基于句间关系的汉语语义块省略恢复*

贾宁[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 100190

Email: gnin_aij@163.com

摘要：语义块是句子的语义构成单位，句子内发生的省略现象可以归结为语义块的省略。本文在句类分析的基础上，从小句间语义块共享关系的角度分析语义块的省略。将语义块的省略分为语义块整块共享形成的省略和语义块部分共享形成的省略，分析了两种情况的特点，并给出了相应的处理算法。测试表明，该算法对于两种省略均有很好的处理效果。

关键词：省略，语义块共享，句间关系

Chinese Ellipsis Recovering Based on Relationship between Sentences

Jia Ning[1,2] Zhang Quan[2]

1. Graduate University of Chinese Academy of Sciences Beijing 100039

2. Institute of Acoustics, Chinese Academy of Sciences Beijing 100190

Email: gnin_aij@163.com

Abstract: A sentence is composed by semantics chunks. So, ellipsis in sentence means ellipsis of semantics chunks. This dissertation tries to resolve ellipsis based on sentence category analysis and relationship of share between sentences. Ellipsis should be divided two categories. The first is ellipsis formed by full semantics chunks share. The second is ellipsis formed by integrant semantics chunks share. An algorithm is represented for ellipsis. Result show that the algorithm is efficient for both two kind of ellipsis.

Keywords: Ellipsis, Relationship between Sentences, Semantics Chunks Share

1. 引言

省略是汉语中的常见现象，适当地使用省略可以使语言的表述更简练，还可以起到连接上下文的作用，使相邻的句子更连贯。省略的恢复对于自然语言处理的许多领域都有重要的作用，如机器翻译、信息抽取、自动摘要等。省略一直是语言学界研究的重要课题，业已取得一定的成果。但是目前的研究多限于理论层面，分析省略的构成、形式等内容，对于如何对省略内容进行恢复的研究并不多见。殷鸿等提出了基于概念模型的省略恢复方法，借助现有的语义模型，从句式意义出发，在语义层面建立了一个省略恢复模型。厦门大学的李旺等引入 Kamp 的语篇表述理

* 本文承国家 973 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项目“句群理解处理理论及其应用”（0654091431）、中国科学院声学研究所“所长择优基金”（GS13SJJ04）。中国科学院青年人才领域前沿项目（O754021432）的资助

论（Discourse Representation Theory DRT），进行了基于 DRT 理论的汉语省略恢复研究。在语篇表述框架 DRS 的基础上，根据上下文的相关信息，对可能的语法空位和语义空位进行填充，以实现对省略的恢复。王厚峰提出了基于句类的真假省略的思想，并通过句类知识和语义块构成知识给出了恢复真假省略的规则。

本章重点分析由于小句间语义块共享造成的省略现象，在句群基础之上对省略进行分析，将省略现象分为语义块整块共享形成的省略和语义块部分共享形成的省略。前一种省略可以通过句间关系信息和句类信息进行判定和恢复，后一种省略的判定和恢复是难点，对这种情况的判定必须深入到语义块内部构成分析，将语义块的核心部分和说明部分分离开，再通过句间关系信息，才能确定发生部分省略的语义块，并实现省略恢复。本章给出了省略恢复的算法，并进行了测试。测试结果表明，对语义块整块共享形成的省略恢复效果很好，对语义块部分共享形成的省略也取得了较好的准确率，但仍需做进一步的研究。

2. 省略与句间关系

省略发生在句子一级的层面，发生省略的通常是一个词，但是本章不从词语的角度分析省略问题，而从语义块的角度来分析。

HNC 理论定义的语义块是一个完整的语义成分，是句子的下一级单位，是句子的直接构成成分。因此从 HNC 的角度研究省略问题，应从语义块着手。先研究语义块的省略问题，再研究语义块内部构成的问题。

HNC 通过句类分析获得了句子的句类代码、句类格式和句间关系等信息。句类代码对应的句类表示式指出了句子由哪些语义块构成，以及语义块间的概念关联关系。句类格式指出了语义块的出现与否及排列顺序。句间关系指出了相邻两个小句间语义块共享的情况。如果发生了语义块省略的情况，句类格式的代码为!3m，“!3”表示有语义块被省略，“m”表示被省略的是哪一个语义块。m=0 表示被省略的是特征语义块，m=1 表示被省略的是 GBK1，m=2 表示被省略的是 GBK2，其它 GBK 依此类推。若省略的语义块不止一个，则被省略语义块的编号都写在“!3”后面，即“!3mn”。本章只讨论广义对象语义块发生省略的情况，特征语义块的省略会导致句类转换，且 pp 类（广义人）概念一般不会出现在特征语义块中，因此不讨论特征语义块被省略的情况。

句间关系信息描述了以逗号为分隔的小句间的关联，小句间的关联根据共享语义块的情况进行分类。前后两个小句间没有共享语义块的，称为列句。后面小句的第一个 GBK 和前面小句的第一个 GBK 共享的，称为迭句。后面小句的第一个 GBK 和前面小句的最后一个 GBK 共享的，称为链句。后面小句的第二个 GBK 和前面小句的第二个 GBK 共享的，称为塔句。后面小句的第二个 GBK 和前面小句的第一个 GBK 共享的，称为环句。没有共享整个语义块而只共享语义块的一部分的，称为半共享句。

例 1：　同日 ~‖，中国驻日本大使王毅 ‖~ 也在东京 ~‖ 紧急约见 ‖ 日本外相町村信孝，+ 向日本 ‖ 提出 ‖ 强烈抗议。

Cn-1Cn-2!0R011J+!311111T3Y30*32J

例子有两个小句，第二个小句是信息转移句 T3J 和一般效应句 Y30 的混合句类，是一个四主块句。它的第一个主语义块即 T3A 被省略了，因此它的句类格式是!31，后面的“1111”表示

该小句四个语义块的排列顺序应为 GBK1+^GBK2+EK+GBK3，“^”是主语义块标志符“向”。第二个小句的 GBK1 之所以省略，是因为第二个小句和第一个小句共享了 GBK1，两句的 GBK1 都是“中国驻日本大使王毅”。因此两个小句是迭句的关系，句间关系符为“+”。

例 2：　上世纪 80 年代 ~‖ ，启功先生 ‖ 先后创立了 ‖ 北京师范大学古典文献学专业硕士点和博士点， +% 2000 年 ~‖ 又与著名学者钟敬文先生等 ‖ 创办了 ‖ 教育部人文社会科学“民俗·典籍·文字”研究基地。

Cn-1!0XY0*21J+%Cn-1!0XY0*21J

例 5-2 中有两个小句，两个小句都是基本作用句 X0J 和一般效应句 Y0J 混合形成的三主块句。第二个小句的两个 GBK 都存在，没有发生整个 GBK 被省略的现象。但是第二个小句 GBK1 的一部分和第一个小句的 GBK1 发生了共享，第二个小句的完整的 GBK1 应为“启功先生又与著名学者钟敬文先生等”。因此两个句子的关系是半共享迭句，句间关系符为“+%”。

句子中省略的内容，一定在前面的内容或者本句的其它部分中出现。口语中有时会出现其它的省略现象，需要结合语境分析，本章不讨论这种情况。例子 5-1 和 5-2 都是省略内容在前面出现过的情况，例子 5-3 是省略内容在本句中出现的情况。

例 3：　{ 今天早上 ~| 出发 }‖ 是 ‖< 老王 | 亲口说的 > 。

!0jDJ#DB={!31T2b1J}#DC=<!32T31J>

例句是一个 DB 和 DC 都是句蜕的是否判断句，DB 是一个原型句蜕，句类代码为 T2b1J，是自身转移句。DC 是一个要素句蜕，句类代码为 T31J，是一个信息转移句。DB 的转移者 T2B 被省略，省略的内容是 DC 的转移者 TA 即“老王”。DC 的转移内容 T3C 被省略，省略的内容就是 DB 这个原型句蜕。

从前面三个例子中可以看出，省略内容在本句中出现的情况比省略内容在前面小句中出现的情况复杂，涉及到语义块之间的概念和逻辑关联。而后面这种情况的信息比较明确，通过句间信息将前后小句串联起来，如果发生了共享的情况，从前面的小句中恢复出后面小句省略的语义块。本文只研究由于小句间共享语义块造成的省略现象，其它的情况在今后的工作中研究。

3. 省略的判定及恢复

处理省略情况的第一步，是判断句子中是否发生了省略现象，哪些语义块被省略了。分别从被省略的对象和发生省略的位置两个角度来分析省略的情况。

从被省略的对象来看，语义块是句子的构成单位，被省略的内容必然属于某一个语义块，可能是整个语义块，也可能是语义块的一部分。为了说明方便，可以把构成句子的语义块分为三种：广义对象语义块、特征语义块和辅块，发生省略的语义块一定是广义对象语义块。如果特征语义块被省略，相当于发生了句类转换，变成了另一种句类，这不在省略现象的研究范围之内。辅语义块不是构成句子的必要成分，无法预期它是否存在，也不存在恢复辅块的问题，因此辅语义块没有省略现象。如果是部分省略，则必须分析出语义块的哪个部分被省略掉了。语义块分为核心部分和说明部分。说明部分不是表达语义必需的成分，可有可无。因此说明部分不会发生省略，即使没有说明部分，也不需要补上。省略情况一定发生在语义块的核心部分。核心部分为并列结构时，被省略的可能是核心的一部分，如例 1 中的例句。

从发生省略的位置来看，省略的语义块可以是句子中的任何一个广义对象语义块。如果是整个语义块被省略，句类格式!3m（m≠0）明确地标出了被省略的 GBK。如果是部分省略，则必须依靠句间关系信息进行判定。如果出现了半共享句符号“%”，说明前后的小句中出现了部分省略的语义块。可能是前面小句中的语义块 GBKm 部分省略，也可能是后面小句中的语义块 GBKn 部分省略，m 和 n 表示 GBK 的序号。通过句间关系符可以确定序号：

表 1 句间关系符与省略语义块的对应关系

句间关系	m	n
迭句	1	1
链句	GBK 的最大编号	1
塔句	GBK 的最大编号	GBK 的最大编号
环句	1	GBK 的最大编号

由于是部分省略，因此前面小句中的语义块 GBKm 和后面小句中的语义块 GBKn 都存在。要确定是哪一个语义块发生了省略，必须分析两个语义块的内部结构。发生省略的部分必定是语义块的核心，因此对比 GBKm 和 GBKn 的核心部分可以确定哪个 GBK 的核心被省略。可能发生的情况包括：

a. GBKm 的核心被完全省略，GBKn 的核心完整

b. GBKm 的核心完整，GBKn 的核心被完全省略

c. GBKm 的核心被部分省略，GBKn 的核心完整

d. GBKm 的核心完整，GBKn 的核心被部分省略

c 和 d 的情况是因为发生省略的语义块的核心是并列结构，有两个或多个并列的核心，其中一个核心被省略。

小句间的语义块共享可能会出现连锁，即多个小句共享一个语义块。比较典型的是迭句的连续共享，所有小句的 GBK1 都发生共享，通常第一个小句的 GBK1 不省略，后面小句的 GBK1 省略。如例 4

例 4： 布什 ‖ 又致电 ‖ “美在台协会办事处处长”包道格，+ 要 [# 包道格 | 亲自转达 | 他对台湾当局的不满 #] ，+ 并用强烈的语调 ~‖ 批评 ‖ 陈水扁。

!0T32J+!0T3XY*31J#B+YC=[# !0T31Y30*21 J#]+Ms!0 X21T32*^21J

例句有三个小句，小句间都是迭句关系。第三个小句和第二个小句共享句 2 的 GBK1，第二个小句和第一个小句共享句 1 的 GBK1，三个小句形成一个共享链，共享的内容都是句 1 的 GBK1“布什”。

判定发生的是哪一种省略及发生省略的是哪些语义块后，即可进入省略恢复的步骤，基本的恢复算法是通过和被省略语义块发生共享的语义块来进行恢复。整块共享的情况可以直接将被省略的语义块恢复为它的共享语义块。部分共享的情况需要对共享语义块进行分解，得到其核心部分后，再使用其核心部分的内容对被省略语义块进行恢复。

恢复语义块省略的详细算法如下：

1. 判断句子中被省略的是完整的语义块，还是语义块的部分。完整语义块转到 2，部分语义块转到 5。

2. 判断被省略的是哪个语义块。

3．向前搜索，在前面的小句中找出与被省略块发生共享的且未被省略的块。成功转到 7，失败转到 4。
4．向后搜索，从后面的小句中找出与被省略块发生共享的且未被省略的块。成功转到 7，失败转到 8。
5．调用前向搜索，成功返回语义块转到 7，失败转到 6。
6．调用后向搜索，成功返回语义块转到 7，失败转到 8。
7．成功结束。
8．失败。

前向搜索
1．省略情况为整块发生省略转到 2，部分语义块发生省略转到 7。
2．找到前面小句中的共享语义块转到 3，没找到转到 14。
3．该共享语义块块是否是被省略的，是转到 4，否转到 5。
4．调用前向搜索，成功返回结果转到 5，失败转到 14。
5．恢复 1 中的被省略语义块。
6．返回恢复的结果。
7．部分省略，找出共享的两个块。
8．获取两个块的核心。
9．判断哪个块发生省略，前面句子的块发生省略转到 10，没有发生省略转到 11。
10．调用前向搜索，成功返回共享块转到 11，失败转到 14
11．分解前面句子的块或前向搜索的结果。
12．判断发生共享的部分，恢复被省略的部分。
13．返回恢复的结果。
14．搜索失败，返回。

后向搜索
1．省略情况为整块发生省略转到 2，部分语义块发生省略转到 7。
2．找到后面小句中的共享语义块转到 3，没找到转到 14。
3．该共享语义块块是否是被省略的，是转到 4，否转到 5。
4．调用后向搜索，成功返回结果转到 5，失败转到 14。
5．恢复 1 中的被省略语义块。
6．返回恢复的结果。
7．部分省略，找出共享的两个块。
8．获取两个块的核心。
9．判断哪个块发生省略，后面句子的块发生省略转到 10，没有发生省略转到 11。
10．调用后向搜索，成功返回共享块转到 11，失败转到 14。
11．分解后面句子的块或后向搜索的结果。
12．判断发生共享的部分，恢复被省略的部分。
13．返回恢复的结果。

14．搜索失败，返回。

4．测试

测试语料选用已做过句类标注的语料，语料来源为互联网，共计 6803 句。语料中发生整块共享的句子对（按小句计）共 3091 个，其中链句 164 个，迭句 2799 个，其它共享句 128 个。发生部分共享的句子对共 101 个，其中迭句半共享 76 个，链句半共享 22 个，其它半共享 3 个。测试只处理迭句、链句、迭句半共享句和链句半共享句。其它共享句和其它半共享句暂不处理。

表 2 省略恢复测试结果

共享类别	实际数量	识别数量	识别正确数量	准确率
+	2799	2799	2794	99.82%
+~	164	164	162	98.78%
+%	76	76	65	85.53%
+~%	22	22	18	81.81%

对测试的结果进行分析，单独的迭句或链句的省略恢复是完全正确的，发生错误的句子都是出现了先部分共享、后整块共享的共享链。当共享链前部的部分共享造成的省略无法正确恢复时，后面整块共享造成的省略也无法正确恢复。部分共享无法正确恢复的情况通过举例说明。

例 5： 联合专家组与青海省农牧、卫生部门 ‖ 进行了广泛的交流， +% 并深入疫区 + 了解 ‖ 疫情发展、兽医防控措施、防控成效和卫生措施执行等情况。

例子中有三个小句，句 1 和句 2 之间为迭句部分共享，其具体共享情况是句 1 的 GBK1 完整，且核心构成为并列结构。句 2 的 GBK1 被省略，其内容应为句 1 的 GBK1 核心两个构成成分之一，但是两个构成成分地位相同，无法确定应该用哪一个构成成分来恢复。句 2 和句 3 形成迭句，共享 GBK1。由于句 2 的 GBK1 无法正确恢复，应此句 3 的 GBK1 也无法正确恢复。

例 6： 一个重要的职责， ‖ 就是 ‖{ 把国家经济社会发展所创造的物质、科技、人力等各种资源 | 转变为 | 保障资源 }， +~% 在必要的时候 ~‖ 迅速转化为 ‖ 后勤保障力。

该例是一个链句半共享句，句 1 的 GBK2 是一个原型句蜕，句 2 的 GBK1 被省略。句 1GBK2 的句类代码为 T4b，是一个变换句，变换句的两个广义对象语义块 T4BC1 和 T4BC2 具有对仗性，区别只是在于具有源流关系。在本例中，无法确定句 1 的 GBK2 的核心，因此无法确定该用 T4BC1 还是 T4BC2 来对句 2 的 GBK1 进行恢复。

例 7： 比如当我们讨论中海油并购优尼科公司一事时 ~‖ ，很多众议员 ‖ 认为 ‖[# 美国企业 ‖~ 在中国石油产业 ~‖ 并无 ‖ 投资 #]， +~% 实际上却有很多；

本例中两个小句形成链句半共享句，但是特别之处在于句 2 的两个 GBK 全部省略，而且全部从句 1 的 GBK2 中恢复。句 1 的 GBK2 是一个块扩，它的 GBK1 和 GBK2 分别对应句 2 的 GBK1 和 GBK2，这种情况比较特殊。

从上面的分析可以看出，恢复部分共享造成的省略，主要难点在于对提供共享信息的语义块的核心进行分析。当语义块是原型句蜕或块扩，或者语义块的核心是并列结构时，比较难于处理，必须再进一步考虑语义块和另一个小句间的概念关联。这将在下一步的工作中进行研究。

与同是使用 HNC 理论的王厚峰的工作相比，王厚峰的研究涉及了所有省略现象，而本文重点研究了语义块共享造成的省略。王厚峰的工作给出了省略消解的规则，但没有给出详细的算法，也没有形成系统，本文则给出了计算机可用的算法并进行了测试。本文的研究范围比王厚峰的工作研究范围要窄，研究的出发点也不相同，但是本文给出的解决方案更具体，更详细，更适合计算机处理，并且通过测试表明了本文的算法是有效的。

5. 结束语

语义块省略在汉语中是一种常见的语言现象，句群中的小句间发生语义块共享则是一种特殊的省略，本章研究了汉语句群中语义块共享造成的语义块省略现象。语义块共享分为完全语义块共享和部分语义块共享，小句间迭句、链句、塔句、环句和其它共享形式等不同的句间关系反映了语义块共享的形式。通过句间关系可以确定前后两个小句发生共享的语义块，再经过对小句的句类格式及对共享语义块的结构分析判定被省略的语义块及进行恢复处理，本章给出了详细的恢复处理算法。其它不是由语义块句间共享造成的省略现象将在进一步的工作中研究。

参考文献

[1]李旺, 李绍滋. 基于 DRT 理论的 汉语省略恢复研究[J]. 计算机工程, 2004, 30(17):39-41

[2]殷鸿, 许威, 赵克, 党建. 基于概念模型的省略恢复研究[J]. 计算机工程, 2007, 33(22):229-231

[3]王厚峰, 汉语指代消解与省略恢复研究[D]. 2000

[4]苗传江, HNC（概念层次网络）理论导论[M]. 清华大学出版社. 2005

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