# 汉英机器翻译中语境知识的表示与应用*

马红妹　　王　挺　　陈火旺

国防科学技术大学计算机学院

国防科学技术大学计算机学院博士生队　长沙　410073

Email: diane_ma@ynmail.com　wonderwang@163.net

**摘要：**　汉语篇章语境知识的表示和获取有助于汉英机器翻译译文质量的提高。本文首先提出了一种汉语篇章语境知识的表示结构——概念关联层次网络（HRNC）；然后给出了 HRNC 的构造算法，它是在汉语篇章分析的过程中动态建立和更新的；最后讨论了应用 HRNC 来解决汉英机器翻译中的几个实际问题。

**关键字：**　汉英机器翻译　语境　汉语篇章分析　译文质量

# REPRESENTATION AND APPLICATION OF THE LINGUISTIC CONTEXT KNOWLEDGE IN CHINESE-ENGLISH MACHINE TRANSLATION

Ma HongMei　Wang Ting　Chen HuoWang

School of Computer, National University of Defence Technology,　Changsha　410073

Email: diane_ma@ynmail.com　wonderwang@163.net

**ABSTRACT:**　The linguistic context knowledge of chinese texts is useful to improve the quality of the translation in Chinese-English Machine Translation. In the paper, it introduces a structure, Hiberarchy Relation Network of Concept (HRNC), to represent the linguistic context knowledge of the text; and gives an algorithm of creating the representation structure. Finally, it discusses the methods to resolve the problems in Chinese-English Machine Translation with HRNC.

**Keywords:**　Chinese-English Machine Translation , linguistic context , chinese text analysis , quality of translation

## 1. 引言

机器翻译研究历经了 50 多年，形成了众多理论、方法和技术，机器翻译的研究成果在某些领域也得到了一定的应用，如科技资料翻译、天气预报翻译、市场报告翻译等等；但是谈到译文质量，却一直没有取得实质性的进展，尤其是汉英机器翻译译文的质量离实用水平相距还比较远。由于汉语和英语属于不同的语言体系，汉语是一种“意合”式语言，缺少形态变化，常通过词语和句子所含意义的逻辑联系来实现连接；而英语是“形合”式语言，形态变化十分丰富，且不同的形态表示着不同的意义，因此汉英机器翻译相比于英汉机器翻译，它面临的困难更多，如汉语中常出现的省略、照应等语言现象在英译时需要补充和加译。而且，由于汉英两种语言差异比较大，单句范围内的汉语分析是无法得到生成英语译文所需的

* 本课题受国家自然科学基金和 863 高科技计划资助。作者马红妹，1974 年生，博士生，主要研究领域：计算语言学、机器翻译；王挺，1970 年生，博士，主要研究领域：计算语言学、机器翻译等；陈火旺，1936 年生，院士，教授，博士生导师，主要研究领域：人工智能、计算机软件。

所有信息的，如名词的数、动词的时态等，因此语境知识的利用和篇章处理是提高汉英机器翻译译文质量的一种必要的手段。

本文首先介绍了语境及汉语篇章语境知识的表示——概念关联层次网络；然后分析了汉语篇章分析及汉语篇章语境知识的动态获取；最后讨论了应用篇章语境知识来解决汉英机器翻译中英语译文生成所面临的动词时态问题、省略主语的确定问题和冠词的确定问题。

## 2. 语境知识的表示

### 2.1 语境

自然语言的意义极大地依赖于它所使用的环境，这个环境因素便是通常所说的语境（context）。语境可分为上下文语境（linguistic context）和情景语境（non-linguistic context）。上下文语境指与本句话有关系的前后语句，凡是出现在该句话之前的语句都是它的上文，凡是出现在该句话之后的语句都是它的下文。因此汉语篇章构成了其中语句的上下文语境。情景语境指说话时的人物、背景，包括说话双方，涉及的人或物，时间处所、社会环境以及说话双方的辅助性交际手段，包括表情、姿态、手势等非语言因素。在汉英机器翻译中，我们讨论的语境主要指前者。

语境可以使词义具体化、单一化，而且还可以对一句话中省略的成分和隐含的意义给出解释。就像一个词语在不同的句子中可以有不同的意义和句法功能一样，一个句子在不同的篇章和语境中也会有不同的意义，特别是在汉语中，某些句子的歧义往往不能在自身范围内消除，需要把它放到更大的语境中才能理解。

### 2.2 语境知识的表示

句子语义的确定依赖于语境，反过来语境知识结构又是句子语义的获取和积累的结果，两者是相辅相成的。我们使用篇章中出现的概念作为构成语境的基本要素，然后通过提取概念与概念之间的关系来描述汉语篇章的语境知识，由于篇章具有层次结构，因此我们称这种语境知识表示结构为概念关联层次网络（Hiberarchy Relation Network of Concept，简称HRNC）。一般的，篇章可划分为：全文、章节、段落、句子 4 各层次，段落内的概念关联最为密切，如下图所示：

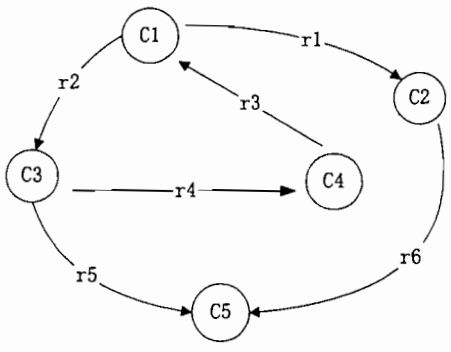


图 1　　段落内概念关联网络

图中的每一个节点是一个概念，具有语法属性、语义属性和所在句子的序号；弧表示概念之间的关联，弧的方向表示概念的支配关系。其中事件类概念节点具有的时态属性用以确

定汉英机器翻译中英语译文的时态，实体类概念节点具有的限定属性用以确定译文中名词的单复数。描述篇章语境知识的概念关联层次网络的定义如下：

HRNC=<$V_T$, R, H>

其中，$V_T$是节点集，表示篇章中的概念；R 是边集，表示概念与概念之间的关联；H 是篇章的层次结构。

$$\forall v \in V_T,\ v = \begin{bmatrix} index & <n1,n2,n3> \\ root & word\ form \\ cat & conecpt\ catigory \\ syn & syntex\ attribute \\ sem & sematic\ attribute \\ restrict & yes/no \\ tense & past/present/future \\ aspect & progressive/perfective \end{bmatrix}$$

$\forall r \in R$，$\exists v1, v2 \in V_T$，有 r(v1, v2)，表示概念 v1 和 v2 具有关系 r。

R 主要包括以下关系：

设有两个概念 $c_1$、$c_2$，两个事件类概念 $e_1$、$e_2$，

（1） 上下位关系，记为 Hyponymy（$c_1$，$c_2$），$c_1$是 $c_2$的下位；

（2） 同义关系，记为 Synonymous（$c_1$，$c_2$），$c_1$与 $c_2$意义相同；

（3） 材料成品关系，记为 Material_Product（$c_1$，$c_2$），$c_1$是 $c_2$的组成材料；

（4） 部件整体关系，记为 Part_Whole（$c_1$，$c_2$），$c_1$是 $c_2$的部件；

【例 1】我刚买了一辆自行车，没骑多久，车胎就坏了。

则有 Part_Whole（车胎，自行车）。 实体类概念之间的上下位关系、材料一成品关系、部件一整体关系体现了篇章词汇语义上的衔接，可以用来说明名词的特指性。

（5） 联想相关关系，记为 Correlation（$c_1$，$c_2$）；

联想相关关系有利于多义词词义的选择。如词语“案”有多个义项：

a） SEM=fact|事情,#police|警（该事情与警方有关）；

b） SEM=furniture|家具,@put|放置（该家具是放置事件的工具）；

如果在一篇文章中出现了“警方”、“刑事组”等相关概念，而在另一篇文章的上下文中出现了“放”、“书”等相关概念，那么根据联想相关关系，“案”在第一篇文章中应该选择义项 a)，在第二篇文章中则应该选择义项 b)。

（6） 动态语义角色关系，记为 Semantic_Role（$c_1$，$e_1$），$c_1$是 $e_1$的动态语义角色；

动态语义角色关系是在事件概念与其所作用的实体概念之间建立连接，动态语义角色关系可以用来判断作为动态语义角色的实体概念所对应的名词的特指性。

【例 2】已经到送信的时间了，邮递员为什么还不来？

虽然不知到“邮递员”是指哪一个人，但是他相对于事件“送信”是确定的，具有特指性。动态语义角色的语义限制也即动态语义角色的值域还可以用来确定省略的成分。

【例 3】我擦了一把汗，继续工作。

第二个分句的事件“工作”省略了作为施事的动态语义角色，它同时又是第二个分句中充当主语，由于“工作”的施事动态语义角色限制为“生物”，所以选择第一个分句中出现的实体概念“我”作为添加的主语。

（7） 重复同指关系，记为 Overlap（$c_1$，$c_2$）；
重复同指关系指概念 $c_1$ 和概念 $c_2$ 所指相同，如果 $c_1$ 和 $c_2$ 分别是两个句子的主语时，则这两个句子的主语相同，说明它们是对同一主题的不同侧面的描写。
【例 4】 (1)我们用歌声送别就要离去、也许永远不会再见的同志。(2)我们用歌声欢迎东方战线上传来的好消息。

（8） 空位指向关系，记为 Gap_Reference（$\Phi$，$c_1$）；
空位指向关系指空位概念 $\Phi$ 与上下文中其它概念发生的指向关系。若句子 j 的空位概念 $\Phi$ 指向句子 i 做主语的概念 $c_1$，空位概念 $\Phi$ 又是句子 j 省略的主语，则 Gap_Reference（$\Phi$，$c_1$）说明两个句子存在主语省略继承关系。
【例 5】(3)我是中国人。(4)$\Phi$来自北京。
(4)中省略的主语继承了(3)的主语“我”，存在空位指向关系 Gap_Reference（$\Phi$，我）。

（9） 时间继承关系，记为 Time_Inherit（$e_1$，$e_2$）；
事件 $e_2$ 位于事件 $e_1$ 的下文，如果事件 $e_2$ 没有明显的时态属性，则事件 $e_2$ 继承事件 $e_1$ 的时态属性。
【例 6】(5)昨天我们一起去郊游，(6)大家玩得很开心。
(6)中的事件“玩”没有明显的时态，于是继承(5)中事件“去”的时态（一般过去时）。

（10）时间先后关系，记为 Time_Precede（$e_1$，$e_2$）；
事件 $e_1$ 先于 $e_2$，如果事件 $e_2$ 为过去时，则事件 $e_1$ 一定为过去时。
【例 7】(7)她打开书包，(8)取出课本，(9)读了起来。
事件“打开”先于事件“取”，事件“取”又先于事件“读”。若判定事件“读”的时态为一般过去时，则可以确定事件“取”和“打开”的“时”都为“过去时”。

H = <n1, n2, n3>，是篇章的物理结构划分，n1 表示篇章包含的章节个数，n2 表示一个章节包含的段落个数，n3 表示一个段落包含的句子个数。

## 3. 汉语篇章分析与语境知识的获取

### 3.1 汉语篇章分析与语境知识表示 HRNC

汉语篇章分析以句子分析为基础，但又不是句子分析的简单相加。汉语篇章分析中的句子分析需要查找 HRNC，也就是说，HRNC 影响着句子的分析；反过来，分析后的句子则需要将其核心概念及其属性写入 HRNC 并更新 HRNC 的内容。这样 HRCN 所表示的语境知识便随着篇章分析的进行而逐渐丰富。汉语篇章分析与 HRNC 的关系如下图：

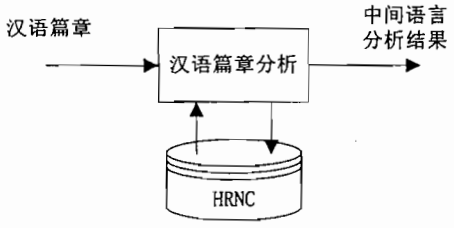


图 2　　汉语篇章分析与 HRNC

当汉语篇章仅由一条语句构成时，此时 HRNC 便等于该句子的语义表示 Sem（Sentence semantic）；当汉语篇章包含 n 条语句时，汉语篇章的上下文语境结构则是由这 n 条语句的语

义动态构造而成。例如，第（i）条语句的语义 $Sem_i$ 依赖于前（i-1）条语句构成的语境结构 $HRNC_{i-1}$ 而确定，然后将确定下来的第（i）条语句的语义 $Sem_i$ 加入 $HRNC_{i-1}$ 并更新 $HRNC_{i-1}$ 的相关内容形成由前（i）条语句构成的新的语境 $HRNC_i$ 。

假设 T(Text)表示汉语篇章，S(Sentence)表示汉语语句，则：

T ::= S|T×S 或 T ::= S|S×T

其中，| 是并操作符，×是笛卡儿乘积操作符，::= 是定义符号。

假设 C（Context）表示语境知识表示结构 HRNC，则 C 的动态构造过程 I 定义如下：

（1） I：S×C→C，当 T ::= S；

（2） I：T×C→C，当 T ::= T×S 或 T ::= S×T；

对于∀t∈T，s∈S，c∈C，有：

I ((t, s), c) = I (s, I (t, c)) 或

I ((s, t), c) = I (t, I (s, c)) .

## 3.2 汉语篇章语境知识表示 HRNC 的构造算法

S1. 初始时 HRNC 为空；

S2. 分析汉语篇章的语句，若分析完毕则退出；

分析当前语句，当无法确定概念的属性时，如实体的限定性、事件的时态等，查找 HRNC，通过与 HRNC 中的概念进行比较，确定当前语句中概念的语法属性、语义属性等；

若 HRNC 为空，则采用缺省值输出；

S3. 将当前语句的分析结果加入 HRNC；

a) 提取当前语句中的实体和事件概念加入节点集 $V_T$；

b) 建立当前语句的概念与节点集 $V_T$ 中已有的概念之间的关联，加入概念关联集 R；

c) 修改或添加 HRNC 中已有概念的节点属性；

S4. 转 S2。

【例 8】 王成是一位作家。他写了二十多部书。

分析第一句之后，HRNC 为：

| 索引 | <n1, n2, 1> | | |
|---|---|---|---|
| $V_T$ | v1 | v2 | v3 |
| | Index <n1, n2, 1><br>Root 王成<br>Cat entity<br>Syn subject<br>Sem human<br>Restrict no | Index <n1, n2, 1><br>Root 是<br>Cat event<br>Syn predicate<br>Sem isa<br>Tense present<br>Aspect simple | Index <n1, n2, 1><br>Root 作家<br>Cat entity<br>Syn object<br>Sem human<br>Restrict no |
| R | | | |

分析第二句之后，HRNC 为：

| 索引 | <n1, n2, 2> | | |
|---|---|---|---|
| $V_T$ | v1 | v2 | v3 |
| | Index <n1, n2, 1><br>Root 王成<br>Cat entity<br>Syn subject<br>Sem human<br>Restrict no | Index <n1, n2, 1><br>Root 是<br>Cat event<br>Syn predicate<br>Sem isa<br>Tense present<br>Aspect simple | Index <n1, n2, 1><br>Root 作家<br>Cat entity<br>Syn object<br>Sem human<br>Restrict no |
| | v4 | v5 | v6 |
| | Index <n1, n2, 2><br>Root 他<br>Cat entity<br>Syn subject<br>Sem human<br>Restrict no | Index <n1, n2, 1><br>Root 写<br>Cat event<br>Syn predicate<br>Sem write<br>Tense present<br>Aspect perfective | Index <n1, n2, 1><br>Root 书<br>Cat entity<br>Syn object<br>Sem book<br>Restrict no |
| R | Correlation（v3，v6）<br>Semantic_Role（v6，v5） | | |

## 4. 语境知识在汉英机器翻译中的应用

由于汉语和英语差异比较大，汉语缺乏严格意义上的形态标志而英语形态变化又十分丰富，这使得汉语的分析以及汉英转换都面临许多难题，仅限于句子的汉语分析方法是不可能获得英语生成所需要的所有信息的，时态、单复数、冠词等问题均需要借助汉语篇章语境知识才能得到深入的分析。我们将从篇章角度分析汉语，应用汉语篇章语境知识表示 HRNC 解决以下问题：

### 4.1 动词时态

汉语没有时态变化，所叙述事件的时间特性是借助时间短语或篇章的语义连贯性来表达的。汉语的时间系统是一个“词汇语法范畴”，包括时相、时制和时态，时相和时态可以通过动词以及句中的时态标志来确定，而时制结构则需要对文本中的时间短语进行分析（另文详述）。如果当前句子出现时间短语，则当前句子采用此时间短语所表示的时态；如果当前句子不包含任何时间信息，则查找 HRNC，查找的原则为优先考虑维持段落中前一句的时态属性；如果在 HRNC 中找不到任何可以确定动词时态的信息，则根据篇章的主题和类别，按照下列规则进行处理：

a) 若描写的是特定的人物则采用过去时；
b) 若描写的是客观的事物和规律则使用现在时；
c) 示范和使用说明使用一般现在时；
d) 内容简介用一般现在时；
e) 报刊标题、图片说明等使用一般现在时和现在进行时；

### 4.2 无主句的主语确定

汉语是主题显著型语言，为保证篇章意义的连贯，同一主题再次出现时常常被省略，表现在句法结构中为常省略主语；而英语是主语显著型语言，主语对于句子结构和词语选择具有统帅作用，省略主语的确定和增译对于谓语动词形态的生成以及整个语句的分析都是十分重要的，翻译时需要根据上下文语境确定省略的主语。

当发现句子缺少主语时，查询 HRNC 中的实体概念，经过语义约束检查和匹配，获得合适的主语进行填充。

如以下几个句子，后一句的主语均来自前一句的主语。

【例 9】 你儿子多大了？七岁了。

How old is your son? *He* is seven.

【例 10】太阳象个红球，慢慢地升起来，发出淡淡的光。

Sun is like a red ball. *It* rises slowly. *It* sends out weak rays.

### 4.3 冠词的选择

冠词的选择不仅和汉语名词有关，还与该名词相对应的英语译词有关，英语译词的可数性、单复数、汉语名词的特指性以及名词前的限定词等都影响着冠词的选择。

首先根据篇章分析确定名词的特指属性和限定性。如果该名词在上下文中是特指的或是作者与读者双方所共知的，则它是限定的。此外，当名词前有数量词和指示词（如“两件”、“那个”）、物主代词（如“我的”）等修饰词或是该名词是专有名词时，它也是限定的。然后确定译词的可数性和单复数。

如果确定该名词短语是限定可数的，而且前面没有其它的限定修饰语，则加上定冠词 the；如果是非限定可数单数名词短语，则加上不定冠词 a/an；如果是非限定可数复数名词短语或是不可数名词短语，则加上零冠词Φ。

## 5. 结束语

汉语是一种语义型语言，汉语的分析较依赖于语义、语境知识的应用。我们提出的概念关联层次网络便是一种汉语篇章语境知识的动态描述，它在篇章分析的过程中不断传递信息，积累信息，体现出了篇章语义的连贯性。汉语篇章语境知识的表示和获取将有助于汉英机器翻译译文质量的提高，为此，我们进一步讨论了应用汉语篇章语境知识来解决汉英机器翻译中的几个实际问题。

## 参考文献


[1] 胡壮麟，语篇的衔接与连贯，上海：上海外语教育出版社，1994

[2] B. Johan and M. Alice, HANDBOOK OF LOGIC AND LANGUAGE, MIT PRESS, 1997.

[3] 黄曾阳, HNC（概念层次网络）理论. 北京:清华大学出版社. 1998. 11

[4] 贾彦德，汉语语义学，北京大学出版社，1999

[5] 赵世开，汉英对比语法论集，上海外语教育出版社，1999

[6] 董振东，董强，知网，http://www.keenage.com/html/c_index.html

[7] 周会平，基于中间语言汉英的翻译系统 ICENT 的研究与实现，博士学位论文，国防科大研究生院，1999