中 文 信 息 学 报
第3卷 第2期 JOURNAL OF CHINESE INFORMATION PROCESSING Vol.3 No.2

自然语言的歧义与机器翻译的对策

俞 士 汶

（北京大学视觉听觉信息处理国家重点实验室）

［摘要］ 歧义是自然语言的特征之一。开发机器翻译系统，不仅要研究一种语言内部的歧义，而且还要研究两种不同语言间的歧义。本文从不同的角度考察了这些歧义现象及其对机器翻译的影响，总结了在机器翻译系统中为了得到恰当的译文所采用的几种办法，并且提出了一种观点或者说方法，它允许译文的歧义同原文的歧义相对应。

一、自然语言的歧义现象

歧义是自然语言的显著特点和普遍现象。开发机器翻译系统，不仅要研究一种语言内部的已经相当棘手的歧义问题，而且还要考察不同语种之间的更为复杂的歧义现象。

1.1 词义所表达的概念的差异

词义和概念既有联系又有区别。在同一种语言内部，词义与概念之间已经不是一一对应的关系[1]，在不同的语言之间，词义同所表达的概念之间的对应关系就更加复杂了。

首先注意到的是多义词的问题。象英语中的名词 spring 表示“春天”，“弹簧”，“泉水”等互不相关的几个概念，可以看作是不同的词，只不过拼法相同而已。这会给译词的选择带来困难，但这是无法避免的，只能将它看成几个不同的词。还要注意一种语言的词与其它语言的词相对应所引起的多义性。例如，英语中的动词 wear 的基本义项是“Carry on one-es person or on some part of it.”因此，当英美人说时，他们不会觉得这些句子中的 wear

He is wearing a hat. （1）

He is wearing a coat. （2）

He is wearing shoes. （3）

的词义有什么差别。但是，当译成汉语时，（2）与（3）中的 wear 应译为“穿”，（1）中的应译为“戴”。当译成日语时，则要更细致地把这3种情况分别译成“かぶろ”、“きる”、“はく”。不难看出，处理这种性质的多义，机译系统的负担是沉重的。

一般认为，概念是思维的单位，它对全人类来讲是共同的[1]。如果能在若干种语言间建立起反映共同概念的语义场，并以这种语义场为指导，编制一部共用的类义词典(thesan-

本文1988年11月7日收到

rus），则会给各种语言的互译带来极大的方便。但是，由于词义具有民族性，用什么样的词来表达语义场中的某个概念，不同的语言又有明显的差异。因此，要建立关于相同概念的的不同语言的词义之间的对应关系。下面的表给出了英语、汉语、日语、俄语中属于“学生场”的有关词的对应关系。

美国	中国				日本		苏联
student	学生	大学生		学生	大学生		студент студентка
		中学生	高中生		高校生		
pupil			初中生		中学生		ученик ученица
		小学生			小学生		

这里只列举了各种语言中最常用的几个单词。一个极其简单的“学生场”中所包含的这些单词的对应关系就如此复杂。不难想象，编一部两种语言参照的百科全书似的类义词典是多么艰巨的任务。

1.2 兼类词所引起的歧义

兼类词的问题在汉语与英语中很突出。例如，

书中还有编辑的错错 （4）

若认为“编辑”是动词，对应的英语是 editing，若认为“编辑”是指人名词，则应该对应 editor，一个常被引用的英语句子是

Time	flies	like	an	arrow	（5）
名词	动词	介词	冠词	名词	
动词	名词	动词			
		形容词			
		副词			

从中取出 3 个合乎语法的词类序列，列在下面：

(a) 名词＋动词＋介词＋冠词＋名词

(b) 名词＋名词＋动词＋冠词＋名词

(c) 动词＋名词＋介词＋冠词＋名词

相应的中文译文为

(a) 时间象箭一样地飞。

(b) Time 苍蝇喜欢箭（Time为苍蝇名字）

(c) 象箭一样地测量苍蝇。

1.3 切分所引起的歧义

关于汉语句子中词的切分问题及其所引起的歧义，已有不少著述，本文不再赘述。日语句子中假名与汉字是连写的，词与词之间不留空格，因此也有切分问题。例如，“大人気”

这个短语由3个汉字组成，既可切分为“大|人気”（中文意思是很受欢迎），也可切分为“大人|気”（中文意思是老成，没有孩子气）。全用假名书写的日语句子因切分产生的歧义更多，例如

くるまではこをはこぶ （6）

这句话，至少有4种不同的切分方法。

英语的词与词是隔开的，但也有切分问题。将 weak point 切分为两个词，译成“弱点”，还能过得去。若将 strong point 切分为两个词译成“强点”，中国人就费解了。最好是将 weak point 与 strong point 都作为切分单位，分别译为“缺点”与“优点”。

1.4 句法结构的歧义

当把一个句子或句子的较大成分分解为较小的成分以描述自然语言句子的结构时，会发现大量的句子有不同的结构，这叫做“同形异构”的歧义[2]。例如

She found a colour plate in that book （7）

中的介词短词 in that book 可能是 found 的状语，也可能是 plate 的定语。又如

He read the letter which she received yesterday. （8）

这一句中的状语 yesterday 可能是修饰 read的，也可能是修饰 received 的。

英语的名词短语内部的不同组合也会引起歧义。英语的名句短语可有如下几种形式

(a) N1 of N_2 of N_3

(b) N1 of N_2 and N_3

(c) N1 and N_2 of N_3

(d) Adj N1 and N_2

(e) N1 of N_2 and N_3 of N_4

可将它们不同的内部结构分别表示为

(a)	N1 of (N_2 of N_3)	(N1 of N_2) of N_3
(b)	N1 of (N_2 and N_3)	(N1 of N_2) and N_3
(c)	N1 and (N_2 of N_3)	(N1 and N_2) of N_3
(d)	Adj (N1 and N_2)	(Adj N1) and N_2
(e)	N1 of (N_2 and N_3) of N_4	(N1 of N_2) and (N_3 of N_4)

类似的问题在日语、汉语中也都存在。

1.5 逻辑上的歧义

有些句子只有一种结构，并且组成这个句子的每个单词的词义又都是确定的，但句子本身仍有歧义，出现了“同构异义”[2]，这也是一种语义上的歧义，为了同单词的多义性相区别，不妨叫做逻辑上的歧义。例如，另一个经常被引用的英语句子是

Every man loves a woman. （9）

不难看出，这句英语可以有两种解释，其一是“所有男人爱同一个女人”；其二是“每个男人爱各自的女人。”

日语中也不乏其例，例如

大学生を対象としこ讲演を行なつた。 （10）

这里的大学生既可以是演讲者所面对的听众，也可以是演讲者谈论的话题

汉语由于形态标志少，造句时“意合法”起作用。因此这类歧义现象更为严重。例如

他刚刚做过外科手术。 (11)

他是大夫？还是病人？两种情况都可以用这句话。

对于这类歧义，如果不对上下文进行分析，不了解有关的背景知识，是没法确定孤立的这么一句话到底的是什么意思。

二、歧义对机器翻译的影响

语言学的研究可分为语法（syntax）、语义（semantics）及语用（pragmatics）三个层次。不同水平的翻译也可分别看成是在这三个层次上进行的。不过，当前机器翻译的适用对象是科技文献，目的在于传达原文的内容。通常以句为单位进行翻译，而不考虑一句话在特定语境中的社会效果。因此，可以认为，当前大多数的机器翻译是立足于语法，向语义的层次发展。在语法与语义的层次上，两种语言之间歧义的对应关系不外乎以下4种情况。

2.1 原文无歧义，译文也无歧义。例如

The Yangtse River is in Asia. (12)

扬子江がアジアにある (13)

长江在亚洲。 (14)

这种情况最简单，无论是人翻译还是机器翻译，都不会碰到什么麻烦。但这种例子极少。

2.2 原文无歧义，译文有歧义

将日语的“高校生”译成中文的“中学生”，将英语的pupil或俄语的ученик都笼统地译成“学生”，笔者认为，中国人都是可以接受的，可以从译文的上下文判断这个“学生”的实际程度。

又如，将日语中含义相左的两句话

私は注射する。 (15)

私は注射してもらう。 (16)

不加区分地译为“我打针”，中国人理解起来也不会有什么困难。

2.3 原文有歧义译文无歧义。

将汉语的“他是学生”译成英语，无论是译作“He is a pupil”还是译作“He is a student”都有犯错误的可能性，因为汉语中“学生”这个词所表达的概念的内涵要比英语的pupil或student所表达的广泛。

将§1中的例句（7）译成“她在那本书中发现了一幅彩色插页”或“她发现了那本书的一幅彩色插页”都难免失之于偏颇。

2.4 原文有歧义，译文与有歧义

将sister译作“姐妹”并不一定贴切，因为中国人习惯上说“她是我姐姐”或“她是我妹妹”，而不说“她是我的姐妹”。但是在不了解原文中sister的确切所指的情况下，将

sister 译作“姐妹”还是比主观随意地译作“姐姐”或“妹妹”要安全些。

将§1中的例句（9）译为“所有男人都爱一个女人”可以说恰到好处，原文有两种的意思，译文也有两种意思，任凭读者去猜测。

假如将下面一句英语

I saw a man swimming on the bridge. (17)

译成中文“我在桥上看见一个游泳的人”，因为译者的脑子里有这样的知识，游泳的场所必须有水，而桥上没没水，所以这么译了。但这样译毕竟加进了译者的主观意见，是否客观地反映了原文的本义，不得而知。如果将 swimming 换成 sleeping，或者将 on the bridge换成 in the lake，那么将这句话译成“我看见一个在桥上睡觉的人”或者“我看见一个在湖面游泳的人”也不能算错吧？两种含义不能兼顾，令人烦恼。但是，如果将例句（17）译成日语

私は橋で泳いでいる人を見た。 (18)

则比较巧妙，英文有两种不同的句型结构，日语有 3 种，完全复盖了英语的两种含义。

从以上分析可知，原文或译文有没有歧义并不特别重要。关键是译文是否正确地传达了原文所包含的全部信息。

三、机译系统中对付歧义的策略

本文并不打算探讨翻译与机器翻译的理论问题，只是考虑为了达到一定的工程目标，为了处理原文中经常出现的歧义问题，机器翻译系统可以采取哪些对策。

3.1 语义分析

早期机器翻译失败的教训告诫人们，仅在词汇转换与句法结构分析的基础上实现机器翻译是不可能成功的，必须进行语义分析。七十年代在自然语言理解方面所取得的成就也鼓励人们在机器翻译中采用语义分析技术。

句法结构的歧义最有可能通过语义分析来消除。例如，§ 2 的例句（17）及下面的两句话

I bought a table with three dollars. (19)

I bought a table with three legs. (20)

都有两种不同的句法结构。但人读这些句子时，并不会感到有歧义。对于这类情况，机译系统可以通过查询知识库（游泳只能在有水的场所进行；美元是用来购物的；桌子是有腿的）来解决两种不同句法结构的取舍问题。

又如，名词短语

珍珠的项链和钻石的戒指

至少有两种不同的组合方式：

（珍珠的项链）和（钻石的戒指）

珍珠的（项链和钻石）的戒指

如果机译系统的知识库中存储了如下知识：项链和戒指都是首饰，珍珠和钻石都是制作首饰的材料，再根据“并列的成分应是同一性质的”语法规则，则会将它正确地翻译成如下的英

语：

a pearl necklace and a diamond ring

从以上应用语义分析的片断，可以推断，语义分析是很有价值的，机器翻译系统应当把语法分析与语义分析很好地结合起来。同时，也应该看到，语义分析有很多困难。第一，机译系统中必须要有一个知识库，但庞大的知识库的开发与管理肯定不是轻而易举的事，应当把知识库控制在一个适当的规模，因此也就不能指望通过查询知识库来解决一切歧义问题。第二即使有了一个足够完备的知识库，机译系统所能解决的大致上也还是局限在这样一个范围，即人在读这些孤立的句子时，能够给以唯一的解释，而不至于认为有歧义[3]。要想解决§1中所列举的，象（7），（8），（9），(10)，(11）那些句子的歧义，则必须进行语境分析。不仅要了解上下文的有关信息，而且要具备有关环境的知识。且不说，目前机译系统有无可能进行如此深入的语义、语境分析，对于一个实用的机译系统，这样做的必要性就值得怀疑。

机器翻译不同于自然语言理解。自然语言理解的中心任务是把有潜在歧义的自然语言输入转换为无歧义的内部表示[4]，可是机器翻译的任务是要得到能正确反映原文内容的译文。理论上当然要求翻译者必须完全透彻理解原文，但在实践上要求每个翻译者在翻译每一篇文章时都达到这么高的境界又是不现实的。科学技术发展迅速，且相互渗透。翻译科技文献时，总会碰到理解不了的内容。这时翻译者只能根据句子的结构和从词典中查到的词义机械地组织译文。这样的译文，普通读者理解不了也没有什么奇怪，但是这个领域的专家来阅读它，凭借他的知识与智慧，却能从翻译者自己也不懂的译文中领悟到原文的内容，吸取知识与信息。另外，原文中若出现“山羊是肉食动物”之类违背常识的句子，不是也要照译吗？总之，在机器翻译中，语义分析也不是一把万能的钥匙。

3.2 保留歧义进行翻译

§2 已经提到，将英语例句（17）译成日语例句（18）是很巧妙的。试想，对于例句（17），如果经过一番语义分析得知，on the bridge 只能修饰 saw 而不能修饰 swimming。在此之后，生成日语句子（18）。但这个日语句子（18）仍是有歧义的。语义分析并没有给译文带来任何实惠。省去语义分析岂不更经济些。

不通过语义分析将例句（15）译成汉语，很可能译成

我看见一个在桥上游泳的人。

这当然令人发笑。但是，如果译成

我看见一个游泳的人，在桥上。

似乎过得去，将例句（7）、(8)、(19)、(20)分别含糊地译成

她发现了一幅彩色插页，那本书。

他看了她收到的信，昨天。

我买了一张桌子，三美元。

我买了一张桌子，三条腿。

在某些应用场合下并非一定不能接受[5]。这样处理的特点是保留了原文的歧义。

词义的选择也可能采用暧昧的办法。如将 wear一律译为“穿戴”。

名词短语的翻译可以采用这种策略。例如，不必通过语义分析以明确名词短语的结构是

(Adj N1) and N2，也不必肯定

peaceful atmosphere and crisis

rapid decision and execusion

的结构是 Adj (N1 and N2)，都可以顺序地译成

和平的气氛与危机

迅速的决策与实行

原文暧昧，译文也暧昧。英美人能懂原文，中国人也能懂译文

笔者将这种“歧义同歧义相对应”的翻译策略应用于日汉翻译，得到了如下一些认识：

1. 日语的不含包孕句的名词短语基本上都可以按原文顺序直译成汉语。

2. 日语与汉语的句子都可以有位于句首的主题（topic），且日语的主题由助词は提示，易识别。日语的主题常可单独译出。例如，

この本は私がすきです。 (21)

译成“这本书我喜爱”，不仅意思对，而且传达了原文强调“この本は”的信息。又如

私は頭が痛い。

译成“我头痛”，就很好。假如对例句（21），（22）作一番语义分析，得出“书是喜爱的对象”及“头是人体的一部分”之类的认识，将（21）、（22）译成“我喜爱这本书”，“我的头痛”，固然也不错，但终究失去了原文的精妙之处。

3. 日语与汉语的句子都可以省略主语。因此，在翻译无主语的日语句子时，既不需要从上下文中将主语找出来，也不需要象译成英语那样将主动句转换成被动句。

4. 汉语句子中时态的形式标记少，中国人也能把握事情的时间关系。将日语译成英语时，时态的分析是一个难题。译成汉语时，没有必要作过分细致的时态分析。

某些情况下，可以绕过语义分析，不等于全盘否定语义分析的必要性。参照下面两句日语，就能明白这一义。

太郎は花子とアメリカに行つた。 (23)

太郎はイギリスとアメリカに行つた。 (24)

在（23）中，太郎和花子之间有并列关系，而在（24）中英国与美国之间才有并列关系。

3.3 限定专业范围

如果机器翻译系统只局限某个专业范围，一词多义的问题，就会显著减少。在机械专业，spring 的译义将是“弹簧”、至少应该优先选择“弹簧”

3.4 优先选择使用频度高的词义、词类

§1的例句（5）中，Time 作为名词，flies作为动词的使用频度高，将这两个词的词类按使用频度确定后，再按照语法规则确定剩下词的词类就容易多了。

3.5 译前编辑与译后编辑

有些机译系统要求对原文作一定的编辑加工，编辑加工的主要目的之一就是消除原文的歧义。译前编辑由只懂原文的人去作，但也对机译系统的要求要清楚。

机译系统的译文难免生硬、牵强，错误也不会少。只要对其出错的规律性有一定了解，

即使不懂或不看原文，也可以对译文进行订正。仍以§2中的（17）为例，如果译文是

我发现一个在桥上游泳的人。

不难判断“在桥上”这3个字安错了位置，移到“发现”之前就合理了。又如，有句译文是：

他在家养病的身体。

这话不合中文习惯，将“的身体”3个字删掉就通顺了。

§4 后记

本文整理了笔者近来学习、考虑机器翻译实现方法的一些心得，或许只能算是一篇读书报告。笔者的衷心希望是能得到机器翻译界的先驱者及同行们的指教。

关于“以歧义对歧义”进行翻译的想法，曾得到马希文教授的启发，在此致以谢意。

参 考 文 献

[1] 刘伶等，语言学概要，北京师范大学出版社，1983年11月第1版，p21-28，p11-p120

[2] 马希文，自然语言理解，计算机工程与应用，1987年第4期，p18-p21

[3] 长尾真，言语工学，株式会社昭晃堂，1986年4月初版4刷p145-p163

[4] Phillip J.Hayes and Jaime G.Carbonell,
A Tutorial on Techniques and Applications for Natural Language Processing, Carnegie-Mellon University, October 1983

[5] 俞士汶，如何看待机译系统译文的质量，计算机信息报，1988年9月6日，总第215期增版，第23页

Ambiguity of Natural Language and the Approaches of Machine translation

Yu Shiwen

(Peking University)

Abstract

Ambiguity is one of the characteristics of natural languages. To develop machine translation system, not only must we analyse the ambiguities inside a language, but also we have to study the ambiguities between two different languages. This paper examines these ambiguities from various angles and discusses their influence on machine translation. In machine translation system some approaches are adopted to obtain a proper translation. This paper includes these approaches and proposes a viewpoint or a method, which enables the ambiguities of translation to correspond the ambiguities of the original.