基于汉英机器翻译的名词回指分析*

——— 句组研究之二

侯　敏[1]　孙建军[2]

([1]北京广播学院 应用语言学系　[2]北京迈创易达有限公司)

Email: houminxx@263.net　mltran@263.net

摘要： 回指是语篇衔接的重要手段，其中名词回指对机器翻译会产生一定的影响。本文在详细分析各类名词回指的基础上，指出其中三类对机器翻译造成的障碍，并提出在句组层面上解决这些问题的算法。

关键词： 汉英机器翻译；名词回指；句组

An analysis of nominal anaphora in view of Chinese-English MT

Hou min*　Sun Jianjun**

(* Applied Linguistics Department, Beijing Broadcasting Institute **Beijing MulTran Technology Corp)

Email: houminxx@263.net　mltran@263.net

Abstract : Anaphora is an important means of discourse cohesion, and nominal anaphora would influence the quality of MT to a certain degree. This paper analyzes all types of nominal anaphora in detail, and points out that there are three of them which would make a trouble to MT. The author proposes some solutions at the level of sentence group.

Keywords : Chinese-English MT , nominal anaphora , sentence group.

一、　话语分析、名词回指与机器翻译

在风雨中走过了五十几年历程的机器翻译，由于人们的渴盼和市场的需求，尽管步履蹒跚，还是摇摇晃晃地迈入了实用化阶段，走进了市场，来到了用户的面前，这就决定了它面对的只能是人们实际用来进行交际的话语，是活生生的、语境化（contextualized）的语料。

以往的机器翻译系统在处理语言时大都是以句子为单位的，对这种非语境化（decontextualized）的句子进行结构及语义的分析研究显然是十分重要的，因为它是构成

* 本文承国家广电总局科研项目(项目号 bw9943)资助。 刘海涛教授对本文提出了很好的修改意见，谨致谢忱。

话语的基本要素。然而我们还必须看到，话语并不是一组句子的简单拼接，话语的形成还有它自身的规律，其中最重要的两种手段、同时也是人们研究得比较多的两个概念就是“连贯”和“衔接”。任何一段人们说/写出的正常的自然话语，都必须是连贯的，否则就达不到交际的目的，连贯是听话人理解话语的前提和假设；而衔接是说话人在产生话语时有意无意运用的能帮助听话人理解这种连贯的手段。（孙玉，1997）回指就是话语衔接的最重要方式之一，有着很强的语篇功能。

在话语中，一个人或一个事物往往要被多次提到，第一次提到时所用的表达式叫先行词，后面出现的与它同指某一事物的表达式叫回指。先行词与回指的相互照应形成了话语意义的连贯。陈平（1987）认为，汉语中的回指形式有三种基本类型：零形回指（zero anaphora）、代词性回指（pronominal anaphora）和名词性回指（nominal anaphora）。如：

(1)每当这时候，木匠婶$_1$总是不声不响地把我领到她$_1$家里，$\varnothing_1$为我擦去满脸的泪痕，$\varnothing_2$为我端上可口的饭菜；夜深了，她$_2$还陪着我说话，$\varnothing_3$哄着我睡觉。在那段艰难的日子里，是木匠婶$_2$给了我们一家最难得的安慰和帮助。　　鲁景超《爱没有终结》

其中，“木匠婶$_1$”是先行词，后面的“她$_1$、她$_2$”是代词性回指，“$\varnothing_1$. $\varnothing_2$. $\varnothing_3$.-是零形回指，“木匠婶$_2$.-是名词性回指。在这三种回指中，代词回指在人际交际中十分重要，因为准确地找到代词回指的先行词是理解话语的基础和关键。然而，由于机器翻译的目标主要是取得与源语同义的目标语译文，而汉语与英语中的代词基本上是一一对应的（如她：she；他：he；它：it；他们\她们\它们：they；这：this；那：that 等等），只要区分好人称代词和物主代词就可以了，所以它在汉英机器翻译中没有太大的问题。而名词回指和零形回指内容较复杂，它们中都有相当一部分内容会给机器翻译带来障碍，而且这些问题只有在句组平面上才能解决。本文重点讨论名词回指的问题，零形回指的问题将另外撰文讨论。

二、　名词回指的类型分析

在进行具体论述之前，我们有必要先区分两个概念。1.语言意义；2.言语所指。语言意义是指一个词汇符号在语言系统中的价值，它所联系的是一类客观事物，没有具体的指称作用。语言系统中的符号（词）正是由于意义不同所产生的区别性才得以相安存在，形成一个系统，如“检察院”、“院”就是这样的有着不同意义的两个语言符号。但语言中的符号一旦进入言语之中，就不一样了，语境把语言中的词与客观世界中的事物联系起来，使得它们有了具体的所指，而且在语境的作用下，在语言中意义不同的词完全可能指称同一个事物。如在下面的句子中，“检察院”和“院”指称的就是同一个客观事物。

(2)（山东省荣成市）检察院在反贪污贿赂斗争中，严把办案质量关，取得较好社会效益。自一九九七年以来，该院共立案查处贪污贿赂等经济犯罪案件七十起，无一起提出申诉。

《人民日报》2000.09.31。

显然，词的语言意义和言语所指之间不具有同一性。正是由于语言意义和言语所指的这种对立，才使得名词回指有了各种不同的类型，否则，名词回指就只能是同形的了。

名词回指与先行词所指称的事物是相同的，但名词回指和先行词在表达式即所采用的语言形式上却既可以相同又可以不同，因此廖秋忠(1986)首先根据名词回指表达式与其先行词表达式之间在语言形式上的关系将其分为三类：1. 同形、2. 局部同形、3. 异形，然

后再根据二者之间的语言意义上的关系把异形分为四类。为了讨论的方便，我们比照徐纠纠（1999）的办法，将汉语中的名词回指直接分为五类：同形、局部同形、同义、上义或下义、部分或整体。其中前两类是从“形”上着眼的，后三类是从“意”上着眼的。下面具体分析。

1. 同形

(3) 一个小伙子暗恋着一个女孩$_1$。女孩$_2$是他的同事，他们在一个办公室里工作。

《浪漫》（《高级汉英语篇翻译》以下简称“高”）

其中女孩$_1$是先行词，女孩$_2$是名词回指，二者的形式是完全一样的。不同的是，女孩$_1$因为在话语中是第一次提到，用的是无定形式，前加“一个”，女孩$_2$前面没有任何修饰成分，但很明显是有定的。

同形的名词回指在汉语和英语中的使用率都是比较高的，根据陆振慧（2002）的统计，这两种语言中的同形回指在所有名词回指中都占到一半以上，相比之下汉语更高一些，其中有生命体的同形回指竟占到所有名词回指的80%。徐纠纠（1999）的统计中，汉语同形回指也占到名词回指的52.74%。我们统计的数据比他略高，占到56%。尤其是，回指离先行词距离越远，同形回指使用的可能性就越大。

2. 局部同形

局部同形回指从逻辑上说可以细分为三类：（1）比先行词少的，可称之为简省类；（2）比先行词多的，可称之为增补类；（3）与先行词部分相同、部分不相同的，可称之为变换类。这三类在语言中出现的频率是不一样的。增补类是极个别的；变换类也不多，且大都是表示人名的，先行词往往是人名的全称，回指是姓加上称谓。如先行词是“艾炳”“赵忠祥”，回指是“老艾”“赵老师”。简省类在局部同形的名词回指中占多数，情况也比较复杂。

(4) 北京大学在蔡元培治校及以后相当长的一段时间里，除了正式招收的学生，还有许多交费的旁听生。北大之大，让你不得不叹服。 《人民日报》2000.10.5

(5) 太子河区人民法院的干警们在执法过程中始终把严格执法与热情服务结合起来，把申判工作与法律宣传教育结合起来。该院所辖地区是城乡结合部，院领导针对实际情况，适时组成巡回法庭，深入乡村开展工作，把法律知识送到田间地头，送到农家炕头。《人民日报》2000.11.12

(6)我是收煤气费的，查一下煤气表。 -------好，表在这儿，您自己看吧。

(7) 那些日子，我的恩师汪曾祺先生刚刚去世，曾祺先生是沈从文先生的入室弟子，兆和先生执意去向曾祺的遗体告别，她说，从文走的时候，我很冷静，没有哭；但这次曾祺离去，我不能接受，因为实在是太突然、太意外了。 《纯净》（高）

(6)是普通名词的简省，(7)是人名称谓的简省。（4）和（5）表面上看来都是机构名称的简省，却有着本质的不同。（4）中的“北大”是个缩略语，“北京大学”和“北大”无论在句中还是在句外，意义都是一样的，（5）中的“该院”或“院”却不是缩略语，“太子河区人民法院”和“该院”的同指关系仅仅在这一语境中存在，离开这一语境，这种联系就不存在了。廖秋忠早在1986年就曾指出过这种现象，并提醒我们注意。

按徐纠纠的统计，这种局部同形的占名词同指的16.85%，我们的统计更低，仅占10%。

3. 同义

同义的名词回指可以是先行词的同义词，如（8），也可以是与先行词同指的短语，如（9），还可以是对先行词的一个比喻，如（10）。

(8) 童话大师安徒生曾经带一个小女孩到林中空地采蘑菇.他事先在有些蘑菇下藏了一件件小东西——一颗包着银纸的糖果、一束蜡制的小花、一枚别针或丝带，等等。小姑娘发现了许多意外的惊喜。安徒生告诉她这些东西都是土地之神放在那里的。《读者》2002.23

(9) 所有关于辛吉斯是否真将退役的传言和揣测都应该停止了，这位22岁的瑞士天才球星7日确定无疑地表示，她8年的职业网球生涯已经画上句号。《环球》2003.2.10

(10) 马上就要过40岁生日的乔丹早些天已经放出话来，如果不能入选首发阵容，就将自动退出今年的全明星赛。但接下来发生的一幕幕情景却让人颇受感动，当飞人第14次、也是首次以替补身份入选全明星阵容时，恰恰是抢走他首发位置的麦格雷迪站了出来："我愿意把自己的首发位置让给迈克尔。"

《体坛周报》2003.2.7

同义的名词回指所占比例仅次于同形的名词回指，在徐文的统计中占21.69%(他不包括比喻的)，在我们的统计中占16%。

4. 上义或下义

上义-下义关系（hyponymy）是指词项间具体与一般的涵义关系，即具体义包含在一般义之中。具体义是下义，一般义、概括义是上义。如"猫"是"动物"的下义词，"椅子"是"家具"的下义词。在话语中，上下义词也可用来指同。大多数情况是先行词为下义词，回指是上义词，如（11），但也有先行词是上义词，回指是下义词的，如（12）。

(11) 总攻打响那天，团长到了主攻连，拍着排长的肩膀说："老战友，你必须在15分钟内拿下1号高地并守到天亮，人在阵地在，人不在了呢，人不在阵地也要在——不！你要给我活着给我回来！"

《"兵"和"酒"》（高）

(12) 古代孟母三迁是为了怕孩子受坏影响，要为自己就没有必要逃避了，后来孟子长大成人后也没听说孟母再搬家。 《我喝我的清茶》（高）

由于上下义关系仅仅是从"义"上着眼，不考虑"形"的因素，而"局部同形"又仅仅考虑"形"的因素，不管"义"如何，这是两个不同角度的分类，所以二者难免会有交叉的地方。如"公鸡"与"鸡"、"朝阳医院"与"医院"，从形式上看，可以说是局部同形，从意义上着眼，又是上下义关系。如果按照廖秋忠的办法，第一层次以形式为准划分，第二层次再以意义为据划分，就不会出现这样的问题了。

5. 部分或整体（包括领有关系）

整体-部分（meronymy）也是词项间的一种涵义关系。如"汽车"和"轮子"之间，"汽车"是整体，"轮子"是部分。下面例(13)中，"父亲"是整体，"眼睛"、"泪水"是部分。领有者和领有物之间的关系与之很相近，我们也把它包括在这里。如（15）。

(13) 父亲听不到，但他知道了我的意思，眼睛里放出从未有过的光亮，泪水和着高粱酒大口地咽下。

《我和我的哑巴父亲》，读者，2002.23

(14) "它不咬人么？""有胡叉呢。走到了，看见猹了，你便刺。这畜生很伶俐，倒向你奔来，反从胯下窜了。它的皮毛是油一般的滑....." 鲁迅 《故乡》

(13)与此同时，郭建伟绕到冰窟南侧假山上，衣服也顾不得脱便跳入湖中，向孩子扑去。

《人民日报》2000.1.12

严格说来，这种关系与上下义关系不一样。“一号高地”和“阵地”之间，离开语境，是具体和一般、下义和上义的关系，但在（11）中，它们确确实实指的就是同一个地方，这种指同，是毫无疑义的。而“父亲”与“眼睛”，可以说，在任何情况下，都不会所指完全相同，“眼睛”只能是“父亲”的一部分。但恰恰是由于二者之间的这种明显的语义联系，在话语中表现出很强的语篇功能，成为了词汇衔接的重要手段之一，所以，我们也把它放在了回指中，但它应该是回指中比较特殊的一类。

在徐文中，部分-整体和上义-下义是作为一类来统计的，占 8%。我们的统计中二者共占 17.9%，其中上义-下义类占 7.5%，部分-整体类占 10.4%。

三、 各类名词回指与机器翻译的关系

名词回指是各种语言中共有的普遍现象(陆振慧 2002)。对机器翻译来说，它是既有利也有弊。就汉英这一语言对来说，由于名词回指在选用类型及使用频率上有所不同，再加上语义系统之间的差异造成的词汇对应上的错位现象，会给机器翻译带来一定的障碍；但另一方面，我们也可以利用某些具有语篇功能的名词回指来弥补系统的不足，以提高机器翻译的准确率。不同类型的名词回指，与机器翻译的关系是不一样的。

1.对机器翻译不构成障碍的名词回指类

在名词回指中占 70%以上的同形、同义的名词回指对机器翻译基本不构成障碍，一般照直译来就可以。而且我们还可以利用同形回指中回指与先行词的同形关系作为词语切分的条件。因为机器不像人，在汉语分词的过程中无论用统计法还是规则法，本来上下文中同形的词语都可能由于前后符号的影响而错分，产生出让人啼笑皆非的结果。如董振东（1999）提到的那五个“薄熙来”。如果我们能利用这种同形关系，在分词时首先将同形词分出来，然后再作其他的切分或合并，也可以提高系统分词的正确率从而提高整个系统的译准率。

2.对机器翻译构成障碍或有影响的名词回指类

局部同形中的简省类、上义或下义与部分或整体这三种名词回指有的对机器翻译有直接的影响，有的会给机器翻译带来较大的障碍。

先说局部同形中的简省类。上面我们说到简省类中可以分成两种情况：一种是先行词为原形式，回指为缩略语，如“北京大学”和“北大”、“复旦大学”和“复旦”，它们之间的联系已经固定下来，无论在静态的语言和动态的话语中，意义及所指都是相同的，这类情况不会给机器翻译带来障碍；另一种就不同了，作为先行词的一部分的回指不是缩略语，而是一个独立的词，如“太子河区人民法院”与“院”、“煤气表”与“表”，这类回指会给机器翻译带来很大障碍，在句子语法中根本就无法处理。当然也不是所有的这类简省都这样，如“朝阳医院”与“医院”，就没有问题。看来，有问题的是那些简省程度比较高且回指本身又是多义词的情况。

上义或下义的同形回指也可以分为两种情况：一种是汉语的上下位词与英语的上下位词是一一对应的，如“一号高地”与“阵地”，这种情况一般不会给翻译带来很大的障碍。另一部分是汉语的上下位词与英语的上下位词不对应的。如汉语中“学校”与“大学、中学、小学”是上下位词，但在英语中与它们对应的“school”与“university 、middle school 、primary school”却不是上下位词。这类情况如果不加处理，很难得到正确的英语译文。

部分与整体的同形回指也会给机器翻译带来影响。部分属于整体，这种意义关系是确定无疑的，但在语言中是不是一定要用形式把这种关系凸现出来，不同的语言是不一样的。英语的习惯是，表示部分或所有物的名词前往往要有物主代词，即这种所属关系是显现的；而汉语却不一定，有时是显现的，如（14），但更多的时候是隐含的，如（13）。那么，在翻译这种句子的时候就要注意两种语言的不同习惯，考虑到英语译文中物主代词的补充，否则也不会得到高质量的译文。

四、 汉英机器翻译中解决名词回指问题的算法

从上面的统计和分析来看，对机器翻译构成障碍或有影响的名词回指类在所有的名词回指中仅占不到30%，其中还有一部分是不存在翻译障碍的。据我们的估算，对翻译有影响的名词回指在所有名词回指中不会超过15%。然而，就这15%，如果不解决，给机译译文的质量也会带来一定的影响。

这些问题由于是在话语中产生的，所以，在句子平面上是无法解决的，只有在句组平面上，尚有望解决。我们解决这些问题的思路是：首先，要以句组作为系统处理的单位[侯敏等，2002]，也就是说，在系统处理某一个句子时，它必须要同时看到前后句子的信息，以前后句子、尤其是它前面的句子作为它的语境，然后，把上述影响翻译的名词回指问题分成两类，采用不同的办法分别处理。

一类是简省类和上下义类，这两种本来就是交义的，问题的性质基本相同，可用一种算法来处理。在句组的环境下，我们可以使用在词条下做规则的办法，在“院”“学校”这样的词条下做一条规则，使它能前溯回访它的上文，如果能找到规则中给定的它所指称的那个词，就以那个词的词义替代它的词义，否则不变。好在汉语中这类和英语错位的上下义词并不太多，是可以枚举的[侯敏等，2002]。如：

院：科学院、法院、检察院、疗养院、医院、设计院……

校：大学、小学、中学、技术学校……

学校：大学、中学、小学、技术学校、学院……

表：煤气表、电表、水表、温度表……

车：公共汽车、卡车、小轿车、摩托车、自行车、火车、三轮车、手推车……

店：饭店、旅店、杂货店、小吃店、商店、服装店……

库：汽车库、军械库、数据库、语料库、书库、粮库、水库……

所：研究所、诊疗所、派出所、变电所、交易所、招待所……

团：代表团、歌舞团、马戏团、话剧团、观光团、旅游团、考察团……

……

另一类是部分-整体类。这个问题的关键是如何在没有领属词的表部分或表所有物的同

指名词前补充出正确的物主代词。那么，可以先根据词典中给出的信息确认同指是生命体的一部分还是非生命体的一部分，然后往同找到它的先行词，根据先行词的性质（单复数、性别）决定采用的物主代词形式。

上面我们以汉英机器翻译为背景，比较详细地讨论了名词同指的各种情况及对机器翻译的影响。其中很多都涉及到概念之间的关系。要想真正从根本上解决这些问题，首先需要建立一个能准确描述这些概念之间关系的知识库作为支撑，否则，即使在句组平面上也是无法解决的。

参 考 文 献

[1] 孙　玉　（1997）《试论衔接与连贯的来源、本质及其关系》，外国语，1997 年 1 期。

[2] 董振东　董强 (2001)《机译研究中的一些误区》，自然语言理解与机器翻译，清华大学出版社 2001 年。

[3] 陈　平（1987）《释汉语中与名词性成分相关的四组概念》，中国语文，1987 年 2 期。

[4] 陈　平（1987）《汉语零形回指的话语分析》，中国语文，1987 年 5 期。

[5] 廖秋忠（1986）《现代汉语篇章指同的表达》，中国语文，1986 年 2 期。

[6] 廖秋忠（1987）《篇章中的管界问题》，中国语文，1987 年 4 期。

[7] 徐纠纠（1999）《叙述文中名词回指分析》，语言教学与研究，1999 年 4 期。

[8] 陆振慧　（2002）《英汉语篇中指同表达的对比研究》，外语教学与研究，　2002 年 5 期。

[9] 黄南松（2001）《现代汉语的指称形式及其在篇章中的运用》,世界汉语教学,2001 年 1 期。

[10]侯　敏　孙建军（2002）《面向汉英机器翻译的句组研究》，机器翻译研究进展，电子工业出版社。

[11] Dong zhen-dong , Bigger Context and Better Understanding -- Expectation on Future MT Technology,　Proc. of the International Conference on Machine Translation & Computer Language Information Processing　pp. 17-25, 1999/6