义结构的软件系统，即句类分析系统[2]。通过系统处理，可以得到表示句子及词语语义的概念符号，从而消解句子及词语的语义模糊。

HNC用句类表达式[5]来描述句子的语义结构，共定义有 57 组基本句类表达式和 57*56 组混合句类表达式。构成句类表达式的是语义块和它们之间的连接符号。语义块是处于词语和句子之间的语言单位，类似于词组，描述句子中相对独立的一个基本单位。

语义块又被分为主语义块和辅语义块，其中主语义块构成句子的主要要素，如句子表述的对象、内容等，主语义块可以分成特征语义块和广义对象语义块两类。谓语所在的语义块往往构成特征语义块，主语和宾语所在的语义块往往构成广义对象语义块。辅语义块是句子的可选要素，描述句子中事件的背景信息，如条件、实施手段等。本文讨论的句群划分规则都将基于这些概念之上。

下面通过一个例子加以说明。本文所举的例子将遵循这样的描述规则：

基金项目：国家“973”计划基金资助项目“自然语言理解的交互引擎研究”(2004CB318104)；中科院声学所知识创新工程基金资助项目“HNC 语言知识处理理论及技术”
作者简介：吴 晨(1979－)，男，博士生，主研方向：自然语言处理；张 全，研究员、博导
收稿日期：2006-03-24 **E-mail：**wuchen@mail.ioa.ac.cn

文字部分为句群(句子)本身，符号串部分为对应句群(句子)在概念空间中的语义表达式，句子的语义表达式用 HNC 句类表达式来表示。对于句群而言，语义表达式为构成这个句群的各个句子的语义表达式的集合。句子的语义表达式以“ SC = ”打头；句群的语义表达式以“ SGC = ”打头。

例 1：一天的下半天 ~|| 没有 || 一个顾客。
SC = Cn&!31jD1J

例子中给出了一个例句以及它的语义表示式。对例句进行了语义块标注，句子共有 3 个语义块组成。~|| 之前为辅语义块；|| 划分了两个主语义块。第 1 个主语义块为特征语义块，第 2 个为主广义对象语义块。语义表示式中 Cn 表示句子中有一个条件辅语义块，它对应 ~|| 之前的部分；!31是句子的格式代码，表示句子中语义块的构成以及位置情况，!31表示句子省略了第一个广义对象语义块，正好对应主语；jD1J 是这个句子的句类表示式，表示这是一个存在判断句。详细的格式代码定义、句类表示式定义、辅语义块表示式以及句子标注规范可以参看“ http://www.hncnlp.com/hncnlp.org/020301.htm ”。这样就把句子语义以及结构以概念符号的形式表示出来了。

2 句群的切分依据

句群是在语义上有逻辑关系、在语法上有密切联系、在结构上有衔接连贯的一群句子的组合[6]，本节将以HNC句子语义形式化表示方法为基础，从构成句群的句子的内在联系出发来讨论这个问题。

2.1 从语义的角度

语义是划分句群的根本依据，语义的转移标志着旧句群的结束和新句群的开始。诚然，在语言空间的音和形上缺乏明确的标记，而人可以了解句子的意义，通过人类思维中的概念联想脉络抓住这个隐现之“ 义 ”[2]，将表述中心一致的语句归集在一起，自觉形成句群。HNC的语言概念空间形式化表示方法则为语言空间隐现语义转变成语言概念空间符号串，使其语义内在的关联显现出来。

本小节将从组成句群的句子之间的语义关系入手来讨论划分句群的依据。当下一个句子超出这些关系，就可以认为是一个新句群的开始。

2.1.1 语义的并列

并列关系即几个句子分别说出几种事物或同一事物的几个方面，彼此之间呈横向的平行关系。这种关系在概念空间中判断规则为：

规则描述 1 句子句类表达式，语句格式一致，并列的句子之间特征语义块概念相同，共享句子的第 1 个广义对象语义块。语句格式即句类表达式中各主语义块排列的位置，如在“ 被 ”字句和“ 把 ”字句中，主语义块所处的位置是不同的，那么它们的语句格式是不同的。

例 2：春天||像||刚落地的娃娃，+~从头到脚都新的，++它||生长着！++春天||像||小姑娘，+~花枝招展的，+笑着，+走着！++春天||像||健壮的青年，+~有铁一般的胳膊和腰脚，+领着|||[#我们||上前去！#]

SGC = jD00J + ~ S04J + + SP *11J
+ + jD00J + ~ S04J + SP *11J + SP *11J
+ + jD00J + ~ jD1J + R41104J[# RC#] = PJ

从构成这个句群各个句子的句类来看，句类结构都相同，每个句子都由一个相互比较判断句(jD00J)领头，并且每个句子的第 1 个广义对象语义块都是“ 春天 ”((wj11c41,fb4))，所有句子的语义核心都是在描述春天，这就构成了并列关系的基本条件。

2.1.2 语义的承接

语义的承接关系即句子按时间先后、方位变化或事情发生发展顺序排列，彼此之间呈纵向的连接关系。这种关系在概念空间中表现为：

规则描述 2 句子句类表达式一致；并列的主句之间特征语义块的特征词语具有相同的概念类别符号[2]，这些概念之间具有同行关联性[2]；并列主句的第 1 个广义对象语义块概念相同。

2.1.3 语义的总分

总分关系即句子之间呈总说和分说的关系。可分为：总说－分说；分说－总说；总说－分说－总说 3 种情况。这种关系在概念空间中表现为：

规则描述 3 接着第 1 个句子的下几个句子的第 1 个广义对象语义块是第 1 个句子第一个广义对象语义块或者最后一个广义对象语义块的一部分，这些部分同属于第一个句子相关广义对象语义块所属的概念的延伸概念，如“ 立法 ”(a51)是“ 法律 ”(a5)的延伸概念。这些组成部分有很高的概念相关度，它们在概念符号体系中处于同一个层次，如“ 立法 ”(a51)与“ 违法 ”(a59)概念相关度很高，并处以同一个层次。或者后续句子的第 1 个广义对象语义块与第 1 个句子相关广义对象语义块概念相同，但是后续句子广义对象语义块中包含修饰类概念对该概念进行限定。

2.1.4 语义的解说

解说关系即句子之间呈解释、说明、补充或举例等关系。解说关系和总分关系非常相似，最大的区别是总分关系不能从分句中推断出总句是什么，但是解说关系可以，所以将其单列为一种关系处理。这一特点会给我们萃取句群的语义概念表达式带来很大的帮助。解说关系在概念空间中表现为：

规则描述 4 构成句群的句子都有相同的作用效应类别，比如同属于广义作用句或者同属于广义效应句，并且这些句子的广义对象语义块概念相同，后面的句子是对前面句子的展开描述。

后面句子第 1 个广义对象语义块描述对象构成的集合小于首句第 1 个广义对象语义块描述对象的范围。即首句第 1 个广义对象语义块所指的概念扩展后包含后面句子第 1 个广义对象语义块所指的概念，如“ 法律 ”(a5)扩展后的概念真包含“ 立法 ”(a51)和“ 违法 ”(a59)。这一点也是解说句与总分句的不同之处。

2.1.5 语义的递进

递进关系即后面的句子与前面的句子之间呈一层进一层、逐层深入的关系。在这里，把句子间的假设、条件、因果关系也归入递进关系，因为这些句子有一个共同的特点，那就是首先给出一个论述或者假设，然后在这个论述的基础上加以深入。这种关系在概念空间中体现为：

规则描述 5 后面的句子是前面句子的效应，从作用效应链[2]的角度来考虑，先是有了前面的“ 作用 ”才产生了后面的“ 效应 ”。这类关系从形式上看，会存在语言逻辑类(lb)的关联词语进行关联，如只有、才能；因为、所以。

例 3：[无论准确也好]&[，鲜明、生动也好]~||，就语言方面讲~||，字眼||总要用得||恰如其分。++这样，表现的概念||才会||准确，+也才能使||[#人||感到||鲜明#]。

ReC & Re & Y0J + + Y0J + X03J[# X03BC#] = X20J

这个句群的前一个句子为效应句，后两个句子是效应句和作用效应句，后两个句子是对前一个效应句所产生的效应的进一步的效应说明。表示语义上的深入递进关系。

2.1.6 语义的选择

选择关系即句子分别说出一件事情，或提出一种情况，句子与句子之间构成选择关系。这种关系下，各个句子所表达的意思是二选一或者多选一的。

一个句子描述的内涵为真，那么其它句子的就为假。在概念空间中这种句群表现为：

规则描述 6 所有句子都会产生一种效应，这种效应的主体是同一对象，或者同一个对象的不同部分。也就是说，这些句子的广义对象语义块的概念主体具有概念的一致性或者概念关联性。

在这种关系下，句子中一定会出现句间逻辑符号(lb)，如：或者，还是等，否则很难准确表示语义的选择。

2.1.7 语义的转折

语义的转折即句子之间在语意上呈相转逆接关系，这就是说，后面的句子不是顺着前面的意思往下说，而是转到了跟顺接意思相反或相对的方面。这种关系在概念空间中的表现为：

规则描述 7 可把这一句群前后划分为两组句子，这两组句子的核心语句描述的是相同的对象，即具有相同的概念表达，但是前后两组句子对于对象的描述是相对的，或者是褒贬不一致的。转折关系的句群中，都会有句间逻辑符号(lb)出现。

转折关系是比较典型的语义关系，但是据从语料中的统计结果来看，转折关系多发生在复句的分句间，发生在句群的句子间的可能性非常小。

2.2 从形式上看

句子和句子组合成句群，最大的组合提示信息就是关联词语，关联词语在概念空间中以 lb(句见逻辑，包括句见逻辑语气说明符)节点统摄。lb 节点及其子节点设置可参看文献[5]。

根据句间逻辑符号可以很好地判断构成句群各个句子的关系，从而划分出句群。在规则 6、规则 7 中都把句间逻辑符号作为判断的依据。

识别句间逻辑符号是计算机划分句群要做的第 1 步，在有句间逻辑符号的前提下：

(1)根据句间逻辑符号来假设句群的构成方式；

(2)通过句群内句子间的语义关系加以验证。

采用这种方式将会大幅度提高句群判断的准确率和处理效率。

3 实验及分析

为了验证句群划分的有效性以及切分规则的有效性，做一个实验，对组成句群的各种语义关系进行了统计，同时采用本文所描述规则对各类句群的识别准确率和召回率进行了统计。

测试数据来源于人民网和新华网 2003 年到 2005 年中随机抽取的 50 篇新闻语料，这些语料总共由 1 200 多个句群组成。

实验结果如表 1 所示，表 1 中列举了测试集中各种语义关系构成的句群的数量、所占百分比以及规则识别正确率和召回率。数量根据句子关系的定义统计得来，识别正确率为根据规则识别正确的句群数量占识别出的句群数量的百分比。

识别召回率为根据规则识别正确的句群数量占总的正确的句群数量的百分比。识别率的总和则考虑测试集中全部句群的识别情况。

表 1 实验结果

句群的构成方式	数量	百分比	识别正确率	识别召回率
并列关系	50%	4.20%	81%	65%
承接关系	411	34.20%	79%	62%
总分关系	165	13.70%	76%	69%
解说关系	280	23.30%	82%	71%
递进关系	104	8.60%	81%	81%
选择关系	30	2.50%	97%	89%
转折关系	2	0%	100%	100%
其它关系	197	16.40%	77%	87%
全部	1 203	100%	82.90%	71.00%

从表 1 中可见，构成句群的句子语义关系中，承接关系是最多的，转折关系最少。其它关系占了 16.4%，其它关系主要为一个句子单独构成一个句群的情况，占了其它关系中的 90%，其它关系中还包括句子间的问答组合关系，如例 4 所示句子。

例 4：在台海形势依然微妙的今天，连战为何选择此时参访大陆？契机在哪里？台湾岛内政局的变化。

由于制定的规则着重考虑了句群成立的充分性，因此识别的正确率较召回率要高。

选择关系、转折关系句群被识别的准确率和召回率都较别的句群高，这主要是由句间逻辑符号给识别带来的帮助引起的。

4 小结

本文对句群进行了一个适宜计算机处理的划分，同时从语义的形式化表示方法入手，制定了一些切分句群的标准。当然，句群的划分并不是绝对的，它受到许多因素的影响，如句群切分的颗粒度的大小，因为句子与句子之间在小范围内存在一定组合关系，所以在大范围内同样会存在这样的关系。本文统计所用的句群都是小颗粒度的句群。

同时，本文所提出的切分规则还存在需要改进的地方，下一步工作重点放在制定句群识别的假设检验策略上，进一步提高句群识别的召回率。

参考文献

1 Fellbaum C. WordNet: an Electronic Lexical Database[M]. The MIT Press, 1998.

2 黄曾阳. 语言概念空间的基本定理和数学物理表示式[M]. 北京: 海洋出版社, 2004.

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

4 董振东, 董　强. 知网 [Z]. 1999. http://www.keenage.com.

5 苗传江. HNC 理论导论[M]. 北京: 清华大学出版社, 2005: 300-315.

6 吴为章, 田小琳. 汉语句群[M]. 北京: 商务印书馆, 2002.