构造一部改进的日语述语语义词典的初步设想

陈治平
解放军外国语学院，河南 洛阳 471003
E-mail：chen_zp95@163.net

摘　要：针对目前的述语语义词典中关于论元的信息无法有力支持嵌套格框架划界问题等缺陷，提出了构造一部改进的日语述语语义词典的初步设想并给出了样例。对设想作了预期评估并对今后的完善作了展望。
关键词：述语语义词典；论元；格框架

A Tentative Idea on Compiling a Revised Semantic Dictionary of Japanese Predicates

Chen ZhiPing
PLA Foreign Languages University , Luoyang, Henan Porv.,471003,China
Email：chen_zp95@163.net

Abstract: In the light of the shortcomings existed in the present semantic dictionaries of Japanese predicates that the arguement infomation can hardly support the issues of the demarcation of embedded case frame ,we propose a tentative idea of compiling a revised Japanese predicative semantic dictionary and provide some examples as well. We also make an expected evaluation to this tentative idea and point to the prospect of improving it.
Key Words: predicate semantic dictionary; argument; case frame

1 研究背景

构造基于格框架的日语述语语义词典即 CJPSD（Case－frame－based Japanese Predicate Semantic Dictionary）的主要目的是为了给自然语言处理系统提供语义信息的支持。主要用于排除句法分析之后仍无法排除的歧义，或与句法分析融合使用，从而得到关于句子更明确的语义信息。

格语法和配价语法在目前的自然语言处理研究中得到了广泛应用。尤其是日语，由于有格助词，为深层格提供了语法标志，因而格语法和配价语法更是成为日语计算机处理研究中的主流技术。高松忍（1981）、田中惠积（1982）、石绵敏雄（1971、1983、1999）均进行过

这样的研究。基于格和配价的电子词典的开发更是曾掀起一股高潮。最有名的是电子词典研究所（EDR）开发的 EDR 系列词典，此外还有 IPA 的计算机用日语基本动词词典等。

综观前人研究，可以说达到了一个理论、技术均较成熟的较高境界。但这并不意味着没有提高的余地。我们认为，目前的述语语义词典关于论元的信息还不是最充分，遇到格框架嵌套的场合，无法给分析系统以有力支持。这是一个很有必要填补的缺陷。此外，对工具格一类的半自由论元的研究较少，日汉双语对译的格框架词典也较少见等等。这些都有待完善。

针对以上有待完善的地方，我们提出构造一部改进的基于格框架的日汉双语述语语义词典的初步设想。限于篇幅，不涉及过多技术细节。

2 基本原则

构造 CJPSD 的几个基本原则：吸纳国内外已有理论成果，采用成熟的操作技术，在前人基础上工作；采用符合自然语言处理需要的名词语义分类体系，边编写边调整，很多时候，需要根据特定的述语来调整分类的标准；只收录现代日语中的基本用言；对于多义词，只要义项中的论元发生了实质变化，都作为不同词条收入。

3 语言模型

CJPSD 的语言模型拟以已经成熟的格语法和配价语法为原型，将两者融合使用。为方便计，以下均称为格框架。语义信息的依据是对真实语料的人工分析，权威的日、汉语传统词典和研制人员的语感。语言模型中与现有述语语义词典类似的不再赘述。介绍一下改进处：

（1）将工具格和处所格（不是空间格）也列为描写的论元。关于工具格和处所格是否为核心论元的问题，不在本文讨论之列。但从实践来看，某些动词（主要是人的有意识动作）的工具格和处所格具有一定的限制性，对语义分析和排除多义性有一定作用。如：

①IE5.0 **で読んで、セーブしてください。**/请用 IE5.0 浏览，并保存。

分析时，机器会面临判别“IE5.0 **で**”是修饰“**読む**”还是修饰“**セーブする**”的问题，这可以通过比较这两个词在词典里的格框架而解决。CJPSD 中“**読む**”的工具应该是“浏览类软件”，“**セーブする**”的工具是“软件部件”，而“IE5.0 **で**” 属于“浏览类软件”，因此，“IE5.0 **で**”与“**読む**”相匹配。

因此，对这类动词要作工具格和处所格的描写。目前还作不到太细，因为工具格和处所格是半自由的，有一定的随意性，不象主体、客体、邻体等论元限制性那么强。可以观察得到，有一些动词的这两个格的语义特征呈多点分布，如

洗濯する：工具格∈〔肢体〕〔化学制品〕〔家用电器〕〔……〕……

（实际词典中的格和语义特征以代码表示，在此为方便起见，用中文表示）

一一描写并非易事，现在我们可以作的是对其在常态下的粗线条的描写，尽量把比较常

见、比较重要的、对语义分析起作用大的描写出来。对一些随意性太大的动词，将其工具格和处所格以自由论元处理。以“読む”为例，可以如下描写：

読む：工具格∈〔视力工具〕〔浏览类软件〕，处所格∈〔〕

（〔〕：完全自由）

这样，我们在CJPSD中拟对述语的主体、客体、邻体、工具、处所等5个论元进行描写。

（2）日汉双语格框架对照。从理论上说，真正的格框架应该是完全抛弃语言表层信息的人类认识共性框架。这也就是所谓的中间语言。但从NLP研究实践来看，这样的程度至少在目前还无法企及。相反，倒是在两种语言的表层与深层之间的一个中间层构造一个所谓带有少量表层特性的深层格框架的对照关系更有益于语言处理特别是机器翻译的研究。有鉴于此，我们在词典中采用日汉双语格框架对照。这样方便日汉机译系统的设计，而且，更重要的意义在于，一种语言的格框架只有分析到了与另一种语言的格框架基本相同的地步，才可以说是真正成功的。以这样的格框架作为机器理解的原语言才可能是合理的。

以“読む”（“阅读”的义项）为例：

読む（〔人〕（が／は），〔作品〕（を））∧（〔视力工具〕∧〔浏览类软件〕，〔〕），→（读，阅读，阅览，看）（〔人〕，〔作品〕）∧（〔视力工具〕∧〔浏览类软件〕，〔〕）

“読む”和“读，阅读，阅览，看”的格框架基本一样。唯一不同的是“読む”的格框架有格标志（が／は）和（を），而“读，阅读，阅览，看”没有，但这属于表层特征。

（3）对论元信息作充分描写，以解决复句中格框架嵌套的问题。这是CJPSD的最主要改进处。

由于日语中有格助词，便于单一格框架的分析，因而目前日语处理系统对日语单句的分析已比较成熟。但若是复句，尤其是一个单句作定语的复句的情形，则会面临一个格框架嵌套的问题，这时格框架的分界问题将成为机器分析的一个瓶颈。如：

②私は、美術や音楽に関する本を読むことも結構であろうが、それよりも、何も考えずに、たくさん見たり聞いたりすることが第一だ、といつも答えています。/我一直这样回答：读有关美术和音乐的书固然好，但是比起这个来，什么也不考虑，多看一看、听一听是最好的。

这个句子中，出现了多个述语及它们的论元，这些论元该属于哪个动词，将成为语义分析中的一个棘手问题。我们认为，各种论元都和其中心述语有着内在的密切联系，因而最主要的判定方法仍是依据论元的信息。这就要求对论元的描写要充分。

为简化起见，我们讨论格框架两层嵌套的简单情形，更复杂的依此类推。设句子中有述语A，它有一个论元a，还有一个相异的述语B，它有一个论元b，两者构成了嵌套关系，机器在判定a和b的归属问题时产生了两种选择。下面分几种情形加以讨论：

1）a和b各自隶属的语义特征集合不相交，则可以直接根据其语义特征与述语匹配，判定其归属哪个格框架。（①句的分析是它的一种简单情形）

2）a和b各自隶属的语义特征集合相交，这时又有两种情形。

I）a隶属的是b隶属的一个子集。可以两种选择同时考虑，先确定a的匹配对象，虽然b的语义特征兼容于述语A和B，但既已确定a为A的论元，则可排除b为A论元的可能性。

II）a,b隶属的是部分重合或完全重合的两个集合。这是最复杂的一种情形，此时单纯用语义特征与述语匹配已失去作用，需要结合更多的语言信息才能解决。

一条途径是运用句法或语用的规则进行定位，如日语中的“は”有总领全句的功能，这要求在复句中，它要与位于句末的述语发生联系，若 a 后附有格标志“は”且该句排列顺序为 a（は），b，B，A 的话，则 a 应为 A 的论元。这样的规则已不属于语义的范畴，可以说是句法的，也有意见认为是语用规则，但可以确定的是，它们对于语义分析的格框架的确定有作用，象上面的规则，对于所有 a（は），……，A 型的复句的嵌套格框架的划分都是起作用的。而在句法或语用分析中，它们的作用则不大。有鉴于此，我们可以将这样的规则编写为附加规则库，加入语义词典中，系统若碰到划分嵌套格框架的问题，即调出参加处理。

如果 a，b 均符合附加规则，此时对 a，b 归属的划分只有通过统计方法解决了。我们认为，特定论元对特定述语有一个依赖度大小的问题，有的述语一出现，它的某一个论元就必须出现，如日语的大多数他动词，有的述语则不一定，如日语的很多自动词，有的述语则很少出现，如“思われる”的主体论元在句子中极少出现，可称为隐形论元。即：

必现论元→非必现论元→隐形论元

由左至右依赖度呈线性递减。

依赖度 D 的计算可以用语料库分析大规模的真实文本而得出，即；

D＝该论元出现次数 n/该论元所依赖的述语充当中心述语的句子或子句总数 N

我们计算出 a 对 A 的依赖度 D（A，a），b 对 A 的依赖度 D（A，b），从而计算出在一个被分析结构 A（a，b）中 a 被匹配为 A 论元的优先度 P（A，a）＝D（A，a）/D（A，a），同理，我们计算出在被分析结构 B（a，b）中 a 被匹配为 B 论元的优先度 P（B，a）＝D（B，a）/D（B，a），比较 P（A，a）和 P（B，a）的大小，按优先度高的优先匹配。同理对 P（A，b）和 P（B，b）的大小也进行计算，以验证 a 的匹配是否正确。

该方法的缺陷是只能保证一定程度的正确率，对于离散的语言现象无能为力，这也是基于语料库方法的通病。因此，系统在分析时，确定性的附加规则的调用应优先于依赖度信息的调用。如果遇到的是离散的、特殊的语言现象，那只能借助于系统对更大范围的上下文语境分析了。但在可预期的将来，我们未必能在这方面取得突破。不过在一种特殊情形下，可以保证完全的正确率，那就是 D（a）和 D（b）之中有一个或全部达到了 100%，也就是 a 或 b 有一个或都是其述语的必现论元的时候，此时的匹配理论上是 100%正确的。

综上所述，为了对解决格框架嵌套问题有帮助，我们应该在 CJPSD 的构造中作好如下工作：

首先是要对论元的语义特征描写的尽可能细，争取将尽可能多的情形转化为 1)的情形，尽量避免两个语义相差较远的述语，其论元的语义特征却归入一个特征集。其次是将一些对论元划分有用的句法和语用规则编写入附加规则库，作为 CJPSD 的组成部分，只要系统在一句之中找到了一个以上的述语便将其调入分析模块。此外，还要运用语料库分析大量的真实文本，计算出每个述语的每个论元对述语的依赖度，作为论元的一个重要信息记入词典。

但限于我们目前的条件，还无法做到对大规模真实文本的自动分析，因此我们只能先对论元的依赖度作一个粗线条的定性描写。我们拟将论元分为三类：

a．必现论元（D＝100%）

b．非必现论元（0 < D < 100%）

c．隐形论元（D≈0）

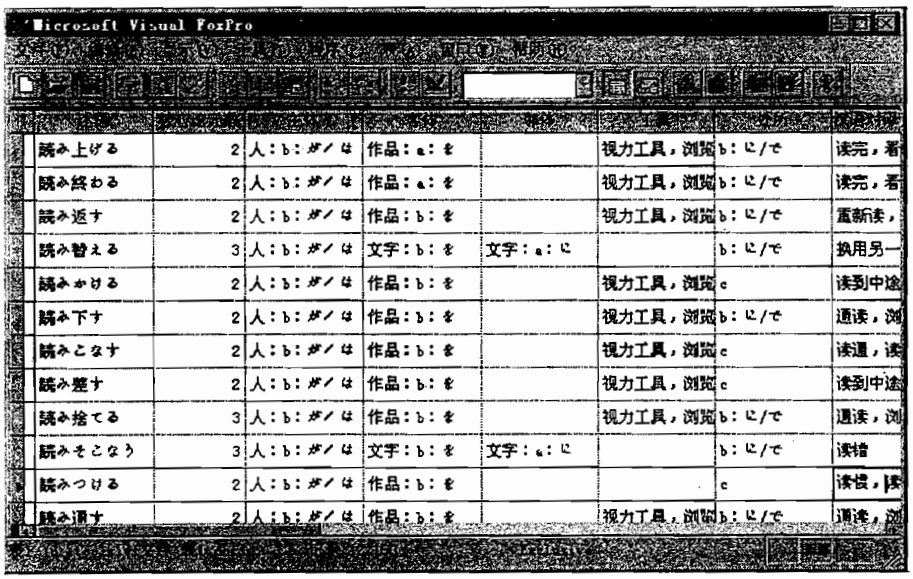
对所有述语的论元按这三个等级进行定性的描写。

4 计算模型

CJPSD 的构造拟采用微软的 Visual Foxpro 关系型数据库管理软件，包括 5 个数据库文件，动词、形容词、形容动词各一个，此外还有一个成语库和一个附加规则库。这样构造移植性较好，可以通过关键字段的匹配与其他词典（如句法词典）连接。VF 也是许多程序语言所支持的一种数据库，对不同系统的兼容性较好。

5 构造 CJPSD 的实践

我们现正在对日语中表示人类活动的 400 个左右的基本动词进行逐一描写，以下是一个样例。



（说明：a,b,c 均为论元依赖度的等级。汉语的对译词有时并不是纯粹的词，而有可能是动补短语或其他短语。为方便读者看，格和语义特征集的名称用汉语表示。）

附加规则库的编写则较为困难。除（三）中提到的外，我们还发现了一些，现正在对它们的适用范围作全面的考察。

6 设想的预期与今后的展望

我们的预期：CJPSD 将在以往的述语语义词典的基础上有所提高，针对第一部分所提的 3 个缺陷作出改善，将对提高语义分析的质量起到一定作用，也会方便 NLP 系统的设计，相

信对汉语在内的其他语言的述语语义词典的编制也有借鉴作用。例如①句中，不通过工具格的语义特征与述语的匹配，就无法排除歧义。又如：

③私たちは、何とか、この本を読むことが一番大切だ、と思われる。

计算机在进行语义分析时会遇到以下的问题：

a.“**私たち**”是“**読む**”的论元还是“**思われる**”的论元？

b.“**思われる**”至少可以有“被认为”和“总觉得”两个意思，应该选择哪个意思？

若调用传统的语义词典无法判定，因为传统词典中“**読む**”和“**思われる**”的论元隶属的语义特征集合部分重合，“**思われる**”（被认为）的客体论元和“**思われる**”（总觉得）的主体论元隶属的语义特征集合也是部分重合。若调用 CJPSD，首先，根据附加规则库中的规则可以判定，“**私たち**”与句末的“**思われる**”同属一个格框架。其次，根据依赖度信息，“**思われる**”（被认为）的客体论元对述语的依赖度等级为 c，而“**思われる**”（总觉得）的主体论元对述语的依赖度等级为 b，因此可以确定“**思われる**”的语义应为“总觉得”。

由此可见，CJPSD 有其自身的价值。尽管如此，我们也必须认识到，目前的这部 CJPSD 仍有很多问题解决不了，语义特征与述语的匹配排除不了全部的歧义，可能还有一些有用的半自由论元未列入描写范围，日汉双语的格框架对照可能还作不到一一完全对应，目前对论元的描写精度还无法保证完全解决划分格框架嵌套的问题。最大的缺陷就是限于条件，无法运用语料库从大规模的真实文本中自动抽取并分析语义信息来编制 CJPSD。这些问题都有待日后解决。

对于今后的展望，一是在编好 CJPSD 的基础上，有条件时用语料库从大规模的真实文本中自动抽取并分析语义信息来完善 CJPSD；二是向广义的格框架模式扩展，编写描写可以作修饰语的名词、形容词、形容动词、连体词等与名词之间语义关系的语义词典；三是可以用同样的方式编写一部基于格框架的汉日双语述语语义词典以及其他语言的语义词典，尤其是汉语的语义词典，因为汉语的语义分析更加复杂，更需要语义词典的有力支持；四是在时机成熟时，尝试编写词汇的语用词典。不少词汇在实际使用时都有语用意义，对理解有重要作用。

参考文献

[1]冯志伟：《自然语言机器翻译新论》，语文出版社，1994 年，北京

[2]陆汝占、靳光瑾：“汉语配价理论与语义计算”《现代汉语配价语法研究（第二辑）》，，袁毓林、郭锐主编，北京大学出版社，1998 年，北京

[3]姚天顺等：《自然语言理解——种让机器懂得人类语言的研究》，清华大学出版社，1995 年，北京

[4]俞士汶、朱学锋：《计算语言学文集（第 4 集）》，北京大学计算语言学研究所，2000 年，北京

[5]詹卫东：“基于配价的汉语语义词典”，《语言文字应用》，2000 年 1 期，北京

[6]郑家恒、黄昌宁：“信息处理用动词细分类研究”，《计算机时代的汉语和汉字研究》，罗振声、袁毓林主编，清华大学出版社，1996 年，北京

[7]石綿敏雄：「現代言語理論と格」，ひつじ書房，1999，东京

[8]国立国語研究所：「分類語彙表」，秀英出版，1965，东京

[9]长尾真：「自然言語処理」，岩波書店，1996，东京