ICENT 汉英机译系统中的语义模型*)

The Semantic Model of ICENT System

齐 璇 陈火旺

(国防科大计算机学院 长沙410073)

Abstract Semantic analyzing is very important in MT system. The usage of semantic knowledge can help to disambiguation. Now semantic knowledge is always used to sense disambiguation. This paper presents a semantic hierarchy which contains semantic primitive, sense and semantic chunk. Then the paper gives the formal definitions of each semantic unit and semantic rule which reflect relations of semantic units. Preliminary experiment was done and the results showed that the usage of the semantic model could increase the accuracy of the system.

Keywords Semantic rule, Well-formed semantic link, Sense, Semantic chunk, How-net

1. 引言

研究汉英翻译技术是让中国走向世界、让世界了解中国的有力支持。目前,汉英翻译软件很少,且翻译质量不尽如人意,究其原因,一个很重要的方面在于汉语缺乏形态特征,仅仅通过句法知识来分析汉语是不够的,会遇到一些难以解决的问题。我们以863测试集测试了环宇通、译星99和我们已开发的 ICENT 汉英翻译系统[1],发现存在几个共同的问题是靠句法知识很难解决,如多义词义项选择问题、施事与受事问题、主动词的确定问题、空位问题等。

针对汉语句法约束较弱而语义约束较强的特点,有必要在汉英翻译中引入语义分析。这样做至少有三个好处:(1)有助于多义词的义项选择;(2)有助于句法结构的消歧;(3)有助于从源语到目标语的转换。目前计算语言学者对自然语言的语义分析大多集中在词义消歧方面,如文[2]、[3],所用的大多为统计的方法。对于如何在机译系统中建立系统的语义模型并利用语义知识解决翻译过程中的各种歧义问题还研究得不是很多。Marsal Gavaldà 在文[4]中提出一个增长的语义文法,但它是受限的,要针对领域提出模型并构建核心文法。苑春法在文[5]中利用语义知识进行句法结构消歧,他针对具体问题具体分析,虽简单但没有一个一般的语义消歧模型,且其方法对连动句子的分析效果有待确定。

ICENT 系统是一个基于中间语言的汉英机译系统。ICENT 中的汉语分析采取语义制导的语法分析方法,即只有通过语义检验的句法成分才能进行产生式规约,规约后的句法成分同时得到其语义描述结构。为进行语义分析,我们提出了 ICENT 的语义模型,建立从语义基元、义项到语义块的层级语义单位体系。给出语义规则的形式化描述,并利用语义规则建立各语义单位之间的关系。

2. ICENT 语义模型的理论基础

ICENT 语义模型的建立是以知网为基础的,对语义块的表达借鉴并扩展了格语法,采用框架形式,在框架中给出各语义单位承担的语义角色。

2.1 格语法(Case Grammar)

格语法是由美国语言学家菲尔默(Fillmore)在1968年发表的《"格"辨》[6](The Case for Case)一文中提出的一种语法分析模式。格语法承认语义在句法中的主导作用,认为动词在句中起中心作用,参与动作的各个体称为"语义格",构成格框架。

格语法最大的特点是承认语义在句法中的主导作用,由格语法分析可以得到句子的深层语义结构,给出各成分的语义角色。这样,有着不同表层形式而含义相同的句子有同样的格框架。格语法适应于汉语的分析,特别对于确定正确的句法结构有很大帮助。但是格语法在汉语分析中存在以下几个缺点:

·*无法解决汉语的连动和兼语句式* 格语法认为动词在句中起中心作用,那么分析句子时首先要确定句子的核心。汉语缺乏形态特征,作为核心的主动词通常也缺乏形态特征,句中同时存在多个动词的现象十分普遍,如何在有多个动词的连动式和兼语式中找出句子的核心是汉语信息处理的一个很难的问题,也是格语法无法解决的问题。

·*短语内部各成分间关系无法确定* 格语法提出的各种格关系都是体词性短语和动词之间的语义关系,对于体词性短语内部、谓词性短语内部和其他短语内部各成分间关系的确定没有给出。

ICENT 的语义模型利用了语义格的思想,对汉语语法中各种形式的短语确定其语义框架形式,在框架中对短语各组成成分给出其担当的语义角色。

2.2 知网(how-net)

知网[7]是由董振东老师提出并建立的一个以汉语和英语的词语所代表的概念为描述对象,以解释概念与概念之间以及概念所具有的属性之间的关系为基本内容的常识知识库。

知网借鉴了概念从属理论的原语概念,提出了1500多个义原,用来描述概念、概念之间的关系以及属性与属性之间的关系。义原分为实体、事件、属性、属性值、数量、数量值、句法特征、次要特征和动态角色等类别。特别地,实体义原、事件义原、属性值义原和数量值义原具有层次性,构成树型分类结构,因此存在上下位的属性继承关系。概念由义原描述,处于第一位的为概念的主要语义特征,这一特征因其层次性而具

*)本课题受国家"863"基金资助,项目编号863-306-ZT03-06-1。齐 璇 博士研究生,研究方向:机器翻译、计算语言学。陈火旺 中国工程院院士。

有上位的继承性。知网对每一个事件义原给出了一个角色框架，列出当某一类事件发生时框架中的必要绝对角色。

知网是针对中文信息处理提出的汉语词汇的语义知识资源，它对概念的划分不是简单的树型分类，而是在主要语义特征之外，辅以次要语义特征描述，从而使得概念间发生联系构成关系网。知网描述了16种关系，比较重要的有上下位关系、部件-整体关系、属性-宿主关系、材料-成品关系、施事/经验者/关系主体-事件关系、受事/内容/领属物等-事件关系、工具-事件关系、场所-事件关系、时间-事件关系、事件-角色关系等。这是知网比《同义词词林》优越的地方，而且知网的双语知识辞典选词较新，符合时代特点。因此，我们选择知网作为I-CENT语义模型的基础。

知网是针对汉语提出的，适合用来进行汉语的语义分析，但知网也有其不完善的地方：

·知网强调了概念之间的关系，但比概念更大的语义单位如复合概念则没有提及。复合概念如何定义、如何表示，没有给出。

·知网给出事物类概念和事件类概念之间的关系，但对事件类概念之间的关系没有给出，不利于解决汉语主动词确定问题。

·知网已有的概念之间的关系描述还远远不满足语义分析的要求，急需补充。

·知网中同时存在两套语义角色符号系统，导致描述的不一致。

针对知网的不完善，并考虑汉英机译汉语分析的要求，我们提出ICENT的语义模型，在ICENT中引入汉语的语义分析。

3. ICENT语义模型

针对知网的不足，同时借鉴格语法的框架描述法，我们建立了ICENT的语义模型。相对于从词素到词、短语、句子、段落直至篇章的句法单位层级体系，语义基元、义项和语义块构成了ICENT语义单位层级体系。语义单位定义如下：

SemUnit::=semantic primitive|sense|semantic chunk

Sense::=name[note]

Semantic chunk::=name(〈HEAD SemUnit〉
〈SEMROLE SemUnit〉*)

name$\in\Psi_{MSP}$，$\Psi_{MSP}\subset\Psi_{SP}$，$\Psi_{MSP}$为主要语义特征原语集合，$\Psi_{SP}$为语义原语集合。

note$\subset\Psi_{SP}$

SEMROLE$\in\Psi_{SR}$，Ψ_{SR}为语义角色集合。

语义原语(semantic primitive)是ICENT语义系统中最基本的语义单位，是构成概念的元素。义项(sense)是语义学术语，为字典词典中同一个条目内按意义列举的项目[8]，在I-CENT中义项表达概念。语义块(semantic chunk)描写复合概念的语义，语义块可大可小，在句法上对应于短语以上的句法单位。

3.1 语义原语和义项

ICENT语义层级体系中所采用的语义原语和义项描述以知网为基础，在实践中有所增减。

ICENT中的语义原语虽是构成概念的语义单位，但不是最基本单位，不等同于语义学里的义素。义项表达概念，包括义项名(name)和义项体(note)两部分。义项名为概念的主要语义特征，用主要语义原语集合中的元素描述；义项体为概念的次要语义特征，用语义原语集合的子集描述。

3.2 语义块

语义块表达复合概念的语义，体现构成复合概念的各概念之间的语义关系。框架结构表示法可以方便地表示语义关系[9,10]，因此，ICENT中语义块采用框架形式表述。框架包括框架名(name)和框架体。框架名描述义块的主要语义特征，用主要语义原语集合中元素描述。这样，语义块框架因其主要语义特征具有属性的向上继承性。框架体由槽值对组成，槽名为构成复合概念的语义单位所承担的语义角色(SEMROLE)，槽值为构成复合概念的各语义单位(SemUnit)，这样由语义角色表达出语义块中各成分的语义关系。语义块结构是一个可嵌套的框架结构，这一特点有利于在句法规约的同时进行语义规约，生成不同层次句法结构的语义结构。

例：被子叠得整整齐齐。其语义结构为：

fold 摺叠|(〈HEAD Sense(叠)〉
〈PATIENT Sense(被子)〉
〈PATIENTATTRIBUTE Sense(整整齐齐)〉)

其中，“被子”的语义角色为“叠”的受事，“整整齐齐”的语义角色为受事的属性。

如前所述，知网中只对义项之间的关系进行了描述，而且这种描述是十分简单和不完备的。较小的语义单位如何组成更大的语义单位、表达更复杂的概念，得到的更大语义单位具有怎样的结构、涉及怎样的关系，知网都不能很好地回答。在ICENT中，这些语义单位间关系的语义描述知识以语义规则形式表现。文[11]对现代汉语计算语言模型中语言单位的频度-频级关系进行了统计分析，指出统计的方法仅对极少数高频的词汇是有意义的，我们无法仅仅通过扩大统计语料库的方法来获得绝大多数单词的上下文搭配信息。因此，我们采用规则的方法描述语义知识，并在学习的过程中辅以统计的方法。

ICENT保留了知网词典中义项间关系的描述，排除了原有的主要语义特征间关系的描述，重新获取语义规则，描述义项间、义项的主要语义特征间、义项与义块间、义块间等各层次语义单位间的语义关系。

语义规则为语义知识的形式化描述，包括合式语义链和语义块模板两部分。合式语义链为语义主要特征原语序列，每一个语义主要特征附着一个限制条件。语义块模板为可由合式语义链生成的语义块的抽象，其语义角色的值不是具体的语义单位，而是合式语义链中相应元素的位置。语义规则的定义如下：

Rule::=Well-formed Semantic Link→Semantic Chunk Template

Well-formed Semantic Link::=$name_0/c_0\cdots name_i/c_i\cdots name_n/c_n$

$name_i\in\Psi_{MSP}$

c_i为限制条件。

不同类型的合式语义链其相互间的修饰约束关系不一样，因而可生成不同结构的语义块。由于ICENT的句法产生式规则待规约项多为二元式，因而语义规则的合式语义链也多为两项。我们以n=2为例，描述语义块的不同结构。设α、β、γ为主要语义特征，$\alpha\in\Psi_{MSP}$，$\beta\in\Psi_{MSP}$，$\gamma\in\Psi_{MSP}$，α/0、β/0为义项的主要语义特征，α/1、β/1为语义块的主要语义特征。当α/1、β/1为语义块的主要语义特征时，设α/1对应的语义块有结构α(〈HEAD h〉〈SEMROLE, r_i〉)，β/1对应的语义块有结构β(〈HEAD h′〉〈SEMROLE, r_i'〉)，则规则形式有：

(1)α+β→β(⟨HEAD 1⟩⟨SEMROLE 0⟩)
(2)α+β→α(⟨HEAD 0⟩⟨SEMROLE 1⟩)
(3)α+β→γ(⟨HEAD NULL⟩⟨SEMROLE1 0⟩⟨SEMROLE2 1⟩),γ!=α && γ!=β
(4)α/0+β→NULL(⟨HEAD NULL⟩⟨SEMROLE 1⟩)
(5)α+β/1→β(⟨HEAD 1.h'⟩⟨SEMROLE,1.r_i'⟩⟨SEMROLE 0⟩)

规则的这五种形式使得句法结构[12]与其语义结构之间有较好的对应关系。第一类型规则对应于偏正结构，α对应的语义单位修饰或限制β对应的语义单位，如例(1)，有实例"小/山羊"；第二类规则对应于述补结构、述宾结构，β对应的语义单位修饰、说明α对应的语义单位，如例(2)、例(3)，有实例"落/下"、"办/工厂"；第三类型规则对应于联合结构、具有体词性谓语的主谓结构，生成新的语义块具有新的语义主要角色γ，α、β对应的语义单位说明γ的内容，如例(4)、例(5)，有实例"太阳/月亮"、"今天/星期一"；第四类规则对应于介词结构，如例(6)，有实例"把/玉米"；第五类规则对应于具有谓词性谓语的主谓结构，α对应的语义单位修饰、限定语义块β的中心成分h，如例(7)，有实例"我/不吃骨头"。

例(1) size|尺寸+livestock|牲畜→livestock|牲畜(⟨HEAD 1⟩⟨MODIFIER 0⟩)

例(2) fall|掉下+Vdirection|动趋→fall|掉下(⟨HEAD 0⟩⟨TREND 1⟩)

例(3) establish|建立+InstitutePlace|场所→establish|建立
(⟨HEAD 0⟩⟨PATIENTPRODUCT 1⟩)

例(4) celestial|天体+celestial|天体→celestial|天体
(⟨HEAD NULL⟩⟨CONT 0⟩⟨CONT 1⟩)

例(5) time|时间+time|时间→be|是(⟨HEAD NULL⟩⟨RELEVENT 0⟩⟨ISA 1⟩)

例(6) patient+crop|庄稼→NULL(⟨HEAD NULL⟩⟨PATIENT 1⟩)

例(7) firstPerson|我+eat|吃/1→eat|吃(⟨HEAD 1⟩⟨AGENT 0⟩)

由于合式语义链有主要语义特征加条件组成，而主要语义特征在其层次分类中有向上继承性，因此，可以对语义规则进行抽象。如何组织规则库，将是我们今后研究的一个问题。

4 初步试验

我们对小学前四册语文课本进行了翻译，一共1456句。首先，我们用已开发的ICENT汉英机译系统(不含语义知识)进行翻译，从中找出315个有明显错误的句子。在这315个句子当中，13%的句子主要由于义项选择错误而导致句子错误；62%的句子主要由于句法结构分析错误而导致句子错误，特别是由于主动词的确定错误而得到错误的句法结构；25%的句子由于其他原因导致错误，包括特殊句型、逻辑关系及生成问题等等。

其次，我们在ICENT中加入语义知识，并对这315个句子重新进行翻译。结果显示，句子读懂率明显上升，至少达到57%。在剩下的错误句子中，主要的错误原因有两个：(1)特殊句型的分析生成错误；(2)长句子中成分间语义关联距离过长，造成结构分析错误。

可见，语义知识对提高译文质量有很大帮助。针对错误原因，今后重点要研究针对特殊句型的语义规则表达，对某些特殊问题，要细化规则。

总结 在机译系统中，语义知识的使用对消歧和提高翻译质量有明显的帮助。本文介绍了ICENT汉英机译系统中的汉语语义知识描述模型，这一模型系统描述了语义单位的各个层级，并刻划了语义单位之间的关系。采用这一模型，我们进行了初步试验，对小学前四册语文课本中的句子进行了翻译。结果显示，这一模型的使用在义项消歧、句法结构消歧和译文生成方面都体现出语义知识对翻译的帮助。

目前我们开发了一个辅助工具，通过人机交互来获取语义规则。今后研究的重点是如何从大规模语料库中自动获取语义规则。规则和统计的方法相结合，尽量自动的前提下通过人机交互提高准确性，成为今后获取语义规则的指导原则。

致谢：感谢董振东老师提供的知网资源。

参考文献

1 齐璇，马红妹，陈火旺. 汉语的语义分析研究. 计算机工程与科学，已录用，待发表
2 Yamaguchi M，et al. Combination of an Automatic and an Interactive Disambiguation Method. COLING-ACL'98，Canada，1998. 1423～1427
3 Dini L，Tomaso V D，Segond F. Error Driven Word Sense Disambiguation. COLING-ACL'98，Canada，1998. 320～324
4 Gavaldà M，Waibel A. Growing Semantic Grammars. COLING-ACL'98，Canada，1998. 451～456
5 苑春法，黄锦辉，李文捷. 基于语义知识的汉语句法结构排歧. 中文信息学报，13(1)
6 Fillmore(1968)，胡明扬译.《"格"辨》，语言学译丛第二辑. 中国社会科学出版社，1980
7 董振东. 知网介绍. http://keenage.com
8 骈宇骞，王铁柱. 语言文字词典. 北京：学苑出版社，1999
9 Zechner K. Automatic Construction of Frame Representations for Spontaneous Speech in Unrestricted Domains. COLING-ACL'98，Canada，1998. 1448～1452
10 Baker C F，Fillmore C J，Lowe J B. The Berkeley FrameNet Project. COLING-ACL'98，Canada，1998. 86～90
11 关毅，王晓龙，张凯. 现代汉语计算语言模型中语言单位的频度-频级关系. 中文信息学报，13(2)
12 朱德熙. 语法讲义. 北京：商务印书馆，1999
13 Lappin S. The Handbook of Contemporary semantic Theory. Blackwell Publishers，1996