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

机器翻译中词典和文法的关系

董振东

(中国软件技术公司)

【摘要】 为机器翻译设计和研制的词典和文法，是机译系统的核心，是机译研究中的关键。词典和文法的关系及其处理一直是机译研究者所关心的问题。本文将从四个方面来论述它们的关系：(1) 个性与共性的关系；(2) 静态与动态的关系；(3) 语法与语义的关系；以及(4)信息与算法的关系。这几个关系中既涉及到词典和文法客观存在的关系，也涉及到我们对两者的主观处理的关系。

为机器翻译设计和研制的词典和文法，是机译系统的核心，是机译研究中的关键。大家知道，面向机译的词典和文法与面向人的词典和文法是很不相同的。原封不动地照搬面向人的词典和文法用于机译系统是断然行不通的。之所以如此，并不是由于面向人的词典和文法不能反映语言的客观现实，而是由于它们尚不能构成一个显性的体系。机器翻译要求其词典和文法必须是一个系统，一个有机的、显性的、精密的系统。这样，机译研究就面临着一个如何科学地反映和处理词典和文法的关系，进而使之成为一个系统。应注意，这里既有客观地反映它们之间的关系的问题，也有科学地处理它们之间的关系的问题。总的说来，有如下四种关系：(1)个性与共性；(2)静态与动态；(3)语法与语义；(4)信息与算法。本文将从上述四个方面做一概要的讨论。

一、个性与共性的关系

在机译系统中，词典是一个词汇数据库，是逐一地、具体地、详尽地词汇特性描述的汇集。它反映了语言中词或词组的词法、句法和语义（还可能包括语音）的特性。另一方面，在机译系统中，文法是一个规则系统，是概括的、抽象的、形式化的文法规则的集合。它反映了语言中词或词组的自身固有特性在进入特定的、许可的结构后所表现的各种句法或语义的关系。因此，可以认为词典所反映并加以处理的是语言中的个性问题，而文法所反映并加以处理的是语言中的共性问题。

例如在汉语中关于名词词组组合我们可以有这样一条规则：

数词＋量词＋名词（如：一头牛）

这是一条高度概括的、体现共性的文法规则。但是，对于具体地对于每一个汉语的名词而言，其可以带的量词是各不相同的，是规定性的，或说是约定俗成的。对于汉语名词的这种个性，必须在词典中给每一个名词逐一加以标记。

本文1988年7月13日收到。

再例如在英语中关于限定词的选用我们可以有这样的规则：

限定词 a ＋辅音为首的词

限定词 an ＋元音为首的词

同样的，这样高度概括的共性规则，也必须靠词典中对每个词的个性描述来保证。即词典中每一个词都应标记其是以辅音还是以元音为首。

上面的例子主要讲的是词典反映个性，文法反映共性。但在机译系统的研制中，不仅仅是“反映”，还存在着一个“处理”的问题。

在讨论研究词典如何处理个性的时候，人们常常提出一个“分类”的问题。他们不仅要把词分为动词、名词、副词等。而且对每一类还要再细分为次类、子类，诸如一个包括类、纲、目、属、科、组等的分类体系。这样一种体系性分类的企图似乎是很科学的，但笔者认为它在理论上恰是对于“个性”这个概念认识不足，同时在实践上会遇到很大的困难，甚至会导致失败的。要知道，词的个性是很强的。个性强具体体现在它们不同于其它词的特性可能是多种多样的。这样，就给分类带来了多种多样划分类、属等的标准。标准愈多，分类愈困难，因为难以排除交叉重叠。同样的两个词可能按甲标准应划为一类，而按乙标准却应分属于不同的类。什么是理想的处理方法？笔者曾多次强调：研究者要摒弃“分类”的观念，代之以树立“属性给定”的观念。即使要作一定的分类，也应遵循“分类宜粗不宜细，而属性给定则是宜细不宜粗”的原则。具体的做法是：面对各种词汇，按词法、句法、语义等特性，列出一份“词汇属性信息表”，然后把所要填于词典中的词或词组逐一地为其填写信息表中所列出的、它应该有的属性项。有人认为“分类”可以体现词的属性的关系如上位、下位等。我们认为“分类”造成的交叉重叠势将扰乱这种关系。而“属性给定”同样可以体现属性的关系。属性关系的真正的描述，应依靠文法的建立。而语义属性关系的真正的描述，应依靠“概念关系文法”来确立。“概念关系文法”是一个知识的形式化描述系统。（关于“概念关系文法”，笔者将另文专门讨论。）这正是共性寓于个性之中。

二、静态与动态的关系

一般地说，词典中列出的词本身的语法和语义等属性是固有的、特定的、具体的。这些属性在没有进入特定的语言环境之前是静态的。词的静态属性只有在进入特定的语言结构后才被激活。这种可以激活词的静态属性的语言结构，就是语法结构，它的规范性是由文法规则加以规定的。同时，在机译系统中，文法是一个系统，多还包含若干个子系统，它们体现了语言处理的推导过程。从上述两个方面看，文法则是动态的。

词作为概念的载体，具有三个基本属性：语义、语法和语音，其中语义属性是基础性、主导性的。词的语义属性由两部分构成：语义核（K），又可称概念语义或范畴语义；语义因子（S），又可称特征语义或抽象语义。

任何一个词都可以这样表述：

$$W::=<M><G><P>$$

这里W表示词，M表示语义属性，G表示语法属性，P表示语音属性。语义属性又可表述为：

$$M::=<K><S_1,S_2,\cdots,S_i>$$

试以如下各词的语义属性为例来作进一步说明：

“纸”

K = 通常呈薄页状，用以承受和呈现书写，图画的对象的物资

S_1：通常软、薄可成形

S_2：可以着色

S_3：本身也可以有多类颜色

S_4：可燃

S_5：极易燃

S_6：可呈现写或画的结果

S_7：多用植物纤维制造

⋮

⋮

“点燃”

K = 使燃烧

S_1：使对象处于燃烧状态

S_2：应以某种易燃物资为工具

S_3：瞬间性动作

S_4：意志性

S_5：可接受动作方式的描写

⋮

⋮

“有”

K = 表示领属

S_1：仅有领属关系，无对象

S_2：对于被领属体无处置性

S_3：仅说明状态

S_4：无意志性

⋮

⋮

上述词的语义属性都可以在词典中静态地加以描述。这样的静态的属性进入语言结构时，常常并不是全部属性同时被激活的。通常的情况是：在进入语言结构时，词 a 的某一属性与词 b 的某一属性被激活。例如：当“纸”和“点燃”两词分别进入如下结构时，它们被激活的语义属性是不同的：

(a)纸被点燃了

(b)他们用纸把树枝点燃了

在例(a)中，“纸”被激活的属性是 S_4(可燃)；“点燃”被激活的属性是 S_1（使对象处于燃烧状态）。但在例(b)中，“纸”被激活的属性却是 S_5（极易燃）；而“点燃”被激活的属性是 S_2（应以某种易燃物资为工具）。我们把在语言结构中词与词之间这样的语义属性激活状态称为“语义撞击”。

语义的合理撞击一方面要取决于词本身的静态属性，另一方面也必须依赖于语法结构的手段。如上例(a)和(b)中的词序以及虚词“被”、“用”和“把”都是规定性的结构手段。当然，语义的合理撞击的决定性因素是词的语义属性。例如下面的例子之所以不成立，主要是“有”这个词缺乏如“点燃”，“写”等的某些属性：

X(c)书被我有了

X(d)我把书有了

在具体的机译系统制的实践中如何处理好词典的静态描述和文法的动态描述的关系？这里当然包含着许多具体的处理技术和诀窍。但最重要的是要把握住这样一点，即以文法的描述为主要环节的系统整体观念。因为词典的静态描述，尤其是其中的语义属性是可以无穷无尽的，但是文法的动态描述是可以有限的。

三、语法与语义的关系

一个好的机译系统的词典和文法都将各自充分地反映其所处理的语言的语法与语义的种种问题。换句话说，词典里包含着语法与语义的问题，文法里也包含着语法与语义的问题。应该怎样处理好这种关系呢？我们认为总的原则是：就词典而言，在为每一词条给定属性时，应尽可能做到语法与语义相互独立；而就文法而言，则不论是算法还是规则却应做到语法与语义的有机结合。

具体地说，在建立词典时，我们为词典划分出独立的信息区。每一个词条都含有如下信息区：词条信息管、词法信息区、句法信息区、语义信息区和词义信息区。应再次强调，务必摒弃那种把语法范畴和语义范畴相混合的分类的观念和做法。

例如，一个汉语动词可能在上述五个信息区内相互独立地登录如下种种信息：

词条信息区：读音、音节数、笔划数、部首、使用频度、同形类别等

词法信息区：词性、构词方式等

句法信息区：句型、名词宾语的格关系样式，与趋向动词连用的样式等

语义信息区：意志性、延续性、方向性、动状性，配价、一般性施受限定和其它特殊格关系要求等其它语义因子

词义信息区：语义核

一般说，前三者的信息应尽可能充分，甚至穷尽。这是可能做到的。但是语义信息则无论是在理论上还是实践上都是不可能穷尽的，因此只能根据可能与需要求加以确定和给出。

另一方面，文法将综合地利用词典中相互独立地给出的语法和语义的信息。虽然构成文法的基本模块或称规则集合的功能在语法或语义上可能有所侧重，但是要想完全做到把语法和语义完全分开，对于文法是不可想像的。这种企图不论在理论上还是在实践上都是不可行的是危险的。这是语言本身的性质所决定的。语言本身就是语法和语义的有机统一，是形式与内容的有机统一。正如第2节所述，语义要依靠语法结构来激活而传输意义；语法结构则要靠其构成成分的语义来建立自己的合理性。我们认为在现阶段我们应着重语义的研究。一方面要挖掘可以利用的语义属性，语义关系，一方面要使机译系统的算法、规则描述以及为此而设计的问题描述语言都能适应语法和语义的结合。

四、信息与算法的关系

迄今为止，现有的各种机译系统中，其词典的信息，以及词典的结构都是依赖于算法的。为语言加工建立的语言模型及相应的算法决定了词典应提供什么信息和应采取什么样的结构，这是正常的做法。但是众所周知，建立一个面向机器的词典是一个工程，尤其对于实用的机译系统而言则更是如此。这样，不同的系统就需要不同的词典。这显然存在一个大量重复劳动的问题。我们应该努力克服这种缺点。

我们认为一个可取的做法是建立一个基本词汇信息数据库，取代现存的依附于各自系统的多类多样的词典。这样的基本词汇信息数据库应具备如下特性：有较好的通用性、适应性和扩充性。当不同的机译系统，不同的算法利用它时，可以方便地根据各自不同的需要来提取该数据库的信息，并可以加以重组、扩充或修订。虽然这种重组、扩充或修订是不可避免的，也是要付出劳动的，但是这样做与那种一个系统自己完全地编制一部词典的做法相比较，还是要经济得多。

建立一个词汇信息库，并使之具有较好的通用性、适应性和扩充性，当然是不容易的。这里不仅仅存在着一个工程问题，更主要的是它的质量。为此，我们认为必须做到如下两点：第一，善于面对全语言最大限度地挖掘可资利用的信息；第二，科学地、独立地安排信息的存储。这里应特别强调的是，面对全语言是保证信息的客观性、科学性的关键。面对全语言就是不是为语言中的某一专门领域所限定。专门领域有关的信息可以用专门领域的词典或词汇信息数据库来反映和描述。基本词汇信息数据库应该是面对全语言的。

The relations between Lexicon and Grammar in Machine Translation

(Abstract)

Don Zhendong

(China Software Technique Corp.)

The lexicon and the grammar developed for machine translation(MT) are the core of the MT system, and the critical problem in MT research and development. Much consideration has been given to the relations between the two. This paper will discuss the relations in four aspects:

<1>individuality and generality;

<2>static and dynamic;

<3>syntax and semantics;

<4>information and algorithm.

The discussion of these aspects will cover the objective existence of the relations and the subjective treatment of them as well.