

语义互联网与辞书编纂¹

冯志伟

(教育部语言文字应用研究所, 北京 100010)

[关键词] 互联网; 语义互联网; 本体词汇; 知识本体; 辞书编纂

[摘要] 本文分析了语义互联网与辞书的关系, 说明了《新编同义词词林》的编写工作是建立在知识本体的基础之上的, 充分考虑到语义互联网建设的需要, 在词典的编纂中, 知识本体起着重要的作用。

[中图分类号] H030; H163 [文献标识码] A [文章编号] 1671-5306(2009)04-0088-07

最近, 鲁东大学在《同义词词林》的基础上, 重新编写一本同义词词林, 叫做《新编同义词词林》, 即将由上海辞书出版社出版。在编写过程中, 他们邀请我作为他们的顾问, 我根据自己在机器翻译研究中设计的 ONTOL-MT 本体知识体系, 参考《同义词词林》的语义代码, 为《新编同义词词林》设计了一个新的代码系统。

1. 《新编同义词词林》语义代码系统设计的4个原则

1982年7月《同义词词林》出版时, 著名语言学家郭绍虞先生为《同义词词林》作序, 他从修辞和文法的角度, 论述了学习词汇的重要性(梅家驹等, 1982)[1]。他引用《文心雕龙》中的“句之清英, 字不妄也”来说明, “古人学文在于记住字和词的用法, 这才是一个真正的难关”, 他明确指出, “学中文的可以不从文法入手, 但是不能不从这些繁多的词汇入手”; 他又指出, 像《同义词词林》“这一类词书, 看似不讲文法和修辞, 但把汉语文法修辞两种学科, 都包赅在内, 经过这具体训练, 比学习语法修辞要好得多, 因为就实用的意义讲, 确实比空谈语法修辞之类的学者要实际”。我完全同意郭绍虞先生的这种看法。《同义词词林》是一个词汇的宝库, 当我们写作时感到词穷而难以表达意思的时候, 查一查《同义词词林》, 我们就会豁然开朗, 从中挑选到恰如其分的词语来表达我们的思想; 《同义词词林》帮助我们排忧解难, 常常使我们体会到“山穷水尽疑无路, 柳暗花明又一村”的快乐。《同义词词林》出版25年以来, 对于中文写作和外文翻译是非常有帮助的, 它成为了我们写作和翻译的好助手。

近年来, 语言信息处理需要进行语义的形式分析, 急需一套能够反映汉语单词语义特征的代码化的语义系统, 而《同义词词林》中的每一个单词都有表示语义的代码, 正好是一个代码化的语义系统, 因此, 语言信息处理学界的专家们把《同义词词林》当做一个宝贵的语言资源, 并且把它改造成为计算机可读的电子文本, 有力地推动了我国语言信息处理的研究。

然而, 《同义词词林》在语言信息处理中的这种作用是郭绍虞先生在他的序中没有提到

[收稿日期] 2009-11-24

[作者简介] 冯志伟(1939-), 云南昆明人。教育部语言文字应用研究所研究员, 中国传媒大学博士生导师, 韩国科学技术院电子工程与计算机科学系教授, 大连海事大学讲座教授, 《中国语文》、《语言科学》、《语言文字应用》、《中国科技术语》等杂志编委, 《中国大百科全书·语言文字卷》编委。主要研究方向为计算语言学和应用语言学, 发表中外文专著20余部, 论文200余篇。

¹ [基金项目] 国家社会科学基金项目(03BYY019)

的，也是《同义词词林》的梅家驹等 4 位编者在编写时没有料到的；《同义词词林》的初衷是为了写作和翻译而编写的，编者并没有考虑到语言信息处理的特殊要求。因此，在语言信息处理中，《同义词词林》的语义代码往往会出现左支右拙、穷于应付的局面。

在这种情况下，我们深切地感到，需要从语言信息处理的需要出发，同时又要考虑到写作和翻译的需要，在《同义词词林》的基础上，重新编写一本同义词词林。鲁东大学多年来一直进行汉语语料库的研究，他们在词语的语义分类方面做了很多有价值的工作，成绩显著，因此，上海辞书出版社委托他们编写了这部《新编同义词词林》。

我参考《同义词词林》的语义代码，为《新编同义词词林》设计了一个新的语义代码系统。我在设计这个新的语义代码系统时提出了如下 4 个原则：

第一，普遍性原则

对于任何两个意义相同的单词，不管这两个单词属于什么语言，它们在新的代码系统中的概念只有一个。

远在 1949 年，美国洛克菲勒基金会的副总裁韦弗(W. Weaver)在讨论机器翻译的时候就提出，当机器把语言 A 翻译为语言 B 的时候，可以从语言 A 出发，通过一种中间语言(Interlingua)，然后再转换为语言 B，这种中间语言是全人类共同的。我们的代码系统中的概念结点也应当是全人类共同的，它们应当适用于不同的语言，应当具有普遍性 (Mitkov. R, 2009) [2]。

在普遍性原则的前提下，在编写不同语言的代码体系时，又应当考虑不同语言的特殊性，不过，特殊性是服从于普遍性的。新的代码系统表示的是语义，具有中间语言的性质，我们要首先考虑普遍性，其次才考虑特殊性 (Hallig. R, 1963) [3]。

目前这个代码系统只在《新编同义词词林》的编写工作使用，只局限于汉语，但是，我们在设计代码体系时，是充分地考虑到它的普遍性的，它应当是多种语言共同的、通用的。

第二，完备性原则

新的代码系统中的概念代码应当具有完备性，它们应当尽量能够覆盖人类在自然语言中表达的所有通用的基本概念 (Roget. P. M., 1851[4]; Berrey. L. V, 1962[5])。

第三，明晰性原则

新的代码系统中的概念代码之间应当是泾渭分明的，它们应当具有明晰的界限，尽量避免交叉或重叠。在使用代码来标注词典的时候，应当尽量把不同的概念明晰地区分开来 (Miller. G 等, 1990[6]; Miller. G, 1995[7];)。

第四，多角度原则

事物从不同的角度观察，可以具有不同的特性，因此，同一个单词也可能具有不同的代码标记，这正说明了事物本身的多义性，应该是正常的。在新的代码系统中，同一个单词可以具有不同的属性，因而可以从不同的角度标注以不同的代码。

这里的第一个原则是“普遍性原则”，我很明确地力图把《新编同义词词林》建立在具有普遍性的“知识本体”(ontology)的基础之上，以便使得《新编同义词词林》能够为新一代的语义互联网 (Semantic Web) 中的“知识本体”这个层次服务。

2. 知识本体在新一代语义互联网中的地位

2001 年互联网之父蒂姆·伯纳斯-李 (Berners-Lee. Tim, 2001) [8]提出如下的 Web 的体系结构，这样的 Web 叫做“语义互联网”，它将是新一代的互联网：

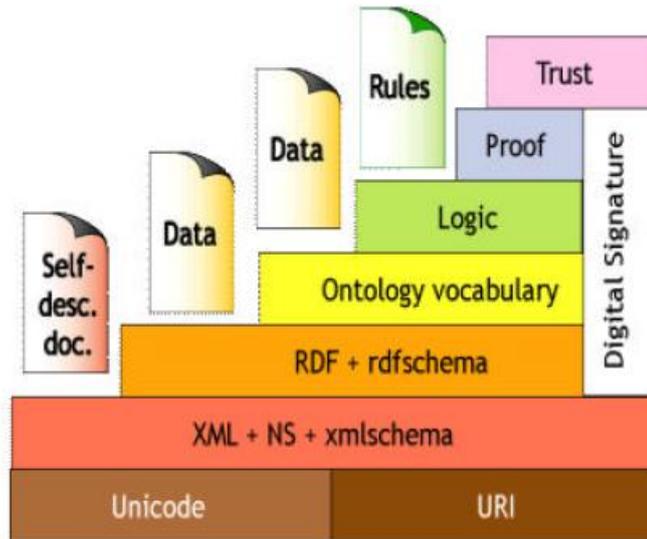


图 1. 语义互联网的体系结构 [8]

在这个语义互联网的体系结构中，Unicode 是国际统一的编码字符集，URI 是英语 Uniform Resource Identifier 的缩写，就是“统一资源定位符”，也被称为“网页地址”（简称网址），是互联网上标准的资源的地址。

XML 是英语 Extensible Markup Language 的缩写，就是“可扩展标记语言”，NS 是英语 Name Space 的缩写，就是“名空间”，xmlschema 就是 XML 模式。

RDF 是英语 Resource Description Framework 的缩写，就是“资源描述框架”，rdfschema（RDF 模式）就是“资源描述框架模式”。其进一步的描述如下图所示：

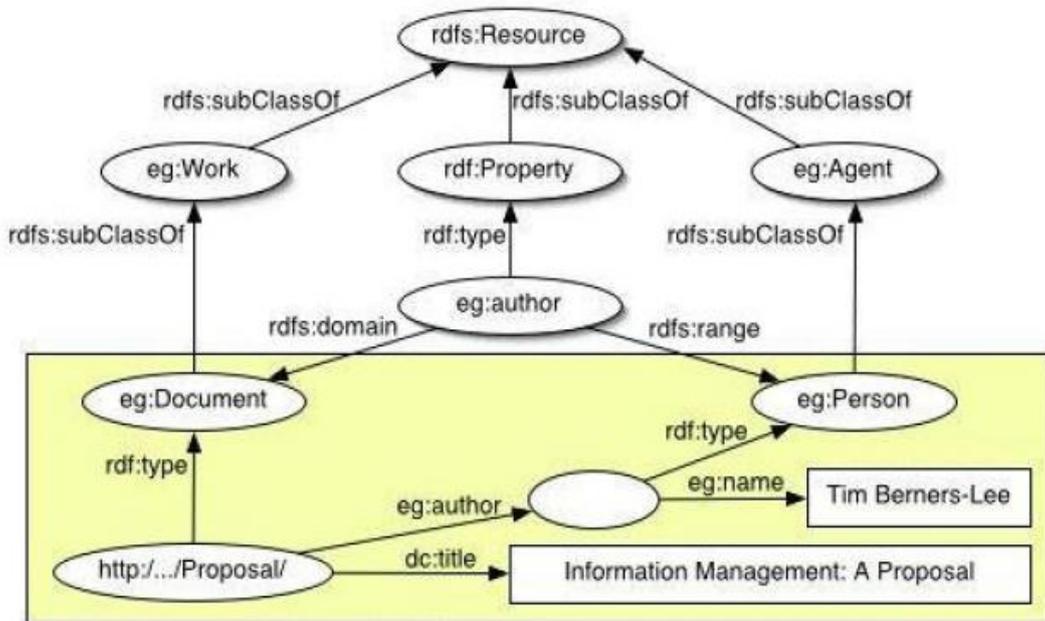


图 2. RDF 和 RDF 模式 (RDF Schema) [8]

RDF 和 RDF 模式中主要的类别、特性和约束如下：

—核心类别：包括 rdfs:Resources, rdfs:property 和 rdfs:Class。所有用 RDF 表达式描述的事物都被看成是 rdfs:Resources 的实例，rdfs:property 用来刻画 rdfs:Resources 实例的所有特性的类别，rdfs:Class 用来定义 RDF 模式中的概念。

—核心特性：包括 `rdf:type`, `rdfs:subClassOf` 和 `rdfs:subPropertyOf`。`rdf:type` 建立资源和类别之间的实例（instance-of）关系模型，`rdfs:subClassOf` 建立类别之间的包容层次模型，`rdfs:subPropertyOf` 建立特性之间的包容层次模型。

—核心约束：包括 `rdfs:ConstraintResources`, `rdfs:ConstraintProperty`, `rdfs:Range` 和 `rdfs:domain`。`rdfs:ConstraintResources` 定义了所有约束的类别，`rdfs:ConstraintProperty` 是 `rdfs:ConstraintResources` 和 `rdf:Property` 的子集，它包括了所有用来定义约束的特性。

因此，RDF 和 RDF 模式是开放的元数据框架，它定义了一种描述计算机可理解的数据语义的数据模型。

语义互联网体系结构中的 Ontology Vocabulary 就是“本体词汇”；Logic 就是“逻辑”，它使用“描述逻辑标记语言”（Description Logic Markup Language）来进行描述；Proof 就是“验证”，Trust 就是“信任”，“验证”和“信任”是网络传输不可或缺的重要因素，用户与用户之间，用户与网络服务器之间要进行验证，要保持诚信；必须要有“验证”和“信任”，用户与网络服务器之间才可以在网络上传输信息。

语义互联网体系结构左侧的 Self-desc.doc 是 Self-description document 的缩写，就是“自描述文档”，Data 是“数据”，Rule 是“规则”。

特别值得注意的是，在这个语义互联网的体系结构中，“RDF+rdf 模式”的上面是“本体词汇”（ontology vocabulary），“本体词汇”的上面是“验证”和“信任”，“本体词汇”处于语义互联网的关键层，它属于“自描述文档”，用于表示语义互联网各种信息的概念和语义。由此可见，“本体词汇”在语义互联网的整个体系结构中起着承上启下的联系作用，处于举足轻重的重要地位。采用“本体词汇”来描述语义互联网中各种资源之间的联系，可以克服目前万维网上的信息格式的异构性、信息语义的多重性以及信息关系的匮乏和非统一性等严重问题（Gomez-Perez. A, 2004）[9]。

2006 年 5 月蒂姆·伯纳斯-李又宣布，经过 10 年的努力，万维网联盟（The World Wide Web Consortium, 简称 W3C）已发布 W3C 推荐标准 80 余份，语义互联网已经具备了为达到成功的目标所需要的所有标准和技术，包括作为数据语言的 RDF、本体语言、查询和规则语言。

2006 年，蒂姆·伯纳斯-李又公布了语义互联网的新的体系结构：

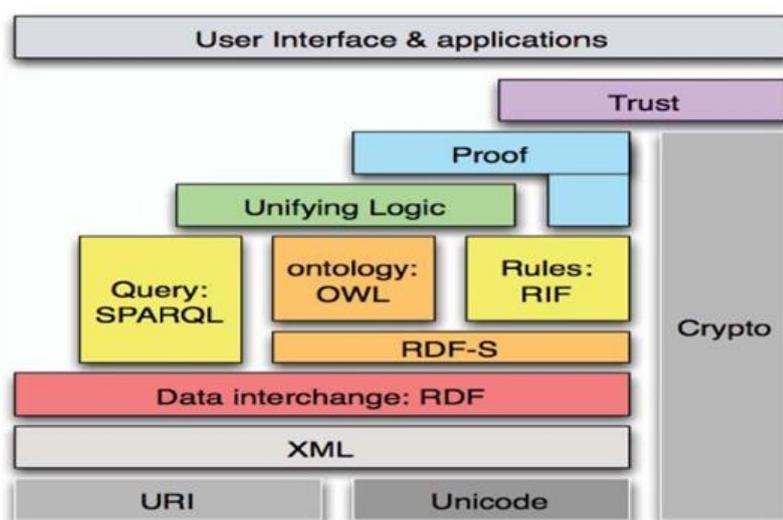


图 3. 语义互联网新的体系结构（Berners-Lee. Tim, 2006）[10]

在这个新的体系结构中，“逻辑”层变成了“统一逻辑”层（Unifying Logic），不再局限于使用特定的“描述逻辑标记语言”。在这个体系结构中，Ontology Vocabulary 变成了

Ontology OWL, 我们可以把它翻译成“知识本体”,“知识本体”直接地处于 Unifying Logic (统一逻辑)和 RDF-S (资源描述框架模式)之间,上承“统一逻辑”,下启“资源描述框架模式”,其承上启下的作用更加明显。由此可见,在 2006 年新公布的语义互联网体系结构中,“知识本体”的重要性更加突出了。

知识本体是描述概念的,在自然语言中用词或词组来表示,在 2001 年的语义互联网体系结构中,把知识本体直接称为“本体词汇”,足以说明知识本体与自然语言词汇之间的密切关系,因此,对于语义互联网来说,词的重要性更加突出了(Gruber. T. R, 1993[11]; Studer. R, 1998[12])。我们在《新编同义词词林》的代码设计中,已经充分地考虑到了词在语义互联网中的重要作用,坚持从知识本体的角度来设计我们的代码体系。我们这样做,一定会受到自然语言信息处理同行的支持。

3. 多语言信息网络时代辞书编纂工作的新使命

据中国互联网络信息中心(CNNIC)统计,截至 2008 年 6 月底,我国的互联网网民人数已经达到 2.53 亿,超过了美国的网民人数,成为了世界上互联网用户最多的国家,成为了首屈一指的互联网大国。

CNNIC 的统计数据最近又显示,截至 2008 年 12 月 31 日,我国网民数达到 2.98 亿人,互联网普及率达 22.6%。宽带网民规模达到 2.7 亿人,占网民总数的 90.6%。我国域名总数达到 16,826,198 个,其中 CN 域名数量达到 13,572,326 个,网站数约 2,878,000 个,国际出口带宽约 640,286.67Mbps。

由于互联网上使用英语之外的其他语言的人数增加得越来越多,英语在互联网上独霸天下的局面已经彻底打破,互联网确实已经变成了“多语言的网络世界”(multilingual Web),“多语言”这个特性使得互联网变得丰富多彩,同时也造成了不同语言之间交流和沟通的困难,互联网上的语言障碍问题显得越来越突出,越来越严重。

语言是信息的最主要的负荷者,如何有效地使用现代化手段来突破人们之间的语言障碍,成为了全人类面临的共同问题。“自然语言处理”(Natural Language Processing, 简称 NLP)技术包括机器翻译技术、跨语言信息检索技术、多语言问答式信息检索技术、多语言文本的自动分类技术、高效的搜索引擎技术;这些技术是解决语言障碍问题的有力手段之一。由于自然语言是人类历史长期发展的产物,带有浓厚的人文色彩,其结构极端复杂,其使用具有随机性和灵活性,在千百年数亿人的频繁使用中,由于历史长期积淀的差异和人们约定俗成方式的不同,自然语言在词汇层、句法层、语义层、语用层都充满了“歧义性”(ambiguity),而且自然语言处理技术又往往涉及到多种语言,需要语言学、计算机科学、数学多学科联合攻关,因此就更加复杂和困难(冯志伟, 2005) [13]。

可以毫不夸张地说,在进入 21 世纪之后,几乎每一个生活在信息网络时代的现代人,都要直接或间接地与自然语言处理技术打交道。不论对于社会政治还是对于经济发展,自然语言处理技术都无疑是一个重要的研究领域(冯志伟, 2009) [14]。

在信息时代,科学技术的发展日新月异,新的信息、新的知识如雨后春笋地不断增加,出现了“信息爆炸”(information explosion)的局面。现在,世界上出版的科技刊物达 165,000 种,平均每天有大约 20,000 篇科技论文发表。专家估计,我们目前每天在互联网上传输的数据量之大,已经超过了整个 19 世纪的全部数据的总和;我们在新的 21 世纪所要处理的知识总量将要大大地超过我们在过去 2500 年历史长河中所积累起来的全部知识总量。据 CNNIC 统计,2002 年底全球的网页总数已经达到 10^9 这样的天文数字,信息量的丰富大大地扩张了人们的视野,人们希望能够准确地、迅速地搜索到自己需要的信息,以自然语言为主要搜索对象的搜索引擎技术,将为解决海量信息的获取问题提供强有力的手段。

从 1954 年美国第一个俄语到英语的机器翻译实验获得初步成功开始,自然语言处理的

研究已经有 50 多年的历史了，在这 50 多年的发展历程中，自然语言处理把语言学、计算机科学、数学、心理学、哲学、统计学、电子工程、生物学等学科融合起来，形成了一门独立的边缘性交叉学科。自然语言处理的范围涉及到众多的部门，如语音的自动识别与合成、机器翻译、自然语言理解、人机对话、信息检索、文本分类、自动文摘，等等。在这 50 多年中，自然语言处理逐渐形成了自己独特的理论和方法，在当代语言学和计算机科学中独树一帜（Mitkov. R., 2009）[2]。

目前，在自然语言处理中的主流技术，是基于词法和句法分析的技术，尽管这些技术在某些受限的“子语言”（sub-language）中也曾经获得一定程度的成功，但是，这些技术难以有效地解决自然语言中普遍存在的“歧义性”问题，因而系统的质量不高，在实际应用中有很大的局限性。为了克服这样的局限性，自然语言处理需要在理论、方法和工具等方面实行重大的革新，其中一个重要的问题，就是在各种自然语言处理系统中，引入语义和概念的信息，以便进一步提高自然语言处理系统的智能。

例如，汉语的“张三吃面包”、“张三吃大碗”、“张三吃食堂”三个句子，它们的句法结构都是“主语-谓语-宾语”的格式，“张三”是主语，“吃”是谓语，“面包、大碗、食堂”是宾语；在汉语中，它们在句法结构上是没有任何差别的。但是，这三个句子的英语译文却各不相同，“张三吃面包”的英语译文是“Zhang San eats the bread”，“张三吃大碗”的英语译文是“Zhang San eats with a big bowl”，“张三吃食堂”的英语译文是“Zhang San eats in the restaurant”。在汉英机器翻译中，如果我们只使用词法和句法信息，按照“主语-谓语-宾语”的格式来翻译，这三个句子的英语译文都是相同的“Zhang San eats the bread”，“Zhang San eats the big bowl”，“Zhang San eats the restaurant”。显而易见，只有第一个句子的英语译文是正确的，而第二个和第三个句子的英语译文都是错误的，也是难以理解的。

但是，如果我们在机器翻译系统中引入关于单词的概念类别的语义信息，比如，“面包”的概念类别是“食物”，“大碗”的概念类别是“餐具”，“食堂”的概念类别是“建筑物”，根据这些概念类别来判定“面包”的语义功能是“受事者”，“大碗”的语义功能是“工具”，“食堂”的语义功能是“地点”，从而泾渭分明地把这三个在语义上不同的句子区别开来，在机器翻译时给它们分别赋以不同的语义功能，分别翻译为不同结构的英语句子：“Zhang San eats the bread”，“Zhang San eats with a big bowl”，“Zhang San eats in the restaurant”。

可以看出，一旦在自然语言处理系统中引入概念语义信息，便可以进行自然语言的“歧义消解”（disambiguation），使自然语言处理系统如虎添翼，把自然语言处理提高到一个新的水平。

近年来，国内外自然语言处理研究者已经逐渐认识到概念语义信息的重要性，开始在自然语言处理系统中引入一些概念语义信息；但是，这些概念语义信息大多数还是零零星星的、片段的、偶发性的，它们难以构成一个完整的系统。可以说，目前大多数自然语言处理系统还没有对概念语义信息进行过全面的、科学的、系统的研究。

《新编同义词词林》的语义代码是建立在“知识本体”（ontology）的基础之上的，我们力图从知识本体的角度出发，对自然语言中的概念语义信息进行全面的、科学的、系统的研究，建立一个比较完善和全面的语义代码系统，从而为机器翻译、信息检索、搜索引擎提供强有力的概念语义信息支持，大大地提高这些系统的智能化水平，推动我国自然语言处理的发展。当然，建立在知识本体基础之上的《新编同义词词林》同时也是一部同义词词典，它同样能够丰富读者的词汇知识，帮助读者提高写作和翻译的水平。这样，《新编同义词词林》不但能够为自然语言处理服务，也能为写作和翻译服务。

可以看出，我们在《新编同义词词林》中提出的基于知识本体的语义代码体系，是有深刻的科学根据的，是人类关于知识本体的研究在信息网络时代的新发展；它的理论意义和实用价值，已经远远地超出了郭绍虞先生在 20 多年前所强调的“修辞和文法”的领域，它将

会在自然语言处理和网络信息处理中发挥巨大的作用，对于互联网（Web）以及语义互联网（Semantic Web）的建设也是很有帮助的。

《新编同义词词林》的语义代码体系可以提供单词的概念类别特征，这些特征有助于提高机器翻译系统的质量以及歧义消解的能力，可以作为高质量的机器翻译词典编制的基础。

在机器翻译中，如果我们根据《新编同义词词林》中具有普遍性的语义代码来标注英语机器词典中单词的固有语义特征，由于汉语和英语的词汇都使用同样的语义代码，它们彼此之间的对应关系将变得非常清晰，这是一种新型的机器翻译系统，可以从根本上改善机器翻译系统的质量。

例如，汉语的“我/用钢笔/写/信”这个句子，使用《新编同义词词林》中语义代码可以标注如下：

“Ab01(人·第一人称)/Cf07(器具·文具)/Ka14(语言活动·书写)/Ck01(创作物·文书)”

这个句子的英语译文为“I / write / a letter / with a pen”，使用《新编同义词词林》中语义代码体系可以标注如下：

“Ab01(人·第一人称)/Ka14(语言活动·书写)/Ck01(创作物·文书)/Cf07(器具·文具)”。

不难看出，汉语句子和相应的英语句子在《新编同义词词林》中的语义代码完全是一样的，只是由于汉语和英语的语法结构的差异，这些标记排列的顺序不尽相同，而这种差异可以通过设计强大的句法语义自动分析软件来解决。

显而易见，在《新编同义词词林》语义代码这个层面上，同一个句子在不同语言中的标记得到了高度的统一，达到了完美的和谐，这就为多语言机器翻译系统的开发提供了有力的语言知识资源的支持。

从数学的角度来看，把A语言翻译为B语言的过程，就是把A的“显拓扑”空间（符号空间），通过“概念”还原到A的“潜拓扑”空间（语义空间），由于“潜拓扑”空间有一个绝对坐标系，使得A和B在“语义空间”上有统一的描述。翻译者只需把语言A的语义通过等价的变换，转成语言B的语义，再与语言B的“显拓扑”空间（符号）产生对应，就完成了翻译（Partee. B. H 等，2009）[15]。传统机器翻译只注重可表达的“显拓扑”（符号）部分，“潜拓扑”（语义）部分非常薄弱，因此在逻辑上难以解决各种类型的歧义问题，很难获得精确的机器翻译效果。《新编同义词词林》中的语义代码体系，为在潜拓扑空间上表达语义提供了巨大的可能性，这是我们进一步开发机器翻译系统的重要保证。

动态的人类知识库是一种智能数据库，人们梦寐以求的智能搜索引擎离不开智能数据库，智能数据库是智能搜索引擎的基础。在搜索中，关键词越多，限制条件越多，搜索的范围越准确。但是，如果数据库没有足够的概念储备以解读所有的关键词，就难以有效地进行这样的搜索。《新编同义词词林》中的语义代码体系是在知识本体的基础上构建的，它可以通过语义空间的逻辑转换，把任何语种下的相关内容从网络中搜索出来。因此，《新编同义词词林》语义代码体系的研究成果还可以在跨语言信息检索、文本自动分类、搜索引擎等系统中得到应用，提高系统的召回率（recall）和准确率（precision）。在搜索引擎中，可以根据《新编同义词词林》中语义代码的概念关联，给用户提供智能搜索提示，从而提高搜索引擎的效率。这样的研究方向有着非常广阔的市场前景和发展潜力。

当然，《新编同义词词林》也完全保持了传统的同义词词典的全部功能。当读者在写作和翻译中发生词穷的情况而难以恰当地表达意思的时候，《新编同义词词林》可以帮助读者从语义查询有关词汇，以便读者从中挑选恰当的词语，这对于写作和翻译仍然是很有帮助的。

《新编同义词词林》属于通用知识本体（common ontology）。这种知识本体从哲学的认

识论出发, 概念的根结点是很抽象的哲学范畴, 例如, 事物、时间、空间、数量、行为、状态、属性等。《新编同义词词林》描述的是人类的常识和普通的概念, 一般不描述专业知识和专业概念。

如果知识本体中的概念结点是专业术语, 那么, 这样的语言知识本体就叫做“术语知识本体”(terminology ontology)。术语是科学技术知识在自然语言中的结晶, 哪里有科学技术, 哪里就有术语, 所以, 术语知识本体对于领域知识的处理是非常重要的。术语知识本体是整个知识本体的一个重要部分, 术语知识本体的建立, 对于语义互联网有着重大的意义。

我们应当接受《新编同义词词林》编纂的经验, 针对语义互联网发展的需要, 在辞书编纂工作中, 开展通用知识本体和术语知识本体的构建工作。这是多语言信息网络时代我国科学词典编纂工作应当特别关注的问题。

总而言之, 在多语言的信息网络时代, 《新编同义词词林》中语义代码体系对于机器翻译、文本处理和搜索引擎等实用系统的开发, 对于语义互联网的建设, 有着良好的发展潜力和广阔的市场前景。我们应当敏锐地考虑到多语言信息网络时代对于辞书编纂的这些新的要求, 在新的背景之下更加科学地部署我国的辞书编纂工作, 使我国的辞书编纂工作能够基本上适应多语言信息网络时代的需要。这是辞书编纂工作在多语言信息网络时代的新使命。

参考文献

- [1] 梅家驹等, 《同义词词林》, 上海辞书出版社, 1982 年。
- [2] Mitkov, R., *The Oxford Handbook of Computational Linguistics*[M] (牛津计算语言学手册), 外语教学与研究出版社&牛津大学出版社, 2009 年。
- [3] R. Hallig & W. von Wartburg, *Begriffssystem als Grundlage für die Lexikographie (Versuch eines Ordnungsschemas)* [M], Akademie-Verlag, Berlin, 1963.
- [4] Roget, P.M., *Thesaurus of English Words and Phrases*[M], United States Book Company, New York, 1851.
- [5] Berrey, L. V., *Roget's International Thesaurus*[M], Third edition, Crowell, London, 1962.
- [6] Miller, G, Beckwith, R., Fellbaum, C., Gross, D., Miller, K., Introduction to WordNet: A on-line lexical database[J], *International Journal of lexicography* 3(4), 235-244., 1990.
- [7] Miller, G., WordNet: a lexical database for English[J]. *Communication of the ACM*, 38(1), 39-41, 1995.
- [8] Berners-Lee, Tim., Hendler, James., Lassila, Ora., The semantic web[J], *Scientific American*, 17 May, 2001.
- [9] Gomez-Perez, A., *Ontological Engineering with examples from the areas of Knowledge Management, e-Commerce and Semantic Web* [M], Springer-Verlag, 2004.
- [10] Berners-Lee, Tim., “Linked Data” (slides), at the TED 2009 conference, CA, USA, 4 February, 2009.
- [11] Gruber, T. R., A translation approach to portable ontologies[J]. *Knowledge Acquisition*, 5(2):199-220, 1993.
- [12] Studer, R., Benjamins. V. R., Fensel. D., Knowledge Engineering: Principle and Methods [J], *IEEE Transaction on Data and Knowledge Engineering*, 25(1-2), 1998.
- [13] 冯志伟, 从知识本体谈自然语言处理的人文性, 《语言文字应用》, 2005 年, 第 4 期。
- [14] 冯志伟, 自然语言处理的形式模型, 中国科学技术大学出版社, 2009 年。
- [15] Partee, B. H., Ter Meulen. A., Wall. R. E., *Mathematical Methods in Linguistics (语言学中的数学方法)* [M], 北京: 世界图书出版公司, 2009.

Semantic Web and Lexicography

Feng Zhiwei

(Institute of Applied Linguistics, Ministry of education)

Key words: Web; Semantic Web; Ontology; Lexicography.

Abstract: In this paper, the author analyses the relationship between Semantic Web and lexicography. He shows that the edition of *New Synonym Thesaurus* is based on the ontology. The editors of this *New Synonym Thesaurus* sufficiently consider the needs of construction of Semantic Web. The ontology plays an important role in lexicography.

【责任编辑 宗世海】