词的语义分类在汉英机器翻译中所起的作用以及难以处理的问题[①]

詹卫东*　刘群**

（*） 北京大学中文系　100871
（**） 中国科学院计算所二室　100080

摘 要：本文较为全面地探讨了词的语义分类在一个汉英机器翻译系统中能够发挥的作用以及难以处理的问题，指出了词语语义属性描述的重要性和实用价值，同时也举例说明了这种方法客观上存在的不足，以期澄清以往对语义问题的一些模糊认识，能为自然语言处理中的汉语语义研究提供一些参考。

Using word semantic knowledge in a Chinese-English Machine Translation system

Zhan Weidong*　Liu Qun**

（*）Dept. of Chinese Language and Literature, Peking University　100871
（**）Institute of Computing Technology,Chinese Academy of Science　100080

Abstract：This paper discusses what roles the semantic classfication information on words plays in a Chinese-English Machine Translation system.The author prensents examples of how to take advantage of semantic classfication information on words for disambiguation in parsing and generation, also points out what difficult problems can not be solved even if semantic information of words is used in the CEMTs.

§1

词的语义分类是标明一个词的语义属性的常用手段之一。描述一个词的语义属性一般包括两个方面：一是这个词自身的语义性质（如该词所属语义类等）；二是这个词的共现成分（co-occurrence）的语义性质（如动词的配价成分所属语义类等）。这两方面属性的描述都可以词的语义分类框架为基础。有关词的语义分类方面的问题，学者们已有过一些探索。问题主要归结为语义分类难以全面，同时分类又难免有交叉，分类的标准不易统一等等[1]，侧重在理论层面进行讨论。

本文试图通过分析在基于规则的（rule-based）汉英机器翻译系统中，词的语义分类信息能起到哪些作用，来为认识上述问题提供一些参考，侧重从实践应用的角度对语义问题作一些探索。我们认为，无论是确立词的语义分类标准，还是判断一个语义分类体系的优劣，首先都应该搞清楚，一个词语语义分类体系，是干什么用的；在一项具体的实践应用中，它确确实实能起多大作用，在哪些方面又是力所不能及的。从这个角度出发，才可能

[①] 本文的研究工作受国家“863”项目（编号 863-306-03-06-2）基金资助。北大中文系陆俭明教授和北大计算语言所俞士汶教授都对本文的写作给予了很多指导，特此致谢。学友常宝宝、刘颖、王斌也对文中内容提出过宝贵意见，在此一并表示感谢。

对词语语义分类的研究真正起到推进作用。

§ 2

根据我们的考察，词的语义信息在汉语分析的各个层面，包括多义词词义判断、短语结构层次和结构关系判定、以及成分之间语义关系的确定等等，都能起到一定作用。本节我们结合一个具体的汉英机器翻译系统[2]来说明词的语义分类信息能够起的作用。

2.1　首先我们来看如何利用词的语义信息确定多义词的意义。

对计算机处理自然语言来说，多义词的概念应该拓宽。可以宽泛地定为只要同形不同义就是多义词。包括了语言学家一般说的多义词，如“开”（“开飞机、开门、开票、开会……”等等）；还包括通常所谓的同形词，如名词的“会”（“一个会”）和动词的“会”（“会打球”）；通常的多音词，如念去声的“好、种”（“好客、种地”）跟念上声的“好、种”（“好人、良种”） 。就计算机而言，这些都需要判定在具体语境中到底属哪一个意思。可以笼统地都称为多义词。

下面以多义动词的义项判定为例说明词语语义分类信息的作用。

（1）　妈妈想女儿

（2）　爸爸想主意

上两例中的谓语动词“想”的意思不一样。判断依据是，例（1）中“想”所带宾语是“女儿”，语义类属“人”，这时“想”是“思念、怀念”义。例（2）中“想”带“主意”做宾语，语义类属“事理”。这时“想”是“思考”义。一般而言，“想”带指人宾语时，就可判定为“思念”义；带抽象的“事理”类宾语时，就可判定为“思考”义。属这类宾语的名词还有“办法、法子、招儿、计策、思路…”等等。正确判断出“想”在上两例中的不同意思后，就能相应地找准对译的英语词了。例（1）中的“想”翻成“miss”；例（2）中的“想”则翻成“think”。

再比如，名词可跟方位词组合形成偏正结构，其中方位词在不同语境下可能有不同意思。请看例句：　（3）　学生们在教室后种了很多树

（4）　学生们在二战后又回到了学校开始正常的学习

上面两句中的“教室后”和“二战后”都是“名词+方位词”的格式，但其中“后”表示不同的意思。在例（3）中是真正表示方位的前后；在例（4）中并不表示方位，而是表示时间的先后。判断的依据是：“后”前的名词属于不同的语义类，决定了“后”的意思。如果词语语义类为具体有形的“具体物”，包括“建筑物”（如“教室、楼房”等等）、“用具”（如“桌子、床”等等）这些名词，它们出现在“后”前，就可判定“后”表示方位。如果是无形的，所指对象跟一段时间相联系的“抽象物”类名词（如“战争、台风”等等）、或者是表示“时点”的“时间”类名词（如“民国、宋朝、仲夏”等），出现在“后”前，就可确定这时“后”表示时间。这样，例（3）中“后”就译为“behind”；例（4）中则译为“after”。

2.2　接下来再看词的语义分类信息在确定短语结构层次时起的作用。

我们以“VP+NP+的+NP”组合格式为例说明。这是汉语句法结构中存在的一个固有歧义结构式。结构层次可以是：　a. [[VP+NP+的]+NP]，整个结构是 NP 短语；

或　b. [VP+[NP+的+NP]]，整个结构是 VP 短语。

但这个歧义格式投射（mapping）出来的实例，并非个个都是歧义的。在实际语料中，这个格式的句子有很多对人而言是单义的。如果要让计算机来判断，就必需让它掌握在何种条件下这个格式的实例是单义的。请看下面例句：

（5）　安装网络系统的人碰到了问题

（6） 安装康柏公司的网络系统碰到了问题

上两句中“安装网络系统的人”和“安装康柏公司的网络系统”都是“VP+NP+的+NP”格式的实例。很显然分别都没有歧义。

这里例（1）应按 a 方式切分。判断的依据是：“安装”是二价动词。要求其施事成分所属语义类为“人”或“集体”，受事的语义类为“设备”。如果按 b 方式切分，就形成“安装”带“人”做宾语，“人”成了“安装”的受事，这与它的受事语义类要求不符，分析失败。按 a 方式切分，“安装”带“网络系统”做宾语，“网络系统”的语义类属于“设备”，符合要求，分析成功。

例（2）则必须按 b 方式切分。理由同上。“康柏公司”语义属“集体”类。“安装”不能带“康柏公司”做受事宾语。

很明显，在这里名词的语义分类以及有价动词对其配价名词的语义选择要求，对得到正确的分析结果起了关键作用。实际上，在分析得到正确的结构层次的同时，结构内部的句法关系和成分之间的语义关系也随之得到了确定。例（1）中，“安装”跟“网络系统”之间是述宾结构关系，语义关系是“动作—受事”，跟“人”则构成“动作—施事”的语义关系。这两例对译的英语是[3]：

（7） Man who installs network system has run up against problem.

（8） Installing network system of Compaq Corp. has run up against problem.

2.3 下面进一步来看词的语义分类信息在确定短语结构关系时起的作用。

我们以“NP+NP”格式为例说明，这一格式内部可以有偏正、并列、主谓等结构关系。其中偏正和并列两种关系是强势关系，出现频度高；构成主谓关系相对是弱势关系，出现频度低，如“鲁迅浙江人、小王急性子”等，我们这里撇开不论。当两个名词先后排列时，计算机判断其内部结构关系是属偏正还是并列，主要的依据就是这两个名词的语义类信息。请看例句：

（9） 我们被外行律师害苦了

例句中“外行律师”是两个名词连用。这里只能是偏正关系，而不能构成并列关系。判断依据是：“外行”跟“律师”虽然都指人，但属不同的语义类。“外行”属“身份”类，“律师”属“职业”类。跟“外行”同类的还有如“庸人、高个儿、急性子、叛徒…”等等；跟“律师”同类的还有如“会计、教师、医生…”等等。我们对两个 NP 不依靠连接词直接构成并列关系（即无标记形式 unmarked-form）采取强限制判断，即认为只有两个名词属同一语义类，才有资格直接形成并列关系，否则只能形成偏正关系。这样例中“外行律师”就只能形成偏正关系，译成“lay lawyer”了。如果例句中的“外行”换成“医生”，跟“律师”同属“职业”语义类，“医生律师”就形成并列关系。需译成“doctor and lawyer”。对两个动词连用、以及形容词连用，如果要构成并列结构关系，我们也可用类似的方法判断。这里就不多举例了。

2.4 最后再来看看词的语义分类信息在判断成分之间语义关系时起的作用。

请看例句：（10） 学生们全都吃食堂

（11） 他习惯于写毛笔

这两例中的述宾结构“吃食堂”和“写毛笔”，动词跟宾语都不是“动作—受事”语义关系。前者“食堂”是“吃”的处所，后者“毛笔”是“写”的工具。译成英语，在述语动词和宾语之间需添上介词。例（10）“吃食堂”译为“eat in dining-room”，例（11）“写毛笔”译为“write with writing brush”。判断依据是宾语名词的语义类，以及动词对配价成分的语义要求。“食堂”属“建筑物”类名词，“吃”可以带这类名词充任处所宾语，译成英文时补出介词“in”。“毛笔”属“用具”类名词，“写”可带这类名词充任工具宾语，译成英文时补出介词“with”。

再如：　（12） 买书的同学可以参加

（13） 安装在桌上的灯亮了

这面例中“买书的同学”和“安装在桌上的灯”都是“VP+的+NP”格式的短语。前者译成英语时，用定语从句形式表达，要添加关系代词“who”，译为“student who buy book”；后者 VP 译成英语不用从句形式而用过去分词形式表达，译为“lamp installed on table”。判断依据是，“VP+的+NP”格式中 NP 的语义类是“人”，并且是 VP 的施事时，就用从句形式表达，从句引导词为“who”；NP 语义类不属“人”，并且 NP 是 VP 的受事时，就用动词的过去分词形式在名词后做定语表达。

词的语义分类信息还可以帮助判断主谓语间的主动和被动关系，在转换生成英语是对确定语态起一定作用。例如：

（14） 这个问题解决了

（15） 蒸汽机是瓦特发明的

例（14）中主语“这个问题”语义类属“事情”，是动词“解决”的受事。译成英文时就需采用被动语态；例（15）中“蒸汽机”语义类属“设备”。这时全句谓语动词是无价动词“是”，“蒸汽机”就跟谓语中的次级动词“发明”发生配价关系，是“发明”的受事。译成英文也需采用被动语态。译文分别为：

（16） This problem has been solved.

（17） Steam engine was invented by Watt.

在这里，判断依据同样是主语名词的语义分类信息，以及谓语有价动词对配价成分的语义选择要求。二者结合起来就可确定主语跟谓语成分之间的语义关系。一般说来，像上面这样分析结果是所谓受事主语句的，转换生成英文时需采用被动语态来表达。

§3

尽管如上述分析所显示的，词的语义分类信息在汉英机器翻译中的确能起不少作用，但有时候词的语义很难分得清楚，语言使用中还牵涉到许多其他因素，仅有语义信息有时仍然会陷入鞭长莫及的困境。实际上，词的语义分类信息在上一节中能发挥作用的场合，也同时都存在着无能为力的另一面。

3.1　对短语结构层次的判定，词的语义分类信息有时就不能有效地发挥作用。

仍以“VP+NP+的+VP”格式为例。这里包括两种困难情形：一是 VP 的施事和受事成分所属语义类可能相同；二是 VP 的受事成分所属语义类不易概括。前一种情形例如：

（18）　批评这个学生的老师是不对的

例句中“批评这个学生的老师”是“VP+NP+的+VP”格式的实例。“批评”的施事和受事所属语义类都是“人”。因此，例（18）即使对人而言，孤立地看也是有歧义的，可以译为“teacher who criticize this student”，或“criticizing teacher of this student”[4]。这里仅靠词的语义分类信息无法做出区分。再看后一种情形的例子。

（19）　改变主意的原因要交代清楚

例句中“改变主意的原因”也是“VP+NP+的+VP”格式的实例。这句对人而言无歧义，但由于“改变”的施事和受事成分所属语义类不易概括，同时“主意”和“原因”都是抽象名词，语义类不易区别开，因此计算机也把例（19）分析出 a 和 b 两种切分结果来。

象例（18）那样的情况，对人也有歧义，目前看来难以处理。可以暂时放一放。而象例（19）这样的情况，对人无歧义，但由于我们对词语的语义搭配性质认识不够，也就是知其然而不知其所以然，也同样难以处理好。这种情况是目前我们做语义分类工作急需解决的难点问题。

3.2　对短语结构关系的分析，词的语义分类信息也有难起作用的时候。

还是以名词跟名词连用的格式为例。请看例句：

（20） 这孩子巧克力饼干吃了一大堆

（21） 公司技术力量雄厚

（22） 中国社会科学杂志太少

（23） 电影事业人才很少

（24） 文化艺术水平不高

上面例（20）是两项名词连用，其他几例都是两项以上的多项名词连用。这时要分析其中的结构关系（也牵扯到结构层次问题），人可以凭语感判断，但人在这方面的语感很难概括清楚教给计算机掌握。例（20）“巧克力”跟“饼干”可以构成偏正关系，也可以构成并列关系，对人而言也有歧义，需依靠更大的语境才能确定。我们也可暂时不处理这类情形的问题；例（21）是“公司”修饰“技术力量”；例（22）是“中国”修饰“社会科学杂志”，“社会科学”修饰“杂志”，“社会”修饰“科学”，属偏正结构套叠[5]的情形；例（23）是“电影事业”修饰“人才”，“电影”修饰“事业”，与例（22）同；例（24）是“文化艺术”修饰“水平”，“文化”跟“艺术”并列。要把这里边谁跟谁是并列关系，谁又是修饰谁都分析清楚了，是非常困难的。症结就在于对抽象名词的语义分类难以把握好。对名词短语自相组合的句法语义规律也难以描述清楚。

3.3　再到确定多义词义项以及判定句子成分之间语义关系的场合，词的语义分类信息更是有可能显得捉襟见肘。仍以多义动词判定义项为例。

（25）　我们送老王

（26）　他看着孩子

例（25）中，“送”是多义词。在这里既可以理解为“赠送”。指“我们送礼物给老王”，译为“prensent”。后面宾语“老王”是“送”的与事；也可理解为“陪着离去的人一起走”。指“我们来送走老王”，译为“see Lao Wang off”。“老王”是“送”的受事。两种情况下，“送”的意思不同，述语跟宾语成分的语义关系也不一样。对人来说，这个例句也是有歧义的。

例（26）中，“看”可以理解为“注视”（这时读去声），译为“look at”；也可理解为“看护、照看”，译为“look after”。跟例（25）一样，这句的歧义对人也是存在的。要确定这种情况下多义词的义项，光有词的语义分类信息是不够的。因为这种多义动词的不同义项允许相同语义类的名词充任其宾语形成不同配价关系。这时光有词的语义信息不可能准确判定多义词的意思。下面的例句又是另一种情形。

（27） 我先开灯，你再开箱子

（28） 老大开飞机，老二开汽车

（29） 哥哥看电影，弟弟看电视

例（27）中，“开灯”跟“开箱子”中的“开”在《动词用法词典》[6]中是归入同一个义项的，但翻译成英文不一样。前者是“turn on”；后者是“open”。这就要求“灯”跟“箱子”分属不同的语义类。但一般说来，“灯、箱子”都可属于“用具”，只不过“灯”是“电用具”，“箱子”不是。这样的差别语义分类要不要区别描写？由此引发出来的一般性问题是，词的语义分类要细到什么程度才比较好？

例（28）的问题跟例（27）类似，尽管“飞机”跟“汽车”都是“交通工具”（不同之处仅在于前者在天上飞，后者在地上跑），但“开飞机”跟“开汽车”译成英文不一样。前者“开”得译成“fly”；后者得译成“drive”。

例（29）不同于前两例，“看”的宾语“电影”、“电视”似乎没什么差别了，但译成英文仍不相同。“看电影”是“see a film”；“看电视”是“watch TV”。看来只能以

用法习惯不同来解释这种情况了。

对类似例（27）—（29）三例的情形，目前我们都不强求通过语义分类信息来进行翻译，而笼统地作为固定搭配来处理。

§4

总的说来，词的语义分类是一项艰巨的工程。应该看到，语义分类对句法分析的控制，比句法规则系统会出现更多的例外现象，存在很多难以照顾到的方面。这正如本文中举例讨论的，在机器翻译的多义词处理、短语结构层次和结构关系判断、语言成分间语义关系的确定等不同层面上，词的语义分类信息都恰恰是既有用武之地，又同时存在无能为力之处。换言之，就是在系统分析中引进了语义信息，并不像通常的在面对一个困难问题时引入一种新方法、新技术那样，可以使该类难题迎刃而解。事实上，在目前水平，利用词的语义分类信息，只能解决个例性质的问题，而不具有解决一类问题的普遍性意义。从这点上说，现有语义分析的理论和技术，其方法论价值就要大打折扣了。

有了上述认识，我们就不应该奢望基于语义的方法能够多么神通广大，务实的态度应该是通过不断努力，让语义信息在局部好用的场合能起到明显的效果。

由此我们认为，对词语进行语义分类，目前切实可行的做法应该是，面向特定的工程目标，先从概念出发建立一个比较粗的分类框架，刻化一定量的词；然后在系统调试中检验初级分类的效果，对层次分类体系的广度和深度进行调整，使分类不断朝合理有效的实用目标逼近。曾经有人说过，搞机器翻译写句法规则，是“三分写，七分调”。其实语义分类也可套上这顶帽子。即先花三分力气分个大概，还得花七分力气来做微调工作。站在用规则方法进行自然语言处理的立场，应特别强调着眼于句法分析来考虑语义分类。语义分类绝非仅仅是概念分类，一个语义分类体系也不会是对任何自然语言都普遍适用的。

附注：

[1] 参见张普（1991）《信息处理用现代汉语语义分析的理论与方法》，载《中文信息学报》1991 年第 3 期；孙宏林（1994）《信息处理用汉语语义词典的描述方法》，载《现代语言学・第三届全国语言学会议论文集》，语文出版社；陈群秀（1996）《信息处理用现代汉语语义分类体系的设计思想》，载《计算机时代的汉语和汉字研究》，清华大学出版社。

[2] 中科院计算所二室跟北大计算语言所自 1994 年开始合作开发汉英机器翻译系统。目前已具一定规模。机译系统的语言知识库中包括一个汉语词语语义分类体系。在机译词典中则根据该分类体系较为详细的记录了汉语名词、动词、形容词的语义类信息，以及有价名词、动词、形容词的配价信息。

[3] 译文目前还有英语词形变化，时态等细节问题没有仔细处理。质量还有待进一步提高。这里给出的结果在文中提及的方面都表现出较好的处理水平。其他方面不一定能照顾得到。本文例句的机译结果都可能存在这样的问题，不一一注明。

[4] 机译系统还可能译成“criticizing this student’s teacher”。汉语结构跟文中后一种译法一样。

[5] 有关句法结构套叠，可参见 陆俭明《汉语句法成分特有的套叠现象》，载《中国语文》1990 年第 2 期。

[6] 参见 孟琮 等《动词用法词典》，上海辞书出版社 1987 年版。