代换法--基于句型和短语的翻译方法

刘孝叔

摘要：代换法是本文作者为小型翻译系统提出的简捷了当的翻译方法。它把源句与正确译句结构之间的对应关系写成带普遍性的转换式，从而使译句的结构在翻译开始时就成为已知。1989 年以来各种基于语料库的翻译方法先后问世。这些方法的生命力在于他们是以经过审订的译句来告诉计算机应该译成什么样子。代换法也是把正确的译句结构告诉计算机，只是这结构是形式化的，更容易执行。

本文介绍了转换式是如何得来的，代换法如何运用转换式来进行翻译。本文也介绍了代换法中尚待研究解决的问题，和笔者对这些问题的浅见。请读者指教。

Conversion Approach -- An approach based on Sentence Structure and Phrase Structure Patterns

Liu Shiao-shu

Abstract: Conversion approach was proposed by the author with the objective of simple andstraight forward translation for a small MT system. This approach is based on the correlation between the structure patterns of the source sentence and its corresponding target sentence, so that the later becomes known at the outset of thetranslation process. Since 1989, various corpus based translation approaches are becoming attractive, because they allow the computer to learn from the solidly proved translations. Conversion approach also tells the computer what the right translation should be, only in the form of a formalized expression, which is easier for the computer to realize.

This paper presents how conversion formulae are obtained and how they are used to perform translation. This paper also presents the problems that have to be further studied and the author's opinion about those problems.

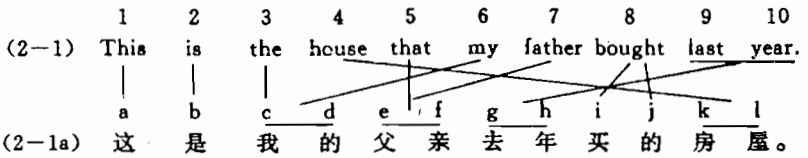
1 导　言

机器翻译已经有 45 年的历史。与它差不多同龄的电视机，复印机，录相机都已经相当成熟，也普及到许多家庭。相比之下，机器翻译的进展显得太慢，更谈不到普及。愈是不普及，关心它的人愈少，必然进步也就更慢。

为了普及机器翻译，笔者力求找到最简单，最直截了当的翻译方法，以便在很便宜，但内存很小的微机上运行。1984 年笔者提出代换法，并以它建立 TECM 英汉翻译系统[1]。到现在为止，这方法还很不完善。但是它与当前正在兴起的基于语料的翻译方法有很多相通之处。可见它还有发展的潜力。因此介绍给读者：一则是抛砖引玉，同时也是向读者求教。

2 翻译的任务

翻译的任务是把源语句 S 所要表达的意思用目标语句 T 说出来，使它在 T 的生活，习俗，文化环境下产生 S 在 S 的环境中所产生的同样效果。这时我们说 T 是 S 的对应句。但是这并不意味 T 与 S 好像投影那样字字对应。大多数情况下，它们的字数并不相等。只有把这些字适当的分组，成为一些单词和短语才能对应起来，而且对应词的位置不见得相同，还有一部份词找不到相对应的词。例如



可以看出：除了 1,2 同 a,b 依次对应之外，其他对应关系混乱，而且 3,5 和 j 找不到对应词。这例子说明了翻译的任务是如下：

(1) 把源句的字适当的切分组合形成适当的短语和单词。

(2) 找出这些短语和单词的适当对应词。

(3) 把这些对应词按适当的顺序排列形成目标句。

前两项任务没有明显的困难。或者说，人们照他自己的想法胡乱做了，也看不出有什么不对。第三项问题就大了。怎样才是适当的排列？照源句排列显然不行。按什么规则排列？语言学家当然有一大套规则，但是计算机如何执行？

3 句型和转换式

所谓句型就是句子内短语和单词的排列顺序。如果我们知道目标句的句型，就等于知道了组成目标句的短语和单词应该如何排列，即是解决了上述第三项问题。

为了使不同句型的数量不致多到令人眼花潦乱，组成句型表达式的变量不宜过多。假设用词类作变量就太复杂了，反而看不出问题。因此笔者选择四个变量，即动词 V，非动词 S，连词 K 和关系词 L。这样句 (2−1) 的句型可以表示为 $S_{1,1}V_{2,2}S_{3,4}L_{5,5}S_{6,7}V_{8,8}S_{9,10}$。每个变量的下标表示第几个字到第几个字。

用传统语法的话来说：my father bought last year 是限定性定性从句，它形容 the house。所以按照中文语法的要求，应该放在 the house 之前。这话可以写成：

(3−1) $S_1V_1S_2LS_3V_2S_4 \rightarrow S_1V_1S_3V_2S_4$ 的 S_2

(3−1)叫做句型转换式。它实际上就是一条语法规则，不过写成表达式比用文字更明确具体，计算机容易执行。附带说明一点：定语从句放在被定的名词短语之前，当然是整个从句前移，而在 (2−1a) 的译文中，把“去年”和“买”掉换了位置。那是另外一个问题，下文还要提到。在 (3−1) 中表明从句 $S_3V_2S_4$ 是保持原来顺序不变。

现在我们看到：(2−1)是一个例句。由于这一例句的提醒，使我们想到：限定性定语从句与被定名词短语的前后顺序，英文与中文不同，这是一个普遍性问题，所以应该把它形式化，变成(3−1)那样的表达式，成为计算机可执行的规则。这就是代换法的核心，也是它与基于语料的翻译方法相通的地方。

4 句型转换表

代换法与基于语料库的翻译方法不同之点在于代换法下用语料库而用一张转换表。转换表比语料库小得多，而且容易得到，容易维护，对小系统很适合。转换表并不包括所有可能出现的句型，它只列入需要变换的句型。在英译汉里大不份译文句型是与源文相同的，只有小部份需要变换。这部份在全文中所占的比重轻重不一。据笔者的统计，技术性文章中需要变换的句数占全文句数的 38.9%到 66.7%。而文学作品则从 2.2%到 53.5%。

1984 年 TECM 系统的句型转换表只有 35 种句型。到现在，由于翻译了不同类型的源文，句型转换表已经增加到 109 种句型。将来可能略有增加，但不会很多。

转换式(3−1)不是无条件的。$S_{3,4}$the house，即是一个名词短语，名词前面可以有若干个副词和若干个形容词，但不能有介词，更不能有其他名词。如果有，(3−1)就不能用。所以句型转换表应该有三栏如表 (4−1)：

表 4−1　　句型转换表

源句型	如果符合条件	目标句型
$S_1V_1S_2LS_3V_2S_4$	S_2 = 一个名词短语 L = that,which	$S_1V_1S_3V_2S_4$ 的 S_2

如果不符合上述条件，就不能执行转换，即保持原来的句型。

TECM 系统的句型转换表实际上不是列成表，而是一套程序。这程序约占 12K 字节，比语料库小得多了。

5 短语转换

从句(2−1)的例子可以看出：源句与目标句字序的差异有两个来源：(1)句型差异，例如句(2−1)中限定性定语从句的位置；(2)短语的差异，例如句(2−1)中限定性定语从句中的时间副词。英语时间副词往往是在一句之末，而中文则必需放在有关动词之前。所以做完句型转换之后，还要做一次短语转换。例如句(2−1) 中的 bought last year 需要变成“去年买”。

一个短语可能是由几个较小的短语组成。尤其是句型表达式中的 S 很可能包括几个名词短语，介词短语，分词短语等等。这也是笔者为什么不用 N 而用 S 代表它的原因。这些短语是否可以看成一个结构，而用一个转换式来表示源短语与目标短语的结构关系？比较单纯的短语是可以的。然而有许多时候，如果短语中包含各种短语情况复杂，就不能用一个转换式来说明一切了。所谓比较单纯是指嵌套层次不大于 1。举一个例来说明如下：

短语 $N_1$of N_2 可以有条件的适用转换式

(5−1)　　$N_1$of N_2→N_2 的 N_1

条件是 N_1，N_2 都是简单名词短语，而且 N_1 不是量词，容器词或种类形式如 sort，type 之类的字。

TECM 系统大约有 4000 条系 (5−1) 那样的短语转换式，它们相当于基于语料库方

法中的“语料碎片”，不过，它们是形式化的，条件明确，更便于计算机执行。

假设有一个短语是 N_1in $N_2$of $N_3$for N_4。显然不可以把它看成 (N_1in N_2) of ($N_3$for N_4) 然后运用(5−1)来转换。三个或更多介词短语接连出现就构成含糊结构，它的意义可以作多种解释，而无法规定某一种是正确或不正确。例如

(5−2) they will request some sample pipes with defects for use in the future 句中 for use in the future 是 sample pipes 的定语，转换式为

(5−3) $S_1$with $S_2$for S_3in S_4↔ for S_3in S_4 的 $S_1$with S_2

译为“他们将申请一些为将来使用的带缺陷的样品管”。但同样结构的另一句

(5−4) they will request some sample pipes with numbers for identification in the future.

句中 for identification in the future 是 numbers 的定语，转换式为

(5−5) $S_1$with $S_2$for S_3in S_4→ $S_1$with for S_3in S_4 的 S_2

译为“他们将申请一些样品管带有便于将来辨认的号码”。同一结构有两个转换式，什么时候该用(5−3)，什么情况下该用(5−5)很难说得清。从翻译角度而言，倒不如以含糊对含糊，干脆不转换，把(5−2)译成

(5−2a) 他们将申请一些带缺陷的样品管，供将来之用。

把(5−4)译成

(5−4a) 他们将申请一些带号码的样品管，以便将来辨认。

即使这样也要把同一个 for S in the future 译成两个样子，如何可能？这只好依赖译后编辑了。

总而言之，短语内部的字序调整是一个很复杂的问题，现在还没有完全解决，本文也不打算多讨论了。结论是：译后编辑不可避免。

6 正确切分

前面§2 把翻译任务分解为三项，上文讨论了第三项。本节将简单的讨论第一项。仍然以(5−2)和(5−4)两句为例。前节问到什么情况下该用转换式(5−3)，什么情况下该用(5−5)？其实这问题不难回答，关键在短语 some sample pipes with $S_1$for S_2in the future 如何切分。可以有两种切分法：

(6−1) some sample pipes with $S_1$for S_2in the future;

(6−2) some sample pipes with $S_1$for S_2in the future。

照(6−1)切分时适用转换式(5−3)，照(6−2)切分时适用(5−5)可惜这回答并没有解决计算机该如何下手的问题。要解决这问题，最简单的办法就是在切分线的地方加一个“,”，即是说需要译前编辑。但是这样细微的问题要求编辑人员在翻译之前就预见到，未免有些过份。因此 TECM 系统设置一种亡羊补牢的办法：先译(5−2)译出来一看，计算机切分不对，按 F8 键，(5−2)句重新显示出来，编辑人员在适当地方加“,”，再按 F1 键，计算机重新翻译，果然好了。当然这只是没有办法的办法，仍属于待解决的问题。

7 多义词问题

关于§2翻译任务的第二项不难解决。一本字典就够了。可惜字典上往往是一字多义。关于岐义的分辨我国的李维等同志[2]已经提出很多方法，这里只说另外两个问题：(1)英文中不同的字有细微差别而对应的中文无法表现这差别；(2)英文中一个字对应中文几个相近但有细微差别的意思，人不难作出选择，而计算机很难选择。

上述第一类问题不很严重。例如英文 blossom 和 flower 中文都译作花。但 blossom 是指能结出果实的花，所以只说花不够准确，但译成能结果的花又太罗嗦。

第二类问题比较麻烦，以 travel 一字为例，它对应中文的旅行，行进，传播，运送等等，试看下列句子：

(7−1) he travels all over the world.他周游全世界

(7−2) train travels faster than ship.火车比轮船跑得快

(7−3) sound can not travel through vacuum.声音不能透过真空而传播

(7−4) this fish travels very badly. 这鱼经不起运输

计算机如何选择恰当的解释？人工智能学家不厌其烦的在语义分析上很下功夫，希望将来能够解决。就目前而论，为了简捷了当，TECM 系统率性把这事留给译后编辑者。考虑到编辑时删除比插入方便很多，所以宁愿让候选的解释过多，而避免把有用的解释遗漏。但是这就造成译后编辑的修改量增大。

8 相似原理和语言差异

前面§2把一个例句普遍化，写成一个句型转换式，即是说，代换法认为结构相同的源句应该对应结构相同的目标句。但是不等于说应该对应字序相同的目标句。代换法也有可能把结构相同的源句译成结构各异的目标句。因为转换式是有条件的，符合某些条件就转换成某种目标句．符合另一些条件就转换成另一种目标句，还有可能每一组条件都不能完全满足，而不进行转换，保持原来句型。

此外，即使两句都转换成同一目标句型，但是经过短语转换之后，这两句的字序也不一样，请看下列例句：

(8−1) I put the book in the drawer.我把书放在抽屉里

(8−2) you drink the tea in the cup.你饮杯中的茶

(8−3) he saw the cook in the room. 他在房间里看见厨师

(8−4) she lives in Vinice.她居住在 Vinice.

这些源句的句型完全一样；但译文都各不相同。有的介词短语在动词之前，有的却在后。这是动词的差异造成的。

我们在承认相似原理的同时，也承认不同的语言有字序上的差异。代换法包括了一些处理字序的规则，但还很不够，不能保证目标句的字序完全正确。

9 翻译速度和质量

TECM 系统的程序总共不到 40K 字节，每翻译一句英文只用 0.25 秒，即每小时可译 14,400 句，大约 288,000 字。但是因为译文质量不高，译后编辑需要修改 30%即大约 86,400 字。改这八万多字可能需要数百小时。因为首先要读懂 28 万多字的源文，想清楚

该怎样译，再看约 30 万字的译文，看清楚有何不妥之处，如何改才好。这等于读 50 万字的书加做笔记的时间。

为了改变这种局面，最根本的途径就是提高翻译质量。如前所述，翻译质量的关键是两点：(1)多义词选择恰当的解释；(2)译文字序符合目标语的习惯。前者困难很多，而后者比较容易见效。凡是源语言中不能逐字翻译的短语都用一个转换式来转换。例如

(9−1)　available to S→S 弄得到

即使不很确切，也比不转换好。

目前 TECM 系统没有放手这样做，因为还有一个使用频率的问题。韦氏大字典上约有 16 万英文字，而一个英国人一辈子真正用到字数不超过三万。R.E.Jones [3] 等人曾经对 10,289 篇摘要，共 446,097 字作了单字出现频率的统计。证明只用到 23,505 个单字，其中有 12,485 个字只出现一次。可见真正常用者只有 11,000 字。为了确保用户的每一分钱都花在有用的地方，TECM 系统的字典只收入在翻译中遇见过的字。翻译第一本书时只有 7,000 多字，现在增加到 14,000 字。对于短语就更加谨慎，要确实常见的才收入。例如上述的(9−1)就没有收。什么叫常见？它是受下列考虑的影响。

TECM 这样的小系统对多义词的分辨能力不可能高明，同形词靠词类来分辨还勉强能做到。如果几个词义都属于同一词类，就只能并列出来让读者自选。为了减少这种现象，唯一的办法就是限制翻译的领域，即针对某专业，甚至某专业中的某用户。笔者曾经试验针对航天料技资料目录来修订字典和短语词典。结果做到翻译 4141 字，译后编辑仅修改 75 字。修改率为 1.82%。题录中动词少，句型结构比较简单，所以比较容易做好。今年试验针对石油管道的技术标准文件为翻译领域。估计在针对范围以外的短语暂不收入词典。希望今年底能得出结果。

10　结束语

代换法可以说是基本句型和短语的翻译方法。它与基于语料的翻译方法有相似之处，只是它不需要语料库而代之以转换式。这种方法的优点是简单，因而迅速，每翻译一句只用 0.25 秒。目前由于多义词词义的选择和译文的字序问题还存在许多问题，所以译文质量不高，需要译后编辑，修改约 35%。笔者认为如果能针对用户的翻译范围来编制字典和短语词典，效果会明显改善，成为廉价而快速的翻译系统。

参考文献

[1]Liu Shiao-Shu, TECM--Translation of English into Chinese on a Micro,in Progress in Machine Translation, ed. Kelly, Sigma Press, 1989.

[2] 李维，刘倬，机器翻译词也辨识对策，中文信息学报，VoL4.No.1,1990

[3] Paul E. Jones, Vincent E. Giuliano, Robert M. Curtice, Automatic Language Processing, 1969.