机器翻译中的消歧策略

朱靖波 王宝库 姚天顺
东北大学计算机科学与工程系
辽宁·沈阳 110006

【摘要】本文首先阐明了英汉机器翻译中所要面对的问题，并论述了英语的三种模式关系：邻接模式关系、搭配模式关系和相关模式关系。然后我们提出了一种基于这三种模式关系的消歧策略，目前已应用于英汉翻译系统中。实践证明，它不仅提高了系统的效率，而且较好的解决了机器翻译中的难题。

关键字：机器翻译，邻接模式关系，搭配模式关系，相关模式关系

An Approach for Disambiguation in Machine Translation

Zhu Jingbo, Wang Baoku & Yao Tianshun
Dept. of Computer Science and Engineering
Northeastern University
Shenyang, 110006
P.R.China

ABSTRACT: The paper first expounds the problems which we must face in English-Chinese machine translation system, and describes three kinds of English pattern relationships: contiguity pattern relationship, collocation pattern relationship and cohesion pattern relationship. We provided an approach for disambiguation based on the three kinds of pattern relationships which is applied in our English-Chinese machine translation system, and is proved to be a way to improve the efficiency of the system perfectly.

KEY WORDS: machine translation system, contiguity pattern relationship, collocation pattern relationship, cohesion pattern relationship

一、引言

任何一种语言，无论是程序语言还是自然语言都有形式和内容两个不可分的部分。在语言形式上，表现为语法，在语言内容上，表示为语义。如果一种语言，语言形式完全决定了语言的意义，那么机器对这种语言的理解是无二义的。汉语和英语都是口头形式和书面形式十分发达的自然语言，在语言特点上有很大的区别。不管在英语分析还是在汉语生成中，都存在许多难免的歧义现象。因此对于任何不同语种之间的机器翻译系统来说，消歧处理过程是一个很关键的环节。

通过对大量的英语句子的分析，文中提出了英语的三种模式关系：邻接模式关系、搭

配模式关系和相关模式关系。然后我们提出了一种基于这三种模式关系的消歧策略，目前已应用于英汉翻译系统中。实践证明，它不仅提高了系统的效率，而且较好的解决了机器翻译中的难题。

二、英汉翻译中的一些问题

§2.1 兼类问题

注：(E)：英语句子　　(C)：汉语译文

例 1:
1(E). I am discussing a problem with TOM .
1(C). 我正与汤姆在一起讨论一个问题.
2(E). This is a cup with a broken handle.
2(C). 这是一个把手破损的杯子.
3(E). To write with a pen !
3(C). 用钢笔写!

在英汉词典中，词语'with'只有一个介词(prep)词性，在句子中带介词短语起修饰作用，但它有多个义项，例子1中给出了'with'的三个义项。这种现象在英汉机器翻译中很常见，称作兼类问题。在我们的英汉翻译系统中采用的是句法语义一体化的分析方法，是比较符合人类理解自然语言的过程。然而，如何去确定词汇的语义呢？对于例子1的兼类问题，我们英汉翻译系统采用词语'with'的邻接模式关系（将在下文中讨论）来解决。邻接模式关系规则描述采用我们自行设计的一种规则描述语言CTRDL（详细书写描述规范见[1]）书写。

a) ^'with'+(1111,1113) => @selectmeaning("与<...>在一起")
b) (n;r,114)+^'with'+(n,¯111) => @selectmeaning("<...>的")
c) (v,21)+^'with'+(1126) => @selectmeaning("用<...>")

这是三条邻接模式关系规则，可用于确定'with'的语义。其中，'+'表示词结点的邻接符，'n'，'r'，'v'分别表示名词，代词，动词。如规则(c)，如果介词宾语是工具(1126)，并且介词短语所修饰的句法成分为动作性动词(v,12)，则词语'with'被译为"用<...>"。实际上，英语在结构上存在四个层次：词、短语、简单句和复杂句，依此，我们在分析策略上采用多层次分析：词法分析、句法语义一体化分析和复句分析。因而消歧策略也具有多层次：词法层消歧、语法层消歧、语义层消歧，也就是说消歧处理过程存在于分析的各层次中。

§2.2 其它歧义现象

例 2:
1(E). Mary likes the beautiful picture hung on the wall which her mother bought yesterday.
1.1(C). 玛丽喜欢墙上那张漂亮的图画，这张图画是她妈妈昨天买的.

1.2(C). 玛丽喜欢墙上那张漂亮的图画， 这个墙是她妈妈昨天买的.

例子2对于分析器来说具有一定的歧义性，因为从英语的一般结构分析来看，分析器不知道她妈妈昨天买了什么。是漂亮的图画还是墙？在英汉机器翻译中称作句法结构歧义问题。通过大量的英语句子分析我们发现，这种句法结构歧义问题大多因为句子中引进修饰性状语。如果例子2中去掉地点状语"hung on the wall"，这样就不存在歧义现象。对于例子2中存在的句法歧义问题可采用搭配模式关系规则来解决（详细技术将在下文介绍）。但也存在一些情况难以处理，如例子：

例 3:
1(E). Mary likes the beautiful toy on the desk which her mother bought yesterday.
1.1(C). 玛丽喜欢桌子上那个漂亮的玩具， 这个玩具是她妈妈昨天买的.
1.2(C). 玛丽喜欢桌子上那个漂亮的玩具， 这张桌子是她妈妈昨天买的.

下文将论述英语中三种模式关系：邻接模式关系、搭配模式关系和相关模式关系。

三、三种模式关系及其消歧处理

系统采用词汇语义驱动算法，详细算法技术参见文[2]。词汇语义驱动不但提升了具备多重属性的词汇在分析中的地位，而且强调各类众多属性之中，语义属性在语言分析中所起的重要作用。

§3.1 模式关系对象描述

模式关系的描述对象是以词为基本对象单位的语句及其中间结构（如语法树，概念语义网等）。每个词语的信息，表示为：

概念（属性1，属性2，...，属性n）

其中概念常用词语本身表示，有时也可以省略，但在句子分析时，描述词结点的概念也常用词结点位置来代替表示。

例4:
<1> ATTRIBUTE-DESCRIPTION('I') = 'I'(the first person,singular number,||||,...)
<2> 如句子：I have some books.
其中^表示词结点'have'，则
ATTRINUTE-DESCRIPTION('I') = I^(the first person, singular number,||||,...)

§3.2 邻接模式关系及其消歧处理

对于任一词结点I，它的邻接模式关系可表示为：

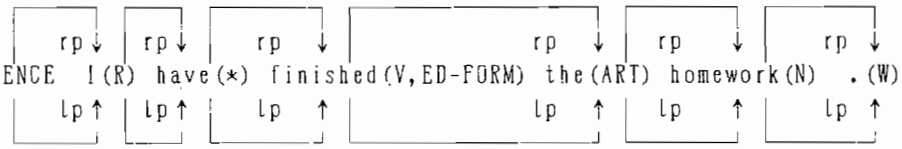
ITEM-DESCRIPTION(I-M)+...+^LEX(I)+...+ITEM-DESCRIPTION(I+N)
==>
ITEM-ACTION(I-M), ..., ITEM-ACTION(I), ..., ITEM-ACTION(I+N)

意义：当词结点i满足ITEM-DESCRIPTION(i)，则对词结点i有关语法语义属性进行操作，其中i满足[I-M,I+N]。‘＋’称为词结点的邻接符，表示不同词结点之间的顺序性和邻接性，在逻辑测试上表示与的关系。^表示当前操作指针的位置。ITEM-DESCRIPTION(i)指出词结点i的某些属性。ITEM-ACTION(i)指出对词结点i有关属性的操作，包括选词性、确定语义等等，LEX(i)表示取词结点I的词语。

例7：

I have finished the homework.

分析：



匹配'have'的邻接模式关系规则：

```
LEX='have'
     (n;r)＋^'have'＋(v,ED-FORM) => @selectcat(aux),^1.TENSE:=PERFECT
```

译文生成：

我已经完成了功课。

其中SENTENCE表示句子的头指针，*表示词语with存在兼类问题，rp表示指向右边词结点的指针，lp表示指向左边词结点的指针，词性aux表示助动词，函数@selectcat用于确定have的词性，ED-FORM表示动词的过去分词或过去式。"^1.TENSE:=PERFECT"表示将右边动词结点(即"finish")的时态设置为完成时态。

§3.3 搭配模式关系及其消歧处理

搭配模式关系所描述的对象并不要求一定相互邻接，但要求在语法或语义关系上存在一定的关系，如两者之间存在施事(AGT)关系，或者主谓关系，或者是父子关系，或者存在某一个兄弟或子孙等等。而邻接模式关系中的对象描述主要注重于词结点本身的属性描述，常用于解决虚词的兼类问题，而搭配模式关系中的对象描述不仅描述词结点本身的属性，而且描述某词结点对另一词结点的语义框架约束或语法结构上共存的约束等。

对于任一词结点I，它的搭配模式关系可表示为：

```
    ^LEX(I), ITEM-DESCRIPTION(I-M), ..., ITEM-DESCRIPTION(I+N),
    ITEM-COLLOCATIONBIND(I-m, RELATION, I+n), ...
==>
    ITEM-ACTION(I), ITEM-ACTION(I-M), ..., ITEM-ACTION(I+N)
```

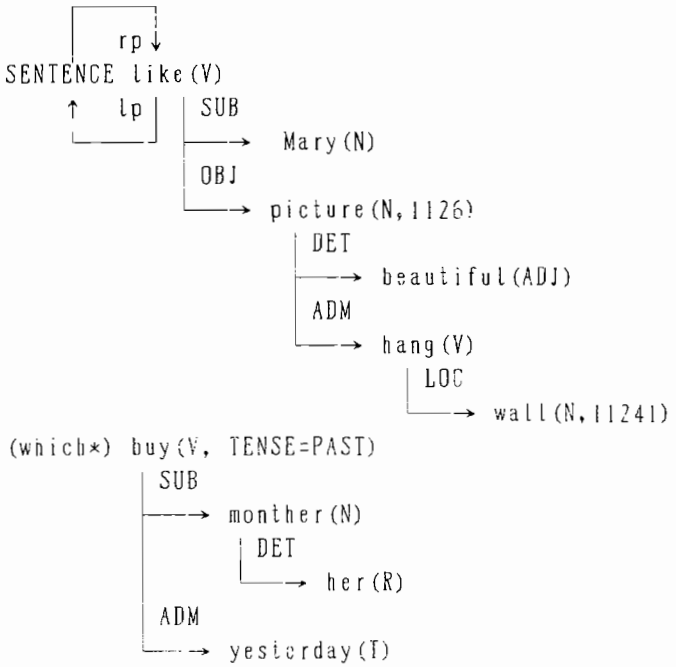
意义：若词结点I的某左边词结点I-m满足ITEM-DESCRIPTION(I-m)，某右边词结点I

+n满足ITEM-DESCRIPTION(!+n)，词结点I+n满足词结点I-m的RELATION关系的约束，其中m，n满足[0...M]，[0...N]，则对词结点I或其他词结点的有关语法语义属性进行操作。其中‘，’称为词结点的与关系符，只表示不同词结点之间是与关系。^表示当前操作指针的位置。ITEM-DESCRIPTION描述词结点的某些属性。ITEM-ACTION指出对某词结点有关属性的操作，包括选词性、确定语义等等，LEX(I)表示取词结点I的词语。

例8：

Mary likes the beautiful picture hung on the wall which her mother bought yesterday.

分析：



匹配'which'的搭配模式关系规则：

```
LEX='which'
    ^1,#1^(n),@prebinding(^1,OBJ,#1^)
=>
    @setsemrela(^1,DET,#1^)
```

译文生成：

玛丽喜欢挂在墙上的那张漂亮的图画，它是她妈妈昨天买的.

若当前词结点(即"which")左边存在某名词结点满足词结点("buy")的OBJ语义关系约束，则在词结点("buy")与该名词结点之间建立DET语义关系。

§3.4 相关模式关系及其消歧处理

相关模式关系与邻结模式关系、搭配模式关系的最大区别在于它所描述的对象可能或根本不存在直接或间接的句法和语义关系，只注重于它们两者的语义属性存在相关性，以此来解决一些歧义问题。

对于任一词结点I，它的相关模式关系可表示为：

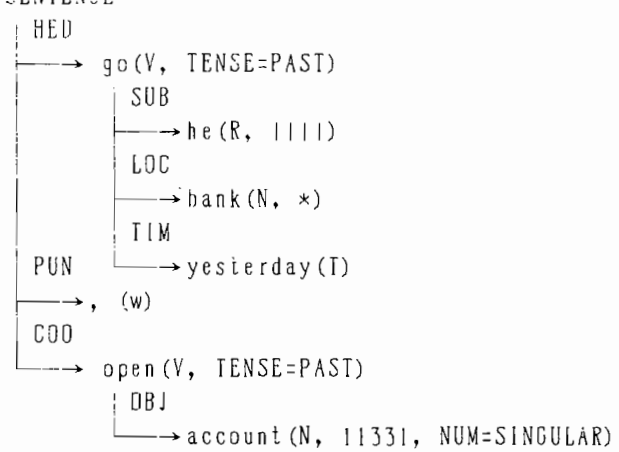
```
^LEX(I), ITEM-DESCRIPTION(I-M), ..., ITEM-DESCRIPTION(I+N),
ITEM-INTERACTION(I-m, ATTRIBUTE, I+n), ...
==>
ITEM-ACTION(I), ITEM-ACTION(I-M), ..., ITEM-ACTION(I+N)
```

意义：若词结点I的某左边词结点I-m满足ITEM-DESCRIPTION(I-m)，某右边词结点I+n满足ITEM-DESCRIPTION(I+n)，词结点I+n与词结点I-m存在语义属性ATTRIBUTE相关，其中m，n满足[0...M]，[0...N]，则对词结点I或某词结点的有关语法语义属性进行操作。其中‘，’称为词结点的与关系符，表示不同词结点之间的与关系。^表示当前操作指针的位置。ITEM-DESCRIPTION描述词结点的某些属性。ITEM-ACTION指出对某词结点有关属性的操作，包括选词性、确定语义等等，LEX(I)表示取词结点I的词语。

例9：

Yesterday he went to the bank, and opened an account.

分析：



匹配'bank'的相关模式关系规则：

```
LEX='bank'
    ^'bank',@searchnode(^,(n;r;v),#1), @interaction(^,RES.113,#1)
=>
    @selectmeaning("银行")
```

译文生成：

昨天他去了银行，开了一个帐户。

由于使用邻接模式关系和搭配模式关系都无法解决"bank"的译文是"河岸"，还是"银行"采用相关模式关系如果还解决不了的话，我们可以认为这个句子本身具有二义性。相关模式关系往往作用于不同分句之间，其中规则中函数@searchnode(^,(n;r;v),#1)表示从其他子树（即根词结点为'open'或','）中查找满足(n;r;v)的所有词结点#1，同时调用函数@interaction(^, RES.113, #1)测试词结点^(即"bank")与词结点#1(即"open"或"account")的语义分类码是否都属于113(即商业金融类)。如果存在满足条件的词结点(如例10中的"account")，函数@selectmeaning("银行")将词语"bank"的译文确定为"银行"。

四、进一步讨论

上文介绍了英语中的三种模式关系：邻接模式关系、搭配模式关系和相关模式关系，及其应用于英汉机器翻译中的消歧处理。实践证明，它不仅提高了系统的效率，而且较好的解决了机器翻译中的一些难题。但是自然语言中许多句子本身存在二义性，只有在上下文中才能确定它的意思。对于单纯从句子本身出发无法解决二义性的句子，系统采用多解输出。有一些特殊情况，如例2，"buy the wall" 虽然也有可能，但与"buy the beautiful picture" 相比较来言，可能性相差太大，对于这种情况，系统目前承认译文1.1(C)是对的。我们希望通过提高系统的消歧能力，以达到提高系统翻译能力的目的。

参考文献

[1] 王宝库，张中义，姚天顺，机器翻译系统中的一种规则描述语言(CTRUL)，Vol.5,No.4，中文信息学报，1991年4月

[2] 姚天顺，王宝库等，"词汇语义驱动方法"，<<机器翻译研究进展>>，电子工业出版社，1992,8

[3] 唐泓英，姚天顺，"基于搭配词典的词汇语义驱动算法"，<<软件学报>>，VOL.6 1995,78-85