清华英汉机译系统中的歧义处理

刘月荣　苏玉宏　陈圣信
（清华大学外语系／自动化系）

摘要： 本文着重讨论自然语言中的歧义问题，概括了我们在机器翻译研究中应用的一些理论及处理问题的方法，提出了"广义近邻"的概念．文章首先讨论了歧义(Ambiguity)的分类，其次讨论了处理歧义的一些策略和方法，包括查字典消歧，语法消歧，语义消歧以及统计消歧．其中部分策略和方法已在清华英汉机译系统中实现，并已取得了良好的效果，有效地提高了机译的准确率．

关键词： 机器翻译　广义近邻　语法消歧　语义消歧

On Resolving Ambiguities in Tsinghua English-Chinese Machine Translation System

Liu Yuerong　Su Yuhong　Chen Shengxin
(The Department of Foreign Languages/Automation, Tsinghua University, 100084)

Abstract: Ambiguity is a key problem in English-Chinese machine translation (ECMT). Its resolution can greatly influence the accuracy of an MT system. This thesis aims at the study of the various resolutions of ambiguity in English-Chinese machine translation. Besides dictionary consulting, syntactic and semantic strategies have played important roles in resolving ambiguities. The concept of "generalized neighboring" has been proposed. The disambiguation mechanisms for resolving ambiguities have proved to be effective when tried in the Tsinghua ECMT system.

Key words: machine translation, generalized neighboring, syntactic disambiguation, semantic disambiguation

一　引言

清华英汉机器翻译系统（简称 THECMT 系统）采用基于规则的方法，并辅之于语料库的知识，实现了程序与规则的完全分离，系统具有良好的开放性．在处理歧义方面，尽可能地利用了语言中的约束关系，引入了广义近邻、语法匹配以及语义分析等模式，采用多层次的处理方法，比较好地解决了歧义问题．本文将就这一问题做如下探讨．

二　语言中的歧义问题

自然语言与人工符号化的语言（如计算机语言）不同，歧义是其固有特性。正如一些语言学家所说，“语言无处不歧义”，机器翻译正是在“歧义的海洋”里游泳，排除歧义贯穿在整个机器翻译过程的始终，其解决的好坏直接关系到译文的质量。

在语言中，歧义主要分为两类：一类是词的歧义，另一类是结构歧义。

1、词的歧义

词的歧义又可以分为两种情况：词类歧义和词义歧义。先来看词类歧义。

词类歧义着眼于同一词属于不同语法范畴，例如，“book”一词做名词用时，含义为“书”，做动词时，其含义是“订购”，二者就分别属于不同的语法范畴，即不同的词类。

英语中词的歧义共有多少种类型？对于这个问题，至今还没有权威的统计资料。我们依据一本英汉小词典做过不完全统计，结果虽很不完全，但却能说明问题。统计结果为下：

英语中词类的兼类类型

(1)	形容词＋副词	alike
(2)	形容词＋副词＋名词	best
(3)	形容词＋副词＋名词＋动词	back, better
(4)	形容词＋副词＋介词	above
(5)	形容词＋副词＋代词	all, any
(6)	形容词＋名词	acid
(7)	形容词＋名词＋动词	advance, average
(8)	形容词＋代词	both
(9)	形容词＋动词	fast, alternate
(10)	形容词＋介词＋连词	after
(11)	副词＋介词＋连词＋代词	as
(12)	副词＋名词	well
(13)	副词＋介词	aboard
(14)	名词＋动词	bargain, abuse
(15)	动词＋助动词	do, have

实际中英语词类兼类类型可能远不止以上十五种。如果能对英语中的词汇的歧义进行科学地分类，并给出每种词类歧义类型的概率分布，对于排歧义处理是相当有帮助的。

词歧义中还有相当一部分属于词义歧义。这是指一个词可以仅有一种词类，但有多种词义。在英语中，这类现象相当普遍，据《牛津英汉大词典》统计，介词“of”就有63种含义，“in”有40种，“strong”一词，在形容人时，含义是“强壮的”，形容建筑物时，含义是“坚固的”。

不难看出，词类歧义考虑的是词在语法上的多义问题，而词义歧义则主要研究词在语义方面的多义问题，对于具体词而言，其歧义往往既包含着词类歧义，也包含词义歧义，是两种歧义的统一体。

2． 结构歧义

结构歧义非常复杂。在含有结构歧义的句子中可能每一个词都具有唯一的词性，如：

[例1] He looked at the girl with a telescope.

这是语言学上的一个著名结构歧义句，有两种不同的解释。

(1) 他用望远镜看那个女孩。

(2) 他看带着望远镜的女孩。

上面句子产生歧义的原因是，介词短语“with a telescope”可看作不同的句子成分，一种是作定语，一种是作方式状语，不过，并非以上这类结构的句子都有歧义，如：

[例2] He looked at the table with four legs.

这个句子就只有一种合理解释，"他看那个有四条腿的桌子"。

由此可以看出，同样的结构，对于由具有不同语义属性的词所构成的句子，并非同样地为歧义句．也就是说，有些结构歧义与语义是密切相关的。

下面再看一个修饰语与并列成分引起的结构歧义。

[例3] fresh milk and butter

可能的两种解释为：(3) 鲜牛奶和黄油（无省略[fresh milk] and [butter]）

(4) 鲜牛奶和鲜黄油（有省略fresh milk and fresh butter）

译文(3)也有结构歧义，它包含了(4)的含义，翻译时可以采用以歧义对歧义的译法。

以上讨论了产生结构歧义的两种情况，实际中的结构歧义远不止上面两种情况。

结构歧义的解决往往要考虑上下文、甚至整个篇章的理解，目前机器翻译对大量的这类问题无能为力．当然，有些结构，即使结合上下文，透彻理解整个篇章，仍然解决不了，自然语言允许多义的存在，象汉语中的双关，本来就是一种修辞手法。

三 歧义的处理

在英汉机译研究中，我们利用了约束关系理论处理歧义．所谓＂约束关系＂，我们指语言中作用与被作用的关系．自然语言中的约束关系包括语篇约束，语法约束和语义约束等，语法约束又有词尾约束，相邻词约束以及词团约束等多种情况。

消歧是一个多层次的处理过程，主要有以下三个层次：

1． 查字典消歧：

在这一部分利用了语篇约束，词尾约束及语料约束，语篇约束通过专业词典优先查找的原则来实现，思路很简单。

词尾约束消歧义，可归纳出三条原则：

(1) 去掉“(e)s”等可恢复原形的，可能是名词和动词。

(2) 去掉“(e)r”，"(e)st”可恢复原形的，可能是形容词或副词。

(3) 去掉“ing”及“ed”可恢复原形的，可能是动词。

关于语料约束的应用，要借助于搭配词典，我们认为这方面重视不够，语料库的思想往往与传统的基于规则的方法截然分开，其实语料库的思想方法完全能被传统的基于规则的方法所采用，下面通过例子来看语料约束的意义：

先看看形容词与名词修饰关系，以“heavy”为例：

heavy	crops	丰收
heavy	fog	浓雾
heavy	traffic	拥挤的交通
heavy	sky	阴郁的天空

下面再看一看动词和名词的关系，以“develop”为例：

develop	a	formula	推导公式
develop	a	film	冲洗一张胶片

从上面可以看出，形容词“heavy”修饰不同名词，意义差别很大，动词“develop”接不同宾语，含义也大大不同，如果仅靠语法分析，即使分析完全正确，也得不到正确的译文，也许有人会说，把上面的这些词团看成惯用语不就行了吗？粗看起来似乎如此，但仔细想一想情况并非那么简单，之所以看成搭配，是因为它们具有“变形”能力，对于形容词＋名词这类搭配，有两种可能变形，如：

heavy sky	阴郁的天空
heavy gray sky	阴郁的灰色的天空
The sky is heavy	天空是阴郁的．

其一中间有插入成份，其二以作表语形式修饰名词．

对于动词＋名词这种搭配，也有两种可能变形，情况类似．

我们把搭配看成可变的，主要的搭配模式以上面两种为主．

语料约束可以解决语法约束及语义对应约束所难以解决的问题，应引起高度重视．

2． 语法消歧

在这一部分，利用了近邻约束、词团约束、及惯用法说明，在此不多重复．我们主要讨论一下“广义近邻”的概念，以及排歧义中的确定性算法．

让我们先来看一看广义近邻这一概念，它有别于位置近邻，例如：

[例4] a man　　[例5] a very very old man

在[例4]中，词“a”与“man”在位置上是相邻的，这是一般意义上的相邻，在[例5]中，“a”与“man”在位置上不相邻，但从含义上是“相邻”的，“man”是该词团的中心词，“a”与“man”存在修饰关系，这类相邻关系为广义相邻，也就是说，一个词“跳过”一些辅助成份而与另外一个词相邻．在英语中，并非相邻的词关系就密切．广义相邻拓宽了我们的视野．这一概念对排歧义和句法分析都很有价值．

在歧义处理中，我们还应用了Marcus的确定性算法的思想，以下例子可说明该问题：

[例6] the red novel book

		名词		名词		名词
歧义情况为	冠词	＋	＋		＋	
		形容词		形容词		动词

如果只考虑近邻关系，不用确定性算法，看到冠词后有名词，于是就急于做决定，认为第二个词是名词，从而就产生了错误，由此可以看到近邻关系的局限。引入广义近邻概念之后，情况则有所不同，看到冠词“the”后，我们并不因为“red”有名词词性就做出它是名词的决定，而是“等待和观察”，跳过“red”，观察到第三个词“novel”也有歧义，继续往下看，直到找到“book”---“the”的广义近邻，才可以做出决定。

把广义近邻的概念和确定性算法相结合，能够提高排歧义的准确率，我们的系统正是运用这种方法，取得了很好的效果。

3. 语义消歧

语义消歧主要利用了动词与名词的施事与受事关系、形容词与名词之间的修饰与被修饰关系以及词典中的语义标注。先看形容词与名词的几种语义对应约束模式.

[例7] a strong man “strong”和 “man”是相邻的，作定语；

[例8] a strong young man “strong”仍然作定语，但中间有插入成份；

[例9] The man is very strong . “strong”以表语形式修饰名词“man”；

[例10] I'd like my son to be strong.

"strong" 仍为表语，但此时是宾补结构 "to be strong" 中"be"的表语；

[例11] We think him strong.

“strong" 为宾语补足语，位于所修饰词 "him" 的后面；

[例12] Strong as he is, he can not lift this heavy box.

“strong" 仍为系动词的表语，但此时并未紧随 "is" 之后，而是位于句首.

这方面与前面讨论类似，但有所不同的是，前面是对具体词而言的，解决语义标注无法解决的问题，而通过搭配来解决，在这里，我们通过词典中的语义标注，通过语义对应关系来解决语义问题，是针对一类问题而不是具体词的。但二者的作用机制是一致的。

动词与名词之间语义约束的模式首先有主－谓关系和谓－宾关系，如以下两例所示：

[例13] He raised a dog.(raise : 饲养，举起)

[例14] The sun rises.

动词 "raise" 和 "rise"的意思分别需要利用名词 "dog" 和 "sun"的语义属性确定.

除了以上两种基本模式外，还有一些衍生出的较复杂的模式，如以下例句所示：

[例15] The old lady has a house to let. (let 的宾语：house)

[例16] The man was too weak to rise. (rise 的主语：the man)

[例17] The prices are reported to have risen again. (risen 的主语：the prices)

[例18] Go and get your hair cut. (cut 的宾语：your hair)

对以上划线的动词进行消歧，同样需要借助它们的主语或宾语的语义属性。但是，由于这些动词的主语或宾语在句中所处的位置不象在例13和例14中那样容易被机器识别，所以，不经过一些深入的语法分析，弄清词与词之间的逻辑关系，消歧的工作将难以进行。

下面来讨论语义消歧的一些局限性。在例13中，我们并不能完全排除“举起”一条狗这一可能，因为狗能被饲养，亦能被举起。然而，在“He raised a big elephant.”中，“raise”的"举起"一意基本上可被排除，因为按常理，大象很重，一般人无法举动。

利用语义消歧存在着两大困难．其一是知识表示问题，如上所述，我们如何表示哪些东西能够被举起，哪些东西不能被举起，而它们的区别界限是模糊的，甚至是因人而异的，张三能够举起的东西，李四未必能举起，因此，要准确地描述语义信息几乎是不可能的；其二是语义作用范围及知识完备性问题，如上文提到的“raise”接动物类名词表示饲养，对于动物这一概念，我们很难给出一个大而全的定义，猪、牛是动物，但“闪电”也可以是动物，要穷举各种可能性几乎不太可能．另外，同一个词，完全可以属于不同的语义范畴，这些不同的语义范畴可能不完全独立，存在交集或包含关系．

在解决歧义时，我们采用了以“语法为主，语义为辅”的原则，对于词类歧义，主要靠语法约束解决问题，对于词义歧义，可用语义约束解决部分问题．

4．统计排歧

有人将每个词的用法作统计，给出每种含义的使用概率，对一句话，可以找出合适的目标函数，使总体概率最大，从而实现排歧义处理．这里隐含着两个假设：统计概率的样本有典型性；小概率事件不会发生．但在自然语言中，这两个假设都很难得到保证．目前统计排歧也取得了一些有意义的结果，但总的来说，只靠统计方法处理歧义是不够的．

对于语法约束、语义约束都难以解决的问题，统计排歧可以做为有益的补充手段，但我们不做大量而复杂的词用法统计工作，而采用“高频先见”的原则，将最常用的词义收在字典的前面，在语法语义处理解决不了的时候，优先选取最常用的词义．这里的概率是估计出来的，当然不可能完全精确，但可以根据实际情况再做调整．

四　结论

不可否认，排歧义是一项非常艰难的工作，没有一种方法可以完全解决这个问题．在歧义处理过程中，我们综合利用了语法和语义知识，但是由于语法研究尚不够深入，字典中语义属性的标注难免有不完善之处，而且语义分析本身就具有不确定性，因而对某些歧义句的处理尚不能令人满意；此外，歧义的处理还常常需要利用上下文的信息，目前我们的系统还缺乏上下文相关处理的能力．这些是我们将来努力的方向．

参考文献

[1] 刘月荣“利用规则消除歧义提高英汉机器翻译的正确率”清华大学硕士论文，1995年3月

[2] 苏玉宏 “英汉机译中规则与实例的结合与实现”清华大学硕士论文，1995年6月

[3] P. F. Brown, S. F. Chen, etc., "Automatic speech recognition in machine-aided translation", *Computer Speech and Language*, 1994. 8.

[4] 陈力为主编《计算语言学研究与应用》，北京语言学院出版社1993年10月 第一版