基于词组本位语法的语义模型[@]

詹 卫 东　　常宝宝　　俞士汶
北京大学中文系　　北京大学计算语言学研究所
北京，100871　　北京，100871
Email: zhanwd@mtgroup.ict.ac.cn　　Email: yusw@pku.edu.cn

(Received: 30 Jan 1998; Revised: 7 Jul 1998)

摘要: 本文在简略阐述我们对几种语义分析理论 —— 配价语法，格语法，论旨理论的认识的基础上，介绍我们研制的一个基于汉语词组本位语法的汉英机器翻译系统所采用的语义模型。对设置语义项目的指导原则和语义分类体系做了概括介绍，并且具体对四万余汉语实词进行了语义分类及语义搭配性质的描述。此外对在汉英机器翻译中引入语义处理机制所起到的效用也举例做了一些说明。

关键词：**语义模型　语义分类　词组本位语法　语义搭配　语义属性　机器翻译**

一、前言

在汉英机器翻译中，引入语义处理机制至少有两个显著作用：（1）有助于得到句子正确的句法结构；（2）使得翻译中的转换环节能够在语义层次上进行。前一条实际上就是能够利用语义信息排除一些歧义句法结构，提高源语分析的质量；后一条则是可以利用语义信息，在不同的翻译结果中挑选较为合适的译法，提高目标语生成的质量。而对于自然语言的语义分析，理论语言学界目前比较成系统的大致有配价语法（Valent Grammar）、格语法（Case Grammar）以及论旨理论（Theta-Theory）等体系。计算语言学界则限于计算机符号处理技术条件的限制，对语义问题的形式化处理与处理句法问题的方式实质上可以说是完全一样的。

这样，从事计算语言学工程实践的研究人员在设计实用的机器翻译系统时，就面临着一个棘手的问题，即只能利用有限的形式化手段来实现理论语言学界已经取得的有关自然语言语义问题的研究成果。从务实的态度出发，一个机器翻译系统在引入语义处理机制时，既要吸收语义研究的理论成果，也要考虑实际的可操作性。

本文介绍了我们目前研制的一个汉英机器翻译系统中的语义处理框架。这个框架就是在分析比较几种不同的语义理论体系之后，在已有的相对比较成熟的句法体系基础上建立起来的。希望能对机器翻译由句法表层向语义深层的发展起到有实质意义的促进作用。

二、配价语法、格语法、论旨理论的扼要辨析

配价语法、格语法、论旨理论是目前理论语言学界讨论颇多的三种分析自然语言语义问题的理论体系。但在这三个名目下，实际上各自内部都并不是一个纯粹的系统，三种理论各自在提出之初就隐含了或多或少的问题，而在发展过程中又有不同支派的分歧和不同阶段的变化，对一些问题存在着错综复杂的不同理解，即使理解相似表述方式上也还有这样那样的差别。本文主旨并不是深究三者异同的方方面面，而是从一些基本问题着眼，粗线条地对三种理论做扼要的异同辨析，从而为做实际的机器翻译系统设计语义处理模型提供一些理论参考。

概括言之，配价语法、格语法、论旨理论尽管直接目标相近，都是关注句内各成分间关联或者说是搭配的问题，但在立论基础（或者说是基本语言观）、对搭配性质的认识以及进行语义分析的具体方式三个方面存在显著差异。

[@] 本文的研究工作得到国家 863 计划的资助。本文所说的系统是指中国科学院计算所二室和北京大学计算语言学研究所目前正在联合研制的一个汉英机器翻译系统。

（1）立论基础

配价语法和作为它的理论基础框架的依存语法（Dependence Grammar），都该归在法国语言学家 Tesniere 的名下。依存语法主张以句子成分间的主从依存关系为纲来刻画句子结构。配价论则专注于语言中动词的价结构，以化学理论中所用的元素“化合价”概念来比附动词与名词的结合能力。直接体现了以成分间依存关系为基础描写句法结构的思想。

格语法由Fillmore在《The Case for Case》中提出之时，美国语言学界的主流仍是Chomsky的转换生成语法。如 Chomsky 一样，Fillmore 也致力于探寻纷繁复杂的语言现象背后带有普遍性的基础要素。但后者不同于前者，另辟蹊径，赋予传统语法的“形式格”以新意，强调对“语义格”的研究，力图以对动名间语义格关系的分析为语法基础理论，从而推演出一整套描写和解释句法语义现象的机制。

论旨理论是以 Chomsky 为代表的生成语法发展到管辖与约束理论（GB Theory）阶段提出的一个原则子系统。不言而喻，其立论基础需要参照整个生成语法理论来看。在不得不关注句法成分间语义关联的背后，仍是强大的模块化的句法分析系统在全面支撑。

（2）对搭配性质的认识

配价语法对搭配性质的认识主要反映为对动词的价性质的争论。配价语法区分动词的“逻辑价”、“句法价”、“语义价”。在三个层面上分析，得到的动词价性质可能一致，也可能不完全一样。取舍难有绝对标准，应视分析目的及不同语言对象而定。

格语法由早期转换生成语法的深层结构不够“深”而来，从注重句法形式转向专注语义格关系，对动名搭配的分析偏重所谓深层概念意义，强调各语言共同的语义格系统。

论旨理论借鉴了格语法的语义格系统，但秉承生成语法一贯的句法中心论传统，在看待动词的论旨属性亦即动名搭配关系上，融入了相当的句法信息，使得对句内成分搭配关系的分析在涉足语义之外，仍然带有浓厚的句法形式色彩。

（3）语义分析的具体方式

在具体进行语义分析时，配价语法需确定动词的价数和配价成分（也即动词支配的补足语）的性质。价数视补足语的多少而定。而补足语的种类和名目，不同学者有不同看法。但就析句的目的而言，都是在动词价性质确定的基础上推导句型系统，分析句内成分间的关系，描写句子结构。因而对补足语的不同处理并无本质差异。此外值得指出的是，补足语并不仅限于名词性成分。

格语法首先确定与具体语言无关的基于概念意义的格系统。在传统语法中充任各种主、宾语成分的名词在格语法中则以各种语义格身份出现，句中动词则被看作是由名词构成的格框架（环境）选择的结果。“简单句中各种不同的格可能出现的各种不同安排方式，就表达了格语法的句型概念。”

论旨理论在词库中记载动词 “论元属性”（相当于动词的价数），标记一个动词的“论旨属性”（相当于动词的补足语的性质），此外还包括论元的“范畴属性”和“句法功能属性”。动词的这些广义搭配信息可以统称为“论元结构”或“论旨网格”，结合生成语法的投射原则及管辖约束等句法处理机制，共同用来控制句法操作，解释语感和歧义问题。

三、立足汉语词组本位语法的语义处理模型概要

基于上述对三种流行的语义理论的认识，并参考汉语学界近年来对名词、形容词的配价问题所做的有益的拓展性研究，同时考虑实现时的客观条件，我们设计的汉英机器翻译系统的语义模型，在以描述动词的语义搭配为主的总架构下，也顾及形容词、名词的语义搭配情况，并且尽量简化语义概念，明确地将语义处理模块定位为句法分析的辅助部件。系统的主干是句法，采用朱德熙先生的词组本位语法体系为组织句法信息的理论指导，而且已有对五万多个汉语词语的语法特征做了详细描述的“现代汉语语法信息词典”作坚实的后盾，这样我们的语义处理框架可以认为是立足于词组本位语法的语义处理模型。

首先我们建立了一个汉语实词的语义分类体系，并对具体实词逐个进行了语义类属性标记。在此基础上，基本以词对类的方式，同时也允许以词对词的方式，描述动词与受其支配的名词、形容词与受其支配的名词、名词与受其支配的其它名词之间的语义搭配关系。

这样的语义处理模型是在利用丰富的句法信息处理表层形式后的辅助分析手段，因而跟格语法不同，不需要太多的深层语义格；同时句法信息跟语义信息尽量明确分工，也不象论

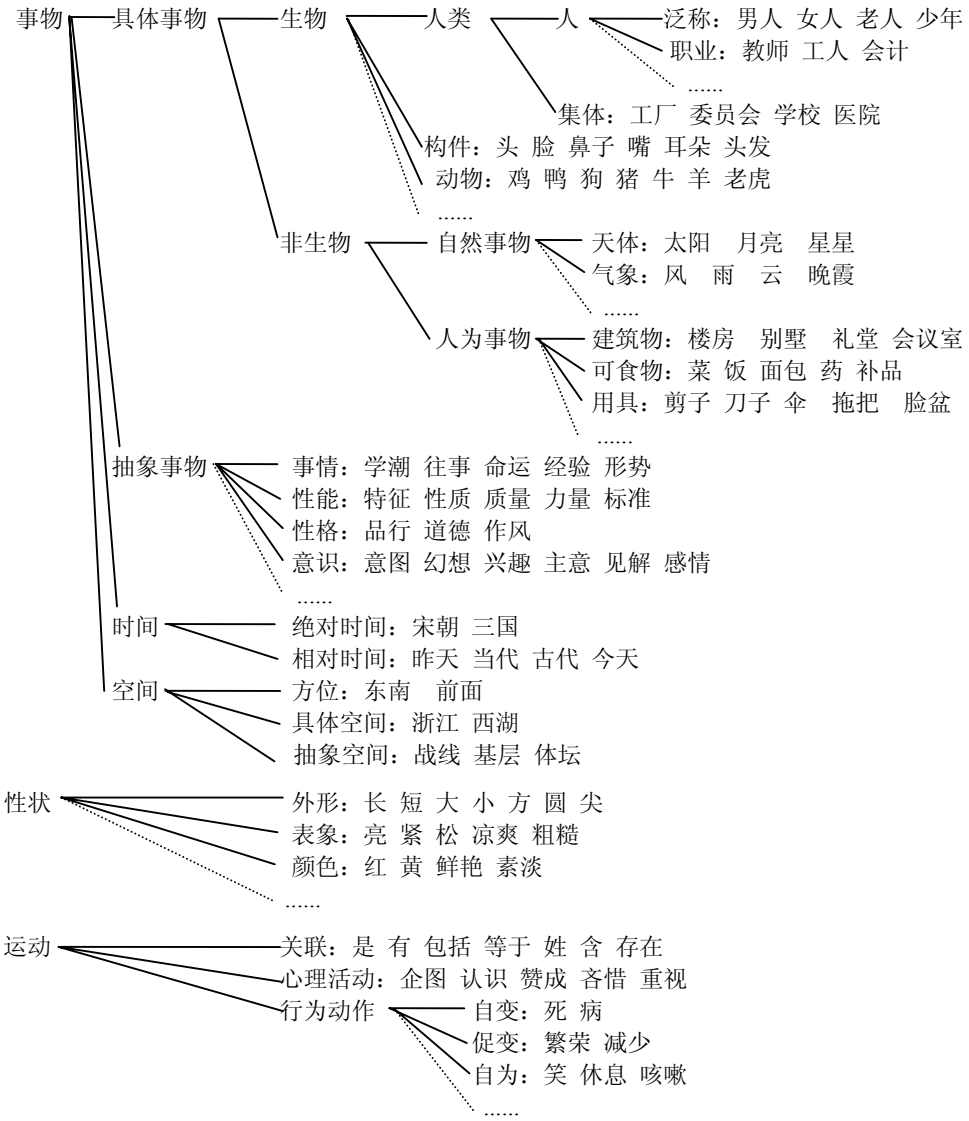
旨理论那样在进行语义处理时还掺进过多的更适于称为句法处理的东西；而跟配价语法相区别的则主要是我们的整套描述方式立足短语结构语法，配价理论则采用所谓依存句法树的模式。

下面我们先来看看语义分类体系。

（一） 语义分类体系

由于我们主要是以动词、形容词为中心描述二者跟名词之间的语义关系，因而对名词采用了相对较深较广的分类，对动词、形容词采用了相对较粗的分类。在名词内部，由于抽象事物的分类标准不易确定，而且实践经验也表明，对抽象事物做过细分类并不利于实际操作，因而我们在分类中只对具体事物进行了层次较多、较细的分类，而对抽象事物采取相对较粗的分类。

我们将汉语实词总的分为事物、性状与运动三个大类。其中事物类又分为具体事物、抽象事物、时间和空间等小类。下图列举了其中的一些细类并给出了例词。



（“：”后面的是例词；虚线表示大类中还有其他没列出的小类）

（二）语义属性描述

有了上述语义分类体系，就可以具体刻画词语的语义属性了。首先是对一个实词确定其语义类归属，接着就是描述词与词之间的语义搭配性质。这主要包括描述一个实词的配价数以及该实词对其语义配项成分的限制要求两个方面。本小节我们着重介绍实词的语义搭配性质。

（1） 配价数

配价数可用来描述一个实词跟其他实词（目前限于名词）间发生语义联系的能力。取值范围为 1、2、3 和 Non。例如：形容词“大”能且仅可能跟一个事物类的词发生语义关联，如：“大树　大雨　声音大”等；名词“儿子”能且仅可能一个事物类的词发生语义关联，如：“老王的儿子”；动词“咳嗽”也是能且仅能跟一个事物类的词发生语义关联。如：“老王咳嗽得厉害”。这些词的配价数就都为 1。而形容词“热情”能跟两个事物类的词发生语义关联，如：“老王对我很热情”；名词“意见”可以跟两个事物类的词发生语义关联，如：“老王对你的意见”；动词“吃”也可以跟两个事物类的词发生语义关联，如：“孩子在吃苹果”。这些词的配价数就是 2。动词“给”可以跟三个事物类的词发生语义关联。如：“老师给了学生一本书”，它的配价数即为 3。动词“例如”经常独用，不跟任何成分搭配。它的配价数就是 Non，表示“无”。

（2） 配项的语义限制描述

对每个具体实词，一般都要求指明它对配项成分（限于名词性成分）的语义限制。语义限制一般用名词所属的语义类来表示。这就是上文所谓的词对类的语义搭配描述方式；对不易从正面限定语义类的，也允许以否定的方式从反面限定；极端的情况下，还可直接由名词自身表示，或者由含某个关键字的名词串来表示（用计算机语言中常用来表示通配的符号“*”标记）。这也就是上文所谓的词对词的搭配关系描述方式。在本文的语义模型中，仅粗线条地描述一个实词可能有的下面三种配项成分的语义限制情况。

I. 主体: 动作或性状的发出者或承当者；事物的参照者，例如：

他把书给我　　动词“给”的主体是“他”，语义要求是“人”

老王的妻子　　名词“妻子”的主体是“老王”，语义要求是“人”

这把刀很快　　形容词“快”的主体是“刀”，语义要求是“*刀|*剑|*斧*”

II.客体: 动作或变化的影响者；事物的关联对象，例如：

突出包围圈　　动词“突出”的客体是“包围圈”，语义要求是“包围圈”|“*围”

突出重点　　动词“突出”的客体是“重点”，语义要求是~“包围圈”~“*围”[1]

III.邻体: 事件中的受益者或受损者，例如：

他把书给我　　动词“给”的邻体是“我”，语义要求是“人”

值得指出的是，由于搭配的不同，一个词在实际使用中往往取不同义项，特别是常用的形容词、名词和动词更是如此，一词多义的现象非常普遍。特别是联系机器翻译的实际，要考虑到语际间意义对应的问题，就更要根据具体情况做区分了。仍以形容词“大”为例，由于跟“树”、“雨”、“声音”三个不同的名词搭配，英语对译选词也不同。分别是“big”、“heavy”、“loud”。这样，从机器翻译的要求出发，在描述一价形容词“大”的配项语义限制时，恐怕就得落实到具体的词来区分多义了。譬如“大”的“主体”至少就应区分出“雨”和“声音”这两个词（即以词对词的方式来描述搭配限制，而对“树”跟“大”搭配这样的情况则可作为比较普遍的类型来处理（即以词对类的方式来描述搭配限制）。这也比较合乎说汉语的人对“大”的英译语感。上面例中“快”也是如此。意为“锋利”的“快”跟意思是“迅速”的“快”同样可根据“主体”成分限制的不同来区分。

能跟动词发生语义联系的名词性成分一般认为还有时间、处所、方式等等多种情况。我们不做详细区分。根据笔者的经验，如果组织汉语句法知识得当，很多涉及这些语义类型的搭配问题，以及其他一些配价语法跟论旨理论都作为语义成分看待的东西如小句宾语、动词性宾语等等，都可在句法层面得到很好的解决。

四、实践：对 4 万余汉语实词的语义归类和语义搭配描述

在上述语义模型基础上，课题组对 4 万余汉语实词进行了语义归类和语义搭配的描述。在确定一个词的语义类归属时，可以根据词的概念义同时参照其句法特征进行操作。一个词允许分属不同的语义类。在描写一个词的语义搭配性质时，倾向是从宽，即在不好确定配项语义类属性时，可归入上层语义类。但对用法受限的词语，描写它的搭配成分则具体到词。

在已经作了语义属性描述的 4 万余汉语实词中，包括名词 27,828 个、动词 10,787 个、

[1] 我们用“~”表示逻辑上的“非”。

形容词 2,640 个。这其中一万多动词基本来自《现代汉语语法信息词典》的动词分库，根据义项分合不同略有些调整。这基本上覆盖了汉语常用动词的范围。下面我们给出动词的部分样例以及一些统计数据，以增进对这个语义模型的直观认识。见图 1 和表 1。

图 1　动词库样例

配价数	Noun	1	2		3		
总词数	9	4782	6881		115		
符合一定配项条件的词数		主体:人类	主体:人类	客体: 人类	主体:人类	客体:人类	邻体:人类
		3539	5060	1407	115	14	115
例词	例如	懊悔	爱 1	想 2	送 1	增援	捐赠

表 1[2]　动词语义属性部分统计数据

上面表一“符合一定配项条件的词数”给出的统计条件比较简单。这里的数据只是想说明我们的语义模型实践结果跟语感一致。在以有的实践基础上，我们可以有计划地开展汉语语义的定量研究。

五、语义信息在汉英机器翻译中的效用

这一节我们通过分析一个实例来说明翻译系统中引入语义处理机制的效果。请看例句：

安装网络系统的人正在想主意。

机器经过一定的句法分析后可得到这样的分段结果：

安装/v [np 网络/n 系统/n] 的/u 人/n 正在/d 想/v 主意/n

为最终产生正确的译文，需要计算机对下面三个问题进行判断：

1）“安装网络系统的人”的抽象句法格式为 “v np u<的> n”。这是一个歧义格式。至少有 a. [[v np u<的>] n]和 b. [v [np u<的> n]]两种组合方式。

2）“网络系统”由两个名词组合而成，内部关系可能是联合，也可能是定中。如何

[2] 表一中例词“爱、想”等加下标表示是该词多个义项中的一个。“主体:人类”表示该词的主体语义类要求是“人类”，余类推。

确定？

3）“想”是多义动词，至少有“思考”和“想念”两个不同的意思需要计算机甄选。

上述三个问题利用句法信息都不易做出判断。而利用我们给出的语义信息，则可得到正确的答案。首先动词“安装”在词典中已经标记其有“客体”配项，且要求语义类是“人为事物”，不能是“人”，这就可以得到第一个问题的答案为 a。“网络”和“系统”两个名词的语义不同类，按照无标记联合结构对内部成分语义同类的要求，可判定它们只能构成定中关系。“想”在取“想念”义时，要求“客体”语义类必须是“人”。跟“主意”搭配的“想”只能被判定为“思考”义。这样，最终我们就得到如下译文：

Man who installs network system is thinking of idea

再经过一些冠词及单复数词形变化等后续处理对译文作适当调整，可将译文修饰得更自然准确。即使不做后续处理，这样的译文也已经大致能传达汉语原句的基本意思了。

六、结束语

本文介绍了一个汉英机器翻译系统的语义处理框架，即在语义分类的基础上对动词、形容词以及名词进行搭配关系的描写，帮助计算机在判定短语结构和确认词语的意义以及分析词语之间的语义关系时能够做得更准确，从而提高译文的质量。

可以看出，我们引入的语义处理框架比较简单，对词语之间语义关系的描述基本上是粗线条的，但粗中有细。系统能够应付相当多的一般性语义问题。不过在需要对词语间语义关系做出精细区分的时候就又可能力不从心了，而这在实际语料中还是会碰到不少的。并且在确定词语之间的语义搭配关系时，限于短语结构语法的描写能力，基本是描述处于同一句法层次上的两个直接成分（IC）之间的语义关系，对处于不同句法层次上的非直接成分之间的语义关系较少兼顾。这一处理模式尽管使译文质量有了一定程度的改善，但同时也存在这类局限性，还需不断探索改进。

致谢：本文所指的汉英机器翻译系统由中科院计算所二室刘群副研究员总体设计开发。他对本文讨论的语义模型提供了大量支持和实际帮助，在此我们特别向他表示感谢。参加我们的语义词典实际开发工作的成员还包括北大计算语言学研究所王惠老师、北大英语系左岩博士、数学系余江生博士、中科院计算所二室刘颖、王斌两位博士及叶煜、张立红等，在此一并致以诚挚的谢意。

参考文献:

1 韩万衡(1997)《德国配价论主要学派在基本问题上的观点和分歧》，《国外语言学》，1997 年第三期。
2 韩万衡(1992)《德语配价语法》，商务印书馆，1992 年版。
3 张烈材(1985)《特斯尼埃的〈结构句法基础〉简介》，《国外语言学》，1985 年第 2 期。
4 李洁(1985)《Kalevi Tarvainen 的〈从属关系语法导论〉》，《国外语言学》，1986 年第 3 期。
5 袁杰(1986)《〈德语动词句法和语义配价词典〉评介》，《国外语言学》，1991 年第 1 期。
6 沈阳、郑定欧(1995)《现代汉语配价语法研究》，北京大学出版社 1995 年版。
7 Fillmore(1968) 胡明扬 译《“格”辨》，载《语言学译丛》第二辑，中国社会科学出版社 1980 年版。
8 杨成凯(1986)《Fillmore 的格语法理论》（上、中、下），《国外语言学》1986 年第 1、2、3 期。
10 Chomsky(1993) 周流溪等译《支配和约束论集》，中国社会科学出版社 1993 年版。
11 徐烈炯(1988)《生成语法理论》，上海外语教育出版社 1988 年版。
12 顾阳(1994)《论元结构理论介绍》，《国外语言学》1994 年第 1 期。
13 顾阳(1994)《生成语法及词库中动词的一些特性》，《国外语言学》1996 年第 3 期。
14 程工(1994)《Chomsky 新论：语言学理论最简方案》，《国外语言学》1994 年第 3 期。
15 汤庭池、张淑敏(1996)《论旨网格、原参语法与机器翻译》，《中国语文》1996 年第 4 期。
16 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
17 陈群秀 张普(1995)《信息处理用现代汉语语义分类体系：属性分类》，载《中文信息处理应用平台工程》，电子工业出版社 1995 年版。
18 鲁川(1995)《现代汉语的语义网络》，出处同上。
19 梅家驹 等(1985)《同义词词林》，上海辞书出版社，1985 年版。
20 朱德熙(1982)《语法讲义》，商务印书馆，1982 年版。

21 朱德熙(1985)《语法答问》，商务印书馆，1985 年版。
22 俞士汶 等(1996)《现代汉语语法信息词典规格说明书》，《中文信息学报》1996 年第 2 期。
23 詹卫东 常宝宝 俞士汶(1997)《现代汉语短语本位语法体系在机器翻译中的应用及其问题》，载《智能计算机接口与智能应用学术会议论文集》，第三届中国计算机智能接口与智能应用学术会议论文集，电子工业出版社 1997 年版。
24 詹卫东(1997)《面向自然语言处理的现代汉语词组本位语法体系》，《语言文字应用》1997 年第 4 期。
25 詹卫东、刘群(1997)《词的语义分类在机器翻译中的作用以及难以处理的问题》，载《语言工程》，第四届全国计算语言学联合学术会议论文集，清华大学出版社 1997 年版。

A Semantic Model Based on Phrase-Standard Grammar of Contemporary Chinese

Weidong Zhan* Baobao Chang** Shiwen Yu**

*Department of Chinese Language & Literature Peking University, Beijing, 100871

Email: zhanwd@mtgroup.ict.ac.cn

** Institute of Computational Linguistics Peking University, Beijing，100871

Email: yusw@pku.edu.cn

Abstract: In this paper, after the similarities and differences of the three theories on semantics, including Valent grammar, Case grammar, Theta theory, are reviewed synoptically, we introduce a semantic model based on phrase-standard grammar of contemporary Chinese for a Chinese-English machine translation system. A semantic classification system for Chinese is built, and more than 40,000 words, including nouns, verbs, adjectives, are marked with semantic attributes, which describe the semantic feature and collocation constraint of a word. Some basic issues on building a semantic dictionary are discussed, and the performance of the CEMT system augmented with semantic information is also shown by an example.

Key words: **semantic model，semantic classification，semantic collocation, phrase-standard grammar, semantic attribute, machine translation**