机器翻译中的语用自动调序[¯]

赵会军

（河北经贸大学 石家庄 050061）

摘要：语用学是研究语言的使用的一门科学，统计机器翻译方法通过词语在语境中的使用情况确定歧义词语的具体含义，这也是语用方法在机器翻译中的自然应用。语用的方法还能够通过分析源语言单元的关联关系来确定源语言的逻辑顺序关系，从而确定目标语言的逻辑顺序关系，这就使得源语言结构和目标语言结构具有可计算性。基于这种可计算性的语用自动调序使得像英汉这样差别巨大的语言之间的自动翻译达到基本可读性目标成为可能。

关键词：机器翻译；语用；调序；可读性

中图分类号：TP391　　**文献标识码：**A

Pragmatic Reordering in Machine Translation

Zhao Huijun

(Hebei University of Economics and Business, Shijiazhuang 050061, China)

Abstract: Pragmatics is a scientific study on the use of languages. In Statistical Machine Translation, according to the use of the ambiguous words and expressions in the context, the exact meaning could be defined by statistical method and this is an unconscious appliance of the pragmatics. In way of pragmatics, the logic relations of the source language could also be defined by analyzing the relevant relations of the source elements, and hence the logic sequence of the target language could be got, the computability between the source and the target languages could be realized. Based on this kind of computability, the pragmatic automatic reordering makes it possible for the readability of the target language in MT between the vastly different languages like English and Chinese.

Key words: Machine Translation; Pragmatics; Reordering; Readability

语用学机器翻译调序系统是基于语用学基本理念构建的。不同于基于规则的机器翻译方法，语用调序系统是在现有统计机器翻译的基础上运用语用学的理念编制规则，正如统计机器翻译不是利用我们常规的语法规则来处理语言结构一样，语用的方法也不是我们所熟知的常规语法，在这一点上与统计机器翻译中运用的一些方法具有一定的相似之处，这也是语用方法能较好地融入到统计机器翻译中的基础。语用方法的根本指导思想是根据人的思维来判断语言结构并根据人的认知来处理语言结构，让机器能够最大限度地模拟人的思维。通俗一点讲就是机器通过语用方法的处理模式能像人一样知道单词和句子的真正意思，而不单单是字符的识别。本文将首先介绍机器翻译中的语用概念，然后论述基于语用方法的语用调序基本原理，与现有统计机器翻译系统的结果对比分析，最后概括实施过程中的一些问题和可能达到的目标。

1 机器翻译中的语用概念

为了让读者更清楚地理解语用学的概念及其理论，在语用学教科书中所举的例子大多数是具有暗示隐含意义或多义模糊的话语，只有通过对前后文或自然环境的综合理解才能知道

[¯]**收稿日期：**　　**定稿日期：**

作者简介：赵会军（1968—），男，教授，主要研究方向为机器翻译、语用翻译。

话语的具体含义，在此不再举例赘述。语用学的核心思想其实可以归结为两个字：语境。语用体现为自然语言的应用。在机器翻译的应用上可以包括前后词语的关联、前后语句的关联、语言文化的差异与关联，语言逻辑的差异与关联等。通过语言本身使用的具体情况来分析源语言，按照目标语言使用规范来约束目标语。

比如我们说出一句话：“我是一个学物理的学生。”机器怎么知道“我”是一个“人”，而不仅仅是一个单纯的字呢？根据语用的方法，机器首先要有“我”的知识：“我是”，“我的”，“我们”，“打我”，“我高兴”，“送我”，等无数关联的词语。根据语用的观点，词在具体语境中的意义是可以相对明确的。这些关联的词语也正是统计机器翻译从语料库所抽取的基本短语对的一部分。以谷歌翻译为例，在词语意义的辨别上差错率不高，原因就在于其基于强大搜索引擎的庞大的双语语料库所抽取的海量词语对，正是这些海量词语对使某个词有了相对确定的含义。即使如此，依然还是有些多义词无法分辨清楚。

然而多数时候谷歌翻译的结果仍然让人感觉一头雾水。原因何在呢？因为谷歌翻译虽然得到了“我”的相对确定的含义，对于“我”有了一定的知识，但并没有进一步拓展知识的适用范围，换句话通俗的话讲就是不知道“我”该如何使用。而语用学正是研究语言是如何使用的一门科学。语言的使用主要体现在句子的层面上。罗列单个词语无法确定该词语将如何使用。只有把词语放到句子中才能确定某个词语的基本意思（歧义的句子还需要放到段落和篇章中予以确认具体含义）。基于常规语法规则的机器翻译研究基本上与语用方法相反，是根据人们总结出来的语法规则来行事，目前的事实证明这种语法规则方法对机器翻译效果虽然有一定的帮助，但在统计机器翻译大踏步向前进的时候却在很大程度上起到了反作用。那么障碍来自何方呢？我们知道，机器翻译需要编制程序来实现自动处理，而根据语法规则编制的程序无法满足千变万化的语言结构变化，往往一个规则却成了另一个规则的掣肘。

西方语言之间（如英法和英德）由于其语言基本结构、基本单元和表达方式相差相对不大，因此机器翻译不考虑语法关系或只考虑简单的语法变化就可以在一定程度上实现一定的可读性，尽管如此，西方语言之间的机器翻译仍然存在很多问题。而英汉之间的差异不仅仅体现在词语形状，更主要的是语言结构方面的。尽管英汉语都有主谓宾定状补，尽管也都可以用乔姆斯基的语法生成规则来分析语言结构，但却无法实现两种语言的千变万化的结构对接。

现有统计机器翻译系统是建立在西方语言之间的翻译理论之上的算法体系，尽管也适用于英汉语言，但并没有一套成熟的理论做指导从根本上来解决这种语言结构差异问题。正如刘群所说：树到树模型，这一类模型试图在源语言和目标语言两方面同时引入语言学意义上的句法结构。这一类模型目前还没有比较成功的尝试[1]。从首个统计机器翻译系统面世到现在，虽然翻译效果不断提高，但仍没有根本性的改变。近年来的成果主要体现在算法上。即使某个算法提高了几个百分点，但与另一个算法叠加的话可能反而会降低几个百分点。纵观语言的纷繁结构和万千变化，单从计算机算法上就某个结构或从某一个角度来提高树对树的准确率的方法似乎对整体翻译效果很难有重大改观。但要把所有的语言结构和变化都融入到一起进行算法融合似乎也是无法做到的[1]。

交际是否成功，就看交际双方对彼此的认知环境是否能显映（manifest）和互相显映（mutually manifest）。关联论把关联定义为“命题和一系列语境之间的关系”，因此关联是依赖语境的[2]。运用语用方法通过对语言使用的原始结构进行分析，使两种语言结构直接对接。对接方式主要通过对源语言的编码和对目标语言的解码方式解决。编码和解码分别依据源语言和目标语言的基本意义单元的使用情况来实现。

2 语用调序基本原理

标志机器翻译成熟的第一个阶段是：实现翻译的可读性目标，也就是读的通顺、读的懂、意思基本准确，可以满足一般性使用，这个阶段目前英汉翻译之间还差的很多。第二个阶段是精准和富有感情色彩阶段，这个阶段只有在第一个阶段实现的基础上才能逐渐实现。

当前，目标语言的调序问题是制约英汉翻译达到第一个阶段的核心问题。代表机器翻译最新发展技术的谷歌、百度、必应等翻译系统在词语的对应上还是相对比较准确的，但在整体句子结构方面还没有达到令人满意的程度。从下例我们可以分析找出问题所在（原句和翻译结果的提取日期为 2013-9-23）。

(1) A US man is $1 million richer after finding a winning lottery ticket inside his glove compartment two months after he first purchased it. (http://nz.news.yahoo.com/a/-/top-stories/19028235/man-finds-1-million-lottery-ticket-in-car-s-glove-compartment/)

谷歌：一名美国男子为 100 万元，找到中奖彩票，他的手套箱内，两个月后，他第一次购买后，更丰富。

百度：一名美国男子赚进 1000000 美元后发现一张中奖的彩票他的手套车厢内的两个月后，他第一次买它。

必应：一个美国人是 $100 万更丰富后发现中奖彩票在他的车厢内两个月后他第一次购买了它。

有道：美国人是富裕 100 万美元的彩票后发现他的手套隔间内两个月后他第一次购买它。

从例句(1)的 4 个译文我们看到，每一个词语的意思基本正确，但整句意思并不通顺，汉语词语基本按照原英文的顺序排列，没有就原英文的语言逻辑按照汉语的语言逻辑顺序进行正确调整，导致汉语语序混乱，意思不明确。赵会军从中英文不同的语言逻辑结构特点出发，把句子结构从空间结构逻辑重组、时间顺序逻辑重组和时空共构逻辑重组三个方面进行划分[3]，较好地说明了英汉语语序不同的问题。下面例句(2) [3]可以帮助理解其中的要点：

(2) You may be stopped-out well before the final top: but that ①is much better than ②watching a massive profit wiped-out by ③staying with the move to the top, ④and then staying with the following collapse.

译文：你可能在最终顶部之前就卖出了：但这比（你的资本）③随着价格移动到顶部，④然后再随着价格崩溃，②眼看着大量的利润被吞噬掉①要好的多。

例句(2)中英文标记数码① ② ③④按照英文顺序编排，而汉语中的③④②①是按照与原英文的对应的汉语的顺序排列。这样的汉语是符合汉语的表达习惯的，意思通顺，逻辑正确。而如果还按照原英文顺序排列则是无法理解的。在人类的翻译实践中，毫无疑问这种语序变化可以做到的。但真正地把这个理论应用于机器自动翻译上还需要程序上能行得通的具体方法做指导。

显然，例句(2)中按照常规的语法规则编制计算机程序算法也是无济于事的，因为语言结构千变万化，人类不可能穷尽句子结构，同样计算机也不能。而语用学的观点可以引导我们将无穷的语言组合变化缩小到一定的可操作范围。何自然把语用翻译从关联理论中得到的一些启示归纳为：1）要翻译，首先要理解原文；2）寻找关联，要靠译者的百科知识、原文语言提供的逻辑信息和词汇信息、原文的文化背景信息等一些对理解原文有用的信息；3）由于原文作者和译者的认知环境不同，作者力图实现的语境效果同译者从原文和语境中寻找关联而获得的语境效果毕竟是两回事[4]。关联论通过明示—推理提供译文最佳的语境效果[5]。正如前面所述，统计机器翻译通过单词和短语的前后关联组合可以将单词和短语相对准确地对应出来。进一步来讲，我们也可以通过分析原语言句子的每个细小单元之间的关联关系将句子重新组合成目标语言的关联关系，并且这种关联关系是可以计算的。

具体到例句(2)我们要抓住时间逻辑顺序这个关键因素，找出时间关联标记，其中，“by”和“then”表示时间前后关系，但英文中这种时间顺序关系与汉语是不同的，英文语序并没有像汉语那样按时间前后顺序排列事件，在自动翻译时可以把表时间的“by”部分的内容做结构调整。

在语言实践中，我们会发现英文语句和汉语语句结构大多不一致，甚至完全相反。最明显的例子就是地址表达上的差异---英文从小到大排列，而汉语从大到小排列。如果如写信那样把英文地址按行表述，那自动翻译时汉语很容易区分出来每个地址单元，只是在翻译时

整个完全反转即可。而如果写成不分行连续形式就要难的多。但我们也可以根据原文的关联标记来自动转换。如：Room 888, 9th Floor, Arts 3 Building, 14 A, Symonds St, Auckland 1142, NZ. 我们把“,”作为关联标记，则可以把汉语按反顺序排列即可。像地址这样的语言是没有语法成分可分析的，因此显然超出了按照NP\VP这种方式来划分语言结构的处理范围。因此需要按照语用学所认为的---只要是现实当中使用的语言就是正确的---这个原则来行事，而不要去管是什么样的语法规则。

总之，我们可以通过原文语言提供的逻辑信息和词汇信息、文化背景信息等，找到原文的关联标记，确定原文的关联关系，按照目标语的关联关系进行重新排列，翻译成符合目标语使用习惯的译文。

3 语用自动调序框架构建

统计机器翻译系统生成的一般过程是：双语语料库→词语标注、短语表→翻译模型（目前主要是利用 NP/VP 结构来进行词语对齐和调序）→目标语生成。由于在双语语料库和词性标注以及短语表的抽取等方面统计机器翻译已经取得了很好的效果，语用调序系统重点关注的是目标语的调序。下面从语料库、词性标注、短语表、调序方法和调序过程等简要说明语用自动调序的框架构建。

语料库、词性标注和短语表 我们自己建立了基本单词表、短语表和各种语言库（如成语、动词、名词、连词、介词、副词、形容词等 20 多个子库），同时主要利用 NLTK 语言包的语料库、词性自动标注系统（我们自己建立了各种特殊词性标记库等）、和由第三方提供的短语表。

调序方法和过程 调序方法不同于一般统计机器翻译所利用基于乔姆斯基的串和树结构模型，而是以线性自然切分的方式。从理论上讲似乎串树结构模型是基于人为制定的语法规则上的，对于千变万化的自然语言和不符合语法规则的语言结构可能不具有普适性，这可能也是目前无论采用多先进和多复杂的算法也还没有找到长距离调序的好的办法、无法达到一般可读性目标的一个原因。而采用线性自然切分语言结构似乎更符合语言使用的自然规律，这种方法并不是完全抛弃现有统计机器翻译的一些短距离调序的巨大成果，而是要充分利用之（由于条件所限，目前还没有衔接好）。

调序过程包括源语言预调序和目标语言规范整理两大步骤。源语言预调序首先对篇章进行句子切分，以自然语言的句号（包括感叹号、换行等）意义单位为一个句子单元。然后对每个句子单元进行扫描并结合各个语言子库抽取语用关联标记信息（如时间关联标记信息等，这些标记信息可能是标点符号、名词、动词、介词等各种词性，也可能是一个很长的短语组），以这些标记信息为节点将句子切割成数个小句，每个小句的位置将由关联标记信息来决定并进行顺序调整。进行词语对齐后，再对得到的结果按照目标语言使用规范进行整理，最后得到目标语翻译结果。整个调序过程如图 1 所示。

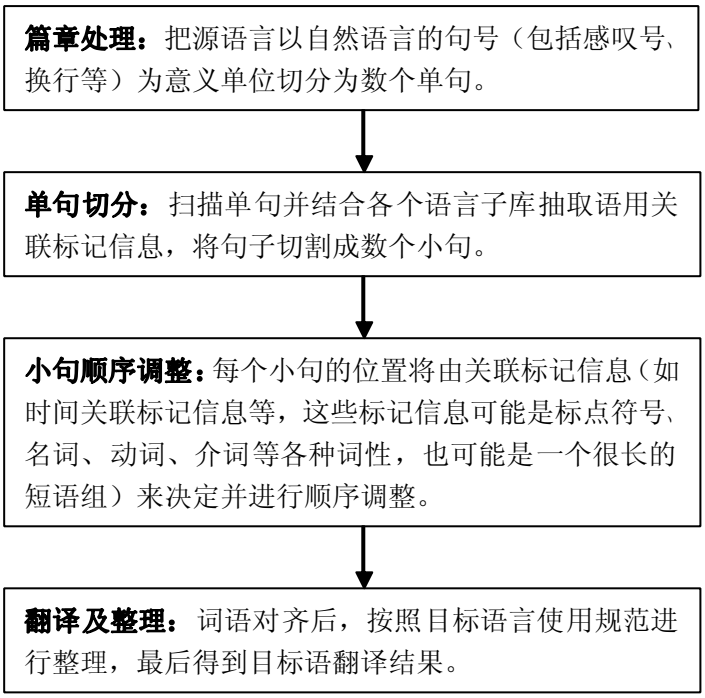

图 1 语用调序过程

结合原文从各个语言子库抽取语用关联标记信息涉及到几十个不同的算法实现，由于篇幅所限，在此不再一一介绍。另外，如何将语用调序系统与统计机器翻译的各个阶段性成果相结合将另文探讨，但基本过程如图 2 所示。

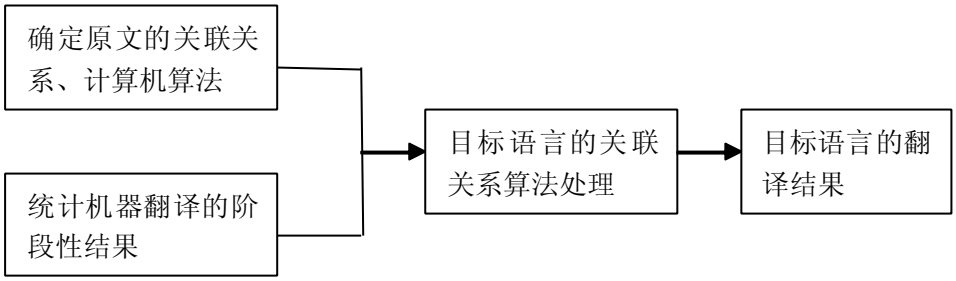

图 2 语用调序与统计机器翻译结合的基本过程

4 语用自动调序实验结果分析

根据语用调序的基本原理和调序框架，经过一段时间的程序设计，我们制作了一个演示性产品来验证理论的可行性。目前机器翻译自动评估方法，如基于 N 元匹配的 BLEU 和 NIST 自动评测方法、基于准确率和召回率的 GTM 评测方法以及若干其他自动评测方法等，主要是基于词对齐的评估方案，而针对词语和语句顺序的合理性和正确性考虑的并不充分，因此语用自动调序结果只能通过人工的方式来评测。我们对近期雅虎网站上的大约 5000 句英文语句的翻译结果进行了人工对比。

下面通过两个例子的几个系统翻译结果的对比（原句和翻译结果的提取日期为2013-9-23，语用自动调序的翻译结果未经过人为加工），分析语用自动调序的优势和存在的问题以及解决方案。

(3) An estimated 34,000 people, or about one in every 120 New Zealanders, were unable to access housing in 2006, according to the latest available census and emergency housing data, say University of Otago, Wellington (UOW) researchers.

（http://nz.news.yahoo.com/a/-/top-stories/19058396/otago-research-shows-34-000-people-missing-out-on-housing/）

谷歌：估计有 34,000 人，大约每 120 个新西兰人之一，在 2006 年无法获得住房，根据最新普查和紧急住房数据，说，惠灵顿奥塔哥大学（UOW）的研究人员。

百度：估计有 34000 人，大约每 120 个新西兰人，无法获得房屋的 2006，根据最新的人口普查和紧急住房数据，说，奥塔哥大学，惠灵顿（UOW）的研究。

必应：估计的 3.4 万，或大约一个在每个 120 新西兰人，到了无法访问在 2006 年，根据最新可用人口普查和紧急住房数据，房屋说奥塔哥大学惠灵顿 (UOW) 的研究人员。

有道：估计有 34000 人,约每 120 新西兰人,无法进入房地产在 2006 年,根据最新人口普查和紧急住房数据可用,说奥塔戈大学的惠灵顿(UOW)研究人员。

语用自动调序：惠灵顿的奥塔哥大学（UOW）的研究人员说，在每 120 名新西兰人约一个，根据最新的人口普查和应急住房数据，2006 年据估计有 34000 人无法获得房屋。

例(3)中，通过分析可以看出，A. 从词语的对应上，谷歌和必应能够把所有的原语言词语都反映到目标语中，百度和有道则缺失了“one”（之一）；B. 从词语的准确性上，基本都可以准确地表到原文的词语含义；C. 从目标语顺序上，必应、百度和有道基本按照原文顺序排列，谷歌则在有的地方进行了微调，效果也好些，而语用自动调序则进行了大幅调整，效果最好；D. 从目标语整体意思表达上，只有语用自动调序符合汉语习惯，逻辑清晰，有较好的可读性。

(4) Mr Barnes had evaded police on the same motorbike in Hawkes Bay during a pursuit that reached speeds of more than 200km/h before the chase was abandoned. （http://www.nzherald.co.nz/accidents/news/article.cfm?c_id=13&objectid=11127188）

谷歌：巴恩斯先生逃避警方在一个追求达到超过 200 公里每小时的速度追前被遗弃在同一摩托车在霍克斯湾。

百度：巴尼斯先生被警察在霍克斯湾一摩托车追求达到超过 200 公里/小时的速度在追逐期间放弃。

必应：先生巴恩斯逃避了警察霍克斯湾同一机车上期间达到小时 200 公里以上的速度在追逐被摒弃了之前的追求。

有道：巴恩斯先生逃离警察在相同的摩托车在霍克斯湾,在追求速度达到 200 公里/小时以上在追逐被放弃了。

语用自动调序：一个追求期间，巴尼斯先生回避在霍克斯湾该摩托车的警察，在大通放弃了之前，达到超过 200km/h 速度。

例(4)中，通过分析可以看出，A. 从词语的对应上，谷歌和必应能够把所有的原语言词语都反映到目标语中，百度则缺失了“evaded”这个关键词语，有道缺失了“during”；B. 从词语的准确性上，语用自动调序把“chase”译成了“大通”，这是由于无法层次提取短语表的内容所致，这个问题是可以通过改善基础条件加以解决的；C. 从目标语顺序上，必应基本按照原文顺序排列，谷歌、百度和有道在有的地方进行了微调，语用自动调序则进行了大幅调整，只有“该摩托车”的位置没有调整过来，这是由于程序调试不到位所致，是可以解决的；D. 从目标语整体意思表达上，只有语用自动调序比较符合汉语习惯，逻辑比较清晰，有一定的可读性。

综合例(3)和例(4)的翻译结果可以看出，语用自动调序可以在相当程度上解决目前统计机器翻译中一直无法很好解决的语言调序问题，这也是机器翻译向前发展的一个瓶颈问题。

通过每天都在进行的大量人工对比，我们发现语用自动调序：A. 目前能够较好地进行词语的对应，丢失原文词语的现象很少；B. 多义词语有时会出现不准确的现象，但如果能够进行层次短语提取，则可以很大程度上解决这个问题；C. 可以更符合目标语的表达习惯，尽管有的地方还需要调整，但整体有一定的可读性。D. 总的来讲，大约有 30%的翻译结果明显优于市面上最好的自动翻译结果，另外约有 50-60%不低于市面上最好的自动翻译结果，

只有约有 10-20%不如市面上最好的自动翻译结果，但是都可以找到解决的办法，主要是由于各个基础数据库不够完善和程序调试不到位所致，并非程序框架和理论上的问题。

5 不足与任务

语用的方法通过分析语言的使用来指导机器翻译的翻译过程，由于条件所限，语用调序系统还存在很多不足：目前各种基础语言库的内容还需要完善；抽取语用关联标记信息的方法和技术仍需提高；人工评测还需要设定具体方案，并需要有权威的专家评测组来评测；由于是第三方短语表，实验语料只能基于单层次短语提取，无法实现多层次短语的调取；程序调试不尽如意等。总之还要不断完善和深入研究。

从语用自动调序试验结果可以看出，把语言学因素加入到统计机器翻译系统中是可行的，对于实现英汉语言的完全自动翻译具有重大的指导意义和实践价值。如果有更好的基础条件的支持，有可能在短时间内完成自动翻译的可读性目标，而不是人们通常所认为的二十年甚至三十年。

参考文献

[1] 刘群. 机器翻译研究新进展 [J]. 当代语言学, 第 11 卷. 2009 年第 2 期: 147- 158.

[2] 何兆熊. 新编语用学概要 [M]. 上海：上海外语教育出版社，2000：181-203.

[3] 赵会军. 金融英语翻译中的逻辑重组 [C] //叶兴国. 第八届全国国际商务英语研讨会论文集《新形势下的商务英语教学与研究》. 上海：上海外语教育出版社，2008：478-485.

[4] 何自然. 语用学与英语学习 [M]. 上海：上海外语教育出版社，1997：193.

[5] 曾文雄. 语用学翻译研究[M]. 北京：武汉大学出版社，2007：57.

作者联系方式：赵会军　河北省石家庄市学府路 47 号 河北经贸大学外国语学院
邮编：050061 手机：13832395550 电子邮箱：golddoctor68@gmail.com