HNC 句群处理研究新进展*

缪建明　张全
中国科学院声学研究所 北京 100190
E-mail: andrewmjm@hotmail.com

摘 要： 句群是比语句更高一级的语言单位，句群的理解结果直接影响篇章的理解结果。给出形式化的句群处理框架、制定相关的句群处理规则，成为了自然语言处理研究，尤其是基于语义的自然语言处理，亟需解决的一项重要问题。本文在 HNC 语境观的指导下，结合句群处理的新研究成果，对句群的语境单元框架进行了详细地阐述，最终形成了一种形式化表示句群处理结果的新方式。最后，通过真实语料的分析，证明这一表述方式有效且可行。
关键词： 语境，领域，句群

The New Progress of the Sentence Group Processing Based on the HNC Context Theory

Miao Jian-ming Zhang Quan
Institute of Acoustics, Chinese Academy of Science, Beijing, 100190, China
E-mail: andrewmjm@hotmail.com

Abstract: As a higher phase of the natural language processing higher than the sentence, the understanding results of the sentence group immediately influence on the understanding quality. The formalized sentence group processing frame and related processing rules have become the important questions in the natural language processing research. This article combines the new research results of the sentence group processing under the HNC context theory instruction. Through the elaboration on the sentence group unit in detail, the method finally forms the new formalized method of sentence group processing. Finally, this method is proved to effective and feasible through the real language demonstration.
Keywords: context, domain, sentence group

1　引言

随着以信息检索、机器翻译为主要研究对象的计算语言学的发展，使得语言学的研究范围进一步扩大。然而，传统语言学对于句群的研究远远不能满足计算语言学的需要，尤其无法满足计算机处理的需要。首先，目前对句群的研究主要集中在多个句子的句间结构关系上，对关系的如何发现、各句子语义内容并未从计算机处理的角度进行研究，使得这一研究成果很难应用于自然语言处理中。同时，句群的处理离不开对语境框架的构建，而国外的研究更多的是探讨语境在

*本文承国家 973 项目“自然语言理解的交互引擎研究”（2004CB318104）、国家科技支撑计划课题“搜索引擎中的语言翻译基础研究”（2007BAH05B02-05）、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”（O654091431）、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

话语中的作用，而国内的研究对语境的分类则做了充分的研究。框架是非尔莫 Fillmore 从心理学中引入语言学的一个概念，是一种与某些经常重复发生的特定场景有关的知识和概念，是某个物体或事件的典型，它能说明情景的主要特征，可变特征以及经验可能表现的特征[1]。

HNC 理论认为自然语言理解是一个从自然语言空间到语言概念空间的映射过程，两个空间各有自己的一套符号体系，HNC 理论正是通过语言概念空间研究语言现象[2]。本文正是在 HNC 的语境观的指导下，详细阐述了语境单元的框架结构，并最终形成了语言空间中的句群处理形式化结果。

2 句群

传统语言学研究认为，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[3]。HNC 理论认为：句群是 HNC 对段落、篇章处理时在句子和段落之间加上的一个过渡层次，对应到自然语言空间，则由“题”来进行统摄，扣题就自然形成句群。“题”就是指一个特定的概念，而概念则是指 HNC 语言概念基元体系中的某些语义表达。

句群在自然语言中没有约定的标记，难于划分，在初步划分句群时，先采用“虚跨逗号，暂停句号，参考分号、问号和叹号，立足于段落标记”的技术手段划分句群，而句群范围的最终认定还需要靠句群的“题”。这一做法也为计算机获取句群描述中心信息提供了可以操作的线索。由此可见，HNC 所谓的句群和语法研究的句群还是具有很大差别的：其一，语法研究的句群以人理解语句为基础判别句群，而 HNC 句群则以计算机理解语句为基础；其二，语法研究句群以语法逻辑为划分手段，而 HNC 句群则以“题”的不同种类进行划分；其三，语法研究之句群必需包含两个或以上的语句，而 HNC 句群则没有此限制，一个语句中的各个小句表达的中心意思相同，同样可以是一个完整的句群。本文所指的句群，均对应 HNC 之句群。

根据 HNC 理论，单句对应一个句类表示式，句群对应一个或多个以上句类句类表示式，句群中的表示每一个句类表示式的语段称为小句。句群的语义结构依据 HNC 理论可以表示为：

$$SG=\sum([lb]+SCi) \quad \text{其中 } SC=[fKn]+GBK1+EK+GBKm(m=2\text{-}4)$$

句群（SG）由一个或多个句类表示式（SC）构成，句类之间可以用句间逻辑说明符 lb 连接。这是句群的第一级层次关系，其中句间逻辑说明符 lb 可以没有。

3 HNC 句群语境观

语境是语言运用的生命，语言表达、语言领会都离不开语境，任何语用现象都只能生存和存在于特定的语境之中[4]。计算机理解句群不同于人类理解句群，从人类交际来说，语境有广义和狭义之分。广义语境包括两个部分：第一是话语自身，简称上下文 context；第二是话语形成过程的外部环境，简称语域 register，通常就把狭义语境叫做语境。人类交际时上下文与语言环境的分野是清晰的，两者相互耦合形成交际语境。但对当前的计算机来说，语言环境所蕴涵的信息是不存在的，不可能形成交际语境，这就要求计算机句群处理时必须在交际语境基础上进行一定简化，我们称之为交互语境的框架[5]。这一框架就是语境，而框架的各个基本构件即定义为语境单元 SGU。任一个语言段落构成的语境都是由有限的基本构件组合而成，而这些基本构件即为语境单元。语境单元是一个三要素的结构体，三要素的名称分别是领域 DOM、情景 SIT

和背景 BAC [2]。

3.1 领域（DOM）

领域 DOM 描述事件的类型，以 HNC 概念基元符号体系中带有人类活动领域的概念进行分类而得到。领域句类 SCD 则是语境单元可计算的关键因素，因为领域句类 SCD 是有限的，而整个人类表述语境则是无限的，通过领域句类 SCD 的不同组合则可表现无限的语境空间。

领域以事件为中心，描述事件核心所归属的范畴[6]。HNC 理论以人类活动为主体划分为十大领域类，领域的划分为分类提供了一个封闭的 HNC 领域类别空间。领域 DOM 的确定即可对应出句群中“题”的确定，而领域的归属则可最终划分出句群的边界。HNC 理论正是利用人类专家设计完成的领域句类知识为指导，对句群进行深入处理的。

领域的判断准则在 2001 年已经初步形成，获取准则如下[4]：

1）领域信息首先取于 Eg//Ep，不计其它；

2）若 Eg//Ep 无领域信息，则取于 El//Er，不计其它；

3）若 Eg//Ep、El//Er 都无领域信息，则取于 C，不计 B 或 A；

4）最后取于 B 或 A。

经过近几年的领域研究之后，我们发现这一领域信息获取准则仍然存在一定的缺陷。其一，各语义块之间虽然有领域优先选择级别，但是如果我们在领域信息贡献率大的语义块部分获得领域信息之后，不计其它贡献率小的语义块部分的领域信息，往往会在复合领域中丢失部分领域信息，使得领域信息缺失，表述不明确。领域经常存在交叉复合的情况，采用原有的领域认定准则必然造成句群语义信息的丢失，不利于计算机的下一步处理。其二，原有领域信息没有充分考虑到辅块所蕴涵的领域信息，也就没有计算出辅块对领域信息的贡献率大小。辅块虽然弱依赖于句类表达式，但是对语境单元中情景和背景信息的提取来说是至关重要的，原有规则没有考虑辅块的作用，必然也会造成语境单元框架 SGU 的内容缺失。其三，领域是对于句群来说的，句群则对应多个句类表示式，多句类之间存在如何确定句群领域的问题？这一问题在原有准则当中未进行思考，而这显然也是领域认定需要解决的问题，否则对句群领域的处理难以形式化。

在此基础上，经过对真实语料的领域判定研究，我们提出了新的领域认定准则：

1） 句群领域必须充分考虑各句类表达式的领域信息；

2） 对于某一个句类，必须对 Eg//Ep、El//Er、 C、B 或 A、fK 各部分进行领域信息认定，在此基础上按照贡献率大小进行领域复合处理，得到句类领域认定；

3） 对多个认定的句类领域按照领域概念树优先级进行归并，归并应该既考虑领域概念的优先级，同时也要考虑领域底层化，最终实现句群领域确认；

4） 若仍没有领域信息，领域归为一般领域。

3.2 情景框图（SIT）

HNC 理论的情景框架不同于语境学之场景语境。场景语境包括说话人（主体）、听话人（客体）、时间、地点和话题等要素。交际中的言语只有与场景语境相一致、相协调，才会获得得体的交际效果[5]。HNC 理论认为：情景要素 SIT 由领域句类 SCD 描述，以事件的特定参与者为中心，包括各参与者以及他们之间的语义关系。每个领域和相关的句类结合都可以生成一个情景框架，每个情景框架都可以根据句类知识给出预期。领域 DOM 确定后，通过领域句类表示式的

归一化组合构成情景框架。

3.3 背景框图（BAC）

背景框架以事件的背景类型为中心，是指语言的褒贬倾向及时间、空间、方式、参照、目的等辅要素情况，分为事件背景（BACE）和表述者背景（BACA）。事件背景主要包括：事件基本信息；陈述内容的时间、空间、方式、途径、因果、参照等信息，与辅语义块概念同构。表述者背景主要包括：表述者本人的情况，与论述型句群对应。

3.4 框架填写

句群的处理，最终以填写有具体现场数据的语境单元框架作为最终的形式化结果。语境单元信息即通过现场的语句处理结果和预期的语境单元框架知识的整合来实现萃取，其萃取遵循“句群切分—领域认定—领域句类知识获取—框架确定—数据整合（框架内容填写）”这一步骤。具体的算法实现如表 1 所示：

表 1　语境单元萃取实现算法

第一步：　初始化句类分析的处理数据	对句类分析处理得到的词语（短语）按照句类贡献大小进行领域排序，得到词语序列 Wr；
第二步：领域认定	结合 HNC 十大领域概念林的优先级别序列 $D^{1\times10}$，形成 领域序列 $Dom(Wr|D^{1\times10})$；当发现段标记符时，进行对应的预期知识检验，获取进一步的领域认定结果 Dom(i)
第三步：领域知识提取	通过 HNC 领域句类知识库，获取领域序列 Dom(i)对应的领域句类知识；
第四步：整合数据	通过预期领域句类知识和现场的句类分析数据的对照，填写对应的语境单元框架内容，实现语境单元的萃取。

4　实例分析

上面我们详细阐述了如何在 HNC 语境观的指导下进行句群处理的相关内容，下面我们将结合具体的句群实例来说明这一步骤的具体操作实现。

①开场之后~||，火箭||跟着||太阳的节奏[=走=]，++双方||大打对攻。②火箭||毫不示弱，+首节~||不断外线开火，++7 次三分出手||竟然命中 6 个，++其中阿泰斯特和巴蒂尔两人||各投中了 2 个。③{本节|还有|2 分 55 秒}时~||~，{阿泰斯特和布鲁克斯|各投中一记三分}后~||，双方||战成||25-25。④不过太阳||~此后~||连得 6 分，++而火箭||~在 2 分钟内~||一分||未得。⑤在这种快节奏的比赛中~||，姚明||难以发挥||作用，+首节~||5 投 1 中，+得 4 分。⑥火箭||~以 30-33~||落后。　　（《姚鲨对决姚明 20 分 14 篮板 火箭负太阳错失赛区宝座》新浪体育讯）

首先根据句群初步切分原则，将本段分为六个句群。依次进行分析，第一句群，领域词语（短语）有三个：开场、火箭、打对攻，其中“打对攻”为特征语义块，权重最高，但“打对攻”这

一短语的HNC符号具有多领域特性（可能是a32a“艺术表演”、a33t“技艺文化的三个基本侧面”、a42a“进攻与防御”、q730e21“比赛的攻”等），而开场和火箭不具有和“打对攻”的照应性，故本句群领域暂给出领域序列。第二句群，领域词语有首节、外线开火、三分出手、阿泰斯特、巴蒂尔、投中，其中单领域的词语有“三分出手、阿泰斯特、巴蒂尔、投中”等，特别是最后一句“阿泰斯特和巴蒂尔两人各投中了2个”更是直接点明了本句群领域为q730e21“比赛的攻”+a339i\31ckne21-“男子篮球”，由此也可指明第一句群的领域为a339i\31ckne21-“男子篮球”。第三个句群，领域词语有本节、阿泰斯特、巴蒂尔、投中，领域判定类似第二句群，明显可合并2和3句群；第四句群，领域词语有得6分，领域为q730e21“比赛的攻守”；第五个句群，领域词语有姚明、首节、得4分，领域为q730e21“比赛的攻”；第六个句群，领域词语为“30-33”，与非领域词语“落后”搭配，可判断领域为q730e56“败”。词语的相关概念符号均来自于HNC词语概念知识库。

由此本段分为四个句群，第一句为句群1，第二句、第三句为句群2，第四句、第五句为句群3，第六句为句群4。我们仅以句群2为例，进一步说明语境单元框架的填写。

（1）领域DOM认定：话题为q730e21+a339i\31ckne21-，自然语言表述为“男子篮球比赛中的进攻”

（2）情景SIT填写：首先提取对应的领域句类表示式

SCD（q730e21）=Cn-1Cn-2R211-6e21(P10-eb3)X-11*22J++Y30-1(Ya0-1)S0-ac26*22J，

领域预期知识为：

Cn-1:=j109;Cn-2:=j219;RB1=RB2:=(p//pe)q730e31;YB:=pq730e31;YC:=50ac26

SCD（a339i\31ckne21-）为虚化领域节点，不配置相应的领域预期知识。现场的句类分析数据内容为SC=S0J+Cn-11R2J+Ya0S0*11J+Ya0S0*11J++Cn-11Cn-12S0J，这样根据预期的领域知识和现场数据，我们可对应填写SIT内容如下。

Cn-11(首节、本节还有2分55秒时)，Cn-2(无对应内容)，RB1（火箭队）、RB2（无对应内容）、YB（阿泰斯特和巴蒂尔）、YC（投中三分），这里有一个数据整合的准则：无对应内容的可以不填写；对应的情景框架中的语义块内容，需要到对应的现场句类的语义块中寻找（例如领域预期中的RB1到现场句类中的R2J中的RB1中找；领域预期中的YB到现场句类中的Ya0S0*11中找）。（相关的自然语言空间标识符、句类表示式内容可参见附录I和附录II）

（3）背景BAC填写：提取现场句类分析中的辅块信息，Cn-11（首节、本节还有2分55秒时）和Cn-12(阿泰斯特和布鲁克斯各投中一记三分后)提取，同时提取文章的作者和篇章信息（《姚鲨对决姚明20分14篮板 火箭负太阳错失赛区宝座》新浪体育讯、2009年04月02日12:32 ），从而可填写具体的背景内容。

上面我们通过具体实例阐述了如何在HNC语境观的指导下，通过预期的领域句类知识和现场的句类分析数据填写语境单元框架的内容，实现语境单元萃取的过程。我们在目前测试的5000字（11 篇）的真实体育语料中，通过框架填写的步骤，都可实现对应的语境单元框架的填写，在此基础上，通过人工校验填写结果，句群的识别准确率达到了91%，结果令人满意，也可证明这一方式可形成句群处理的有效形式化表述方式。目前HNC理论尚未能完全设计出全部对应领域概念（300多一级领域，50000多领域概念）的语境单元框架知识，但这一框架的设计原则和步骤已经确定，也设计了人类专业活动领域的部分语境单元框架知识库。我们可利用这一知识库，对真实的句群语料进行处理分析，一方面可以检验和完善这些已经设计完成的框架，另一方面也

可以为尚未设计完成的领域概念提供一定的知识积累。未来，我们将尽快建成一个大型的完整的语境单元框架知识库，在更大规模的真实语料环境下进行验证，最终使得计算机的自然语言处理水平达到一个新的水平。

5 总结

本文在HNC语境观的指导下，结合近几年句群处理研究的新成果，提出了语境单元框架这一句群处理形式化表述方式，并给出了具体的填写步骤。这种方法不需要大量真实语料统计形成处理规则，而是通过人类专家设计完成的预期语境单元框架知识和现场分析的句类分析得到的句类知识进行整合，通过合理运用预期知识与现场知识，从而实现计算机对句群的自动处理。这些语境单元框架知识通过领域为分类预先装入计算机知识库，可以很方便地通过领域概念为索引来提取。从具体示例的分析处理结果来看，这种方法能够有效地解决目前地句群处理的问题，为句群处理提供了一种新的形式化方法。

参 考 文 献

[1] 刘澍心. 语境构建论. 湖南:湖南人民出版社. 2006.

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

[3] 吴晨. 浅析句群划分的基本依据. 中文信息处理的探索与实践(第三届 HNC 与语言学学术研讨会论文集). 北京：海洋出版社. 2006

[4] 黄曾阳. 自然语言理解的20项难点. 中国科学院声学研究所内部资料. 2001.

[5] 史秀菊. 语境与言语得体性研究. 北京:语文出版社. 2004.

[6] 晋耀红. HNC(概念层次网络)语言理解技术及其应用. 北京：科学出版社. 2006.

附录 I：

相关的标记符号：

‖	语义块边界标记符号
‖~	辅语义块界标记符号
|	句蜕中语义块的间隔符号
++	列句之间的间隔符号
+	迭句之间的间隔符号
[= =]	特征语义块向后分离成分的间隔符号

附录 II：

相关的句类表达式：

R211-6e21(P10-eb3)X-11*22J=RB1+R211X+RB2

Y30-1(Ya0-1)S0-ac26*22J=YB+Y30S0+YC

S0J=SB+S0+SC

R2J+RB+R

Ya0S0*11J=YB+Ya0S0