

词汇语义学与知识本体

冯志伟

教育部语言文字应用研究所

在语义自动分析中，单词本身的语义信息是很重要的，根据“组成性原则”，句子的语义是由构成该句子的单词的语义以及这些单词之间的语义关系组成的。因此，我们在自然语言处理中，应该重视词汇语义的研究。

1. 词汇语义学

语言中的词汇具有高度系统化的结构，正是这种结构决定了单词的意义和用法。这种结构包括单词和它的意义之间的关系以及个别单词的内部结构。对这种系统化的、与意义相关的结构的词汇研究叫做词汇语义学 (Lexical Semantics)。

从词汇语义学看来，词汇不是单词的有限的列表，而是高度系统化的结构。

在继续讲述词汇语义学之前，让我们首先引入一些新的术语，因为迄今为止我们用过的这些术语都过于模糊。例如，对于“词”(word)这个术语，目前已有各式各样的用法，这增加了我们澄清其用法的难度。因此我们将使用“词位”(lexeme)这个术语来替代“词”这个术语，词位表示词典中一个单独的条目，是一个特定的正字法形式和音素形式与一些符号的意义表示形式的组合。词典(Lexicon)是有限个词位的列表，从词汇语义学的观点看来，词典还是无限的意义的生成机制。一个词位的意义部分叫做“涵义”(sense)。

词位和它的涵义之间存在着复杂的关系。这些关系可以用同形关系、多义关系、同义关系和上下位关系来描述。

——同形关系

形式相同而意义上没有关系的词位之间的关系叫做同形关系 (homonymy)。具有同形关系的词位叫做同形词 (homonyms)。

例如：bank 有两个不同的意思：

① 银行 (financial institution)。在句子 “A **bank** can hold the investments in an account in the client's name.” 中的 bank 就具有这个意思，我们把它叫做 bank1。

② 倾斜的堤岸 (sloping mound)。在句子 “As the agriculture development on the east **bank**, the river will shrink even more.” 中的 bank 就具有这个意思，我们把它叫做 bank2。

Bank1 和 bank2 在意义上没有联系，在词源上，bank1 来自意大利语，而 bank2 来自斯堪底纳维亚语。

同形词可以分为两种：

- 同音异义词 (Homophones): 发音相同但是拼写法不同的词位。例如，wood — would; be — bee; weather — whether。
- 同形异义词 (Homographs): 正词法形式相同但是发音不同的词位。例如，bass [bæs] — bass [beis]。bass [bæs] 是一种皮肤带刺可食用的鱼，叫做“狼鲈”，而 bass [beis]: 表示低音。

在自然语言处理中，我们应该重视同形关系的研究。

- 在拼写校正时，同音异义词可能会导致单词的拼写错误。例如，把 “weather” 错误地拼写成 “whether”。
- 在语音识别时，同音异义词会引起识别的困难。例如，“to”、“two”和“too”发音相同，在识别时难以区分。
- 在文本—语音转换系统 (Text-To-Speech system, 简称 TTS 系统) 中，同形异义词由于发

音不同，会引起转换的错误。例如，bass[bæs] 和 bass [beis]。

——多义关系

一个单独的词位具有若干个彼此关联的涵义的现象，叫做多义关系现象 (polysemy)，具有多义关系的词位叫做多义词，这意味着，在一个多义词中的各个涵义是彼此相关的，而同形词各个涵义是不相关的。

例如，英语的 head 是一个多义词。它具有如下的涵义：

- ① 包括大脑、眼睛、耳朵、鼻子和嘴的身体部分。
- ② 物品的最前端。例如，“the head of the bed” (床头)。
- ③ 头脑。例如，“Can't you get these facts into your head?”中的 head。

②的涵义是从①的涵义的引申，③的涵义是①的涵义的缩小。各个涵义之间是有联系的。

判断涵义是否有联系的方法有两个：

• 词源判断法 (Etymology criteria)：词源上有联系的是多义词，词源上没有联系的是同形词。bank1 和 bank2 的词源不同，因此，它们的涵义没有联系，应该是同形词。

• 共轭搭配法 (Zeugma criteria)：把待判断的两个涵义用连接词组合到一个句子中，如果句子成立，则判断它们是多义词，否则，则判断它们是同形词。

例如，在句子 “Which of those flights serve breakfast?” 中，“serve”的涵义是“供应食品” (“to offer food to eating”)，在句子 “Does ASIANA serve Philadelphia?” “serve”的涵义是“提供服务” (“to work for”)。共轭搭配法使用连接词 and 把它们构成句子* “Does ASIANA serve breakfast and Philadelphia?”，这个句子是个病句，我们觉得有些怪，我们似乎不能把 serve 同时应用于 breakfast 和 Philadelphia，这说明 serve 的这两个涵义之间的联系不是很明显的，但是我们还不能决定它们究竟是多义词还是同形词。

可见，与词源判断法比较起来，共轭搭配法只能作为我们判断多义词或同形词的参考，它还不是非常过硬的判断手段。

不过，这种共轭搭配法可以帮助我们测定不同多义词的各个涵义之间的语义距离。

我们来研究下面的句子：

① “They play soccer.” 这里的“play”表示“进行某种使人愉快的体育活动” (“to do sport that passed the time pleasantly”)。

② “They play basketball.” 这里的“play”的涵义与上句相同。

③ “They play the piano.” 这里的“play”的涵义表示“操作某种乐器” (“to perform a music instrument”)。

④ “They play doctors and nurses.” 这里“play”的涵义是“孩子们在游戏中扮演某个角色” (“Children amuse oneself by pretending to be some roles in the games”)。

我们使用共轭搭配法造出如下的新句子，从而判定各个涵义之间的语义距离：

“They play soccer and basketball.”——这个句子可以成立，说明句子①和句子②中 play 的涵义最接近。

*“They play soccer and the piano.” ——这个句子有点儿怪！这说明句子①与句子③中 play 的涵义相距比较远。

** “They play soccer and doctors.”——这个句子非常奇怪！！这说明句子①与句子④中 play 的涵义相距很远。

在用共轭搭配法造出的这些新句子中，句子越是奇怪，它的可接受程度就越差。可见，共轭搭配法是判断词位的语义距离的一种有效的手段。

在语言学中，区分同形词和多义词是很重要的。不过，在自然语言处理中，由于同形词和多义词实际上都是一个词具有一个以上的涵义的现象，它们都属于词义的歧义问题，我们一般没有必要区分同形词和多义词，我们把它们都作为词义排歧 (Word Sense

Disambiguation, 简称 WSD) 的问题来处理。

——同义关系

在传统语言学中, 如果两个词位具有相同的意义, 那么, 就说它们之间具有同义关系 (Synonymy)。这样的定义显然过于笼统, 缺乏操作性。

在自然语言处理研究中, 我们可以根据可替换性 (substitutability) 来定义同义关系: 在一个句子中, 如果两个词位可以互相替换而不改变句子的意思或者不改变句子的可接受性, 那么, 我们就说这两个词位具有同义关系。这样的定义显然具有可操作性。

例如, 句子“**How big is that plane?**”和句子“**Would I be flying on a large or small plane?**”中的 **big** 和 **large** 可以互相替换, 而不会改变这两个句子的意义或改变它们的可接受性, 我们就说 **big** 和 **large** 具有同义关系。

不过, 如果我们坚持这种可替换性一定要在一切的环境中都具有, 那么, 英语中的同义词的数量就很少了。因此, 我们对于可替换性的要求不能太过于严格, 只要求在某些环境下可替换就可以了。也就是说, 我们宁愿给同义关系一个比较弱的定义, 这样做比较现实。

可替换性与下面 4 个因素有联系:

- 多义关系中的某些涵义的有无

例如, 句子“**Miss Kim became a kind of big sister to Mrs. Park's son.**”是可以接受的, 而句子 ***“Miss Kim became a kind of large sister to Mrs. Park's son.”**就显得有些怪。其原因在于, 第一个句子中的 **big** 这个多义词的多个涵义中有 **older** 这个涵义, 而 **large** 这个多义词的多个涵义中, 没有 **older** 这个涵义, 因此, 在这样的环境下, **big** 和 **large** 不能相互替换。

- 微妙的意义色彩的差别

例如, 句子“**What is the cheapest first class fare?**”是可以接受的, 而句子 ***“What is the cheapest first class price?”**就显得有些怪。其原因在于, **fare** 比较适合于描述某些服务中需要支付的费用, 而 **price** 通常适合于描述票据的价格, 因此, 第二个句子中用 **price** 来替换 **fare** 就显得有些奇怪。

- 搭配约束的不同

例如, 句子“**They made a big mistake.**”是可以接受的, 而句子 ***“They made a large mistake.”**就显得有些怪。其原因在于, 当描述 **mistake** 比较严重时, 往往使用 **big** 而不用 **large**, 也就是说, **mistake** 倾向于与 **big** 搭配, 而不倾向于与 **large** 搭配。

- 使用域的不同: 使用域 (register) 是指语言使用中的礼貌因素、社会地位因素以及其他社会因素对于词语使用的影响。使用域的差别也会影响到同义词的选择。

在自然语言处理研究中, 同义词的意义色彩差别、搭配约束和使用域对于译文的质量有明显的影 响, 我们应该考虑到这些因素, 正确地选择恰当的同义词。

——上下位关系 (Hyponymy)

如果两个词位中, 一个词位是另一个词位的次类, 那么就说它们之间存在上下位关系 (hyponymy)。car(小汽车)和 vehicle(交通工具)间的关系就是一种上下位关系。上下位关系是不对称的, 我们把特定性较强的词位称为概括性较强的词位的下位词(hyponym), 把概括性较强的词位称为特定性较强的词位的上位词(hypernym)。因此, 我们可以说, car 是 vehicle 的下位词, 而 vehicle 是 car 的上位词。

我们可以使用受限的替换来探讨上下位关系的概念。

我们来考虑下面的蕴涵式

This is a X => That is a Y

在这个蕴涵式中, 如果 X 是 Y 的下位词, 则在任何情形下, 当左边的句子为真时, 右边新产生的句子也必须为真, 例如。我们有:

This is a car => That is a vehicle

在这里，新生成句子的目的并不是作为原句的替换，而仅仅是作为对是否存在上下位关系的一种诊断测试。所以，这只是一种受限的替换。

在知识本体 (ontology)、分类体系 (taxonomy)、对象层次 (object hierarchy) 中，上下位关系是很有用的。知识本体通常是指对一个领域或微世界 (micro-world) 进行分析而获得的概念系统的规范说明。分类体系是指把知识本体中的元素排列成包含结构的树状分类的一种特别的方式。计算机科学里的对象层次 (object hierarchy) 是基于这样的概念：知识本体中的对象被安排在一个分类体系中，并能够接受或继承分类体系中它们的祖先的特征。当然，只有当分类体系中的元素事实上是带有可继承特征的复杂结构的对象时这才会有意义。因此，上下位关系本身并不能组成一个知识本体、范畴结构、分类体系或对象层次。然而，上下位关系可以作为这类结构的近似表示。在我们下面讨论知识本体和词网 (WordNet) 的时候，都会涉及到上下位关系的概念。

2. 知识本体

如果我们对于一个领域中的客体进行分析，找出这些客体之间的关系，获得了这个领域中不同客体的集合，这一个集合可以明确地、形式化地、可共享地描述这个领域中各个客体所代表的概念的体系，它实际上就是概念体系的规范，这样的概念体系规范就可以看成这个领域的“知识本体” (ontology)。

人们很早就开始研究知识本体，因此，知识本体有很多不同的定义，这些定义有的是从哲学思辨出发的，有的是从知识的分类出发的，最近的一些定义则是从实用的计算机推理出发的。

牛津英语词典对于知识本体 (ontology) 的定义是：“对于存在的研究或科学” (the science or study of being)，这个定义显然是非常广泛的，因为它试图研究存在的一切事物，为存在的一切事物建立科学。不过，这个定义确实是关于知识本体的经典定义，它来自哲学研究。

什么是事物 (things)？什么是本质 (essence)？当事物发生改变时，本质是否仍然存在于事物之中？概念 (concept) 是否存在于我们的心智 (mind) 之外？怎样对世界上的实体 (entities) 进行分类？这些都是知识本体要回答的问题，所以，知识本体是“对于存在 (being) 的研究或科学”。

远古希腊时代，哲学家就试图研究当事物发生变化的时候，如何去发现事物的本质。例如，当植物的种子发育变成树的时候，种子不再是种子了，而树开始成为了树。那么，树还包含着种子的本质吗？巴门尼德 (Parmenides) 认为，事物的本质是独立于我们的感官的，种子在表面上虽然变成了树，但是，它的本质是没有改变的，所以，在实质上种子并没有转化为树，只不过是我们的感官原来感到它是种子，后来感到它是树。亚里士多德 (Aristotle) 认为，种子只不过是还没有完全长成的树，在发育过程中，树的本质并没有改变，只是改变了它存在的形式，从没有完成长成的树 (潜在的树) 变成了完全长成的树 (实在的树)。种子和树的本质都是一样的。知识本体就要研究关于事物的本质的问题。亚里士多德还把存在区分为不同的模式，建立了一个范畴系统 (system of categories)，包含的范畴有：substance (实体)，quality (质量)，quantity (数量)，relation (关系)，action (行动)，passion (感情)，place (空间)，time (时间)。这个范畴系统是最早的概念体系。

在中世纪，学者们研究事物本身和事物的名称之间的关系，分为唯实论 (realism) 和唯名论 (nominalism) 两派。唯实论主张，事物的名称就是事物本身，而唯名论主张，事物的名称只不过是引用事物的词而已。在中世纪晚期，大多数学者都倾向于认为，事物的名称只是表示事物的符号 (symbol)，例如，book 这个名称只不过是用来引用一切作为实体的“书”的一个符号。这是现代物理学的一个起点，在现代物理学中，采用不同符号来表示物理世界

的各种特征（如，速度的符号为 V，长度的符号为 L，能量的符号为 E，等）。这些用符号表示的特征，实际上都是物理学中的概念或范畴。

1613 年，德国哲学家郭克兰纽（R. Goclenius）在他用拉丁文编写的《哲学辞典》中，把希腊语的 on（也就是 being）的复数 onta（也就是 beings）与 logos（含义为“学问”）结合在一起，创造出 ontologia 这个术语。ontologia 也就是英文的 ontology，这是西方文献中最早出现的 ontology 这个术语。1636 年，德国哲学家卡洛维（A. Calovius）在《神的形而上学》中，把 ontologia 看成“形而上学”（metaphysica；英文为 metaphysics）的同义词，这样，他便把“ontologia”与亚里士多德的“形而上学”紧密地联系起来。法国哲学家笛卡尔（R. Descartes）更是明确地把研究本体的第一哲学叫做“形而上学的 ontologia”，这样，ontologia 便成为形而上学的一个部分了。德国哲学家莱布尼兹（G. von Leibniz）和他的继承者沃尔夫（C. Wolff）更是从学科分类的角度，把 ontologia 归属为形而上学的一个分支，使 ontologia 成为了哲学中一个相对独立的分支学科。ontologia 这个术语，在哲学中翻译为“本体论”，在自然语言处理中，从应用的角度出发，我们认为翻译为“知识本体”更为恰当。因此，在本文中，我们统一地使用“知识本体”这个术语。

德国哲学家康德（Emmanuel Kant）也研究知识本体，他认为，事物的本质不仅仅由事物本身决定，也受到人们对于事物的感知或理解的影响。康德提出这样的问题：“我们的心智究竟是采用什么样的结构来捕捉外在世界的呢？”为了回答这个问题，康德对范畴进行了分类，建立了康德的范畴框架，这个范畴框架包括 4 个大范畴：quantity（数量），quality（质量），relation（关系），modality（模态）。每一个大范畴又分为 3 个小范畴。Quantity 又分为 unicity（单量），plurality（多量），totality（总量）3 个范畴；quality 又分为 reality（实在质），negation（否定质），limitation（限度质）3 个范畴；relation 又分为 inherence（继承关系），causation（因果关系），community（交互关系）3 个范畴；modality 又分为 possibility（可能性），existence（现实性），necessity（必要性）。根据这个范畴框架，我们的心智就可以给事物进行分类。从而获得对于外界世界的认识。例如，本文作者冯志伟属于的范畴是：unicity, reality 和 existence，这样，我们就认识到：冯志伟是一个“单一的、实在的、现实的”人。在数据库中，我们可以根据康德的方法给事物建立一些范畴，从而根据这些范畴来管理数据。例如，我们给人事管理数据库建立“姓名，性别，籍贯，职业”等范畴，使用这些范畴进行人事管理。可以看出，康德对于范畴框架的研究，为知识本体的研究奠定了坚实的基础。

1991 年，美国计算机专家尼彻斯（R. Niches）等在完成美国国防部高级研究计划局（Defense Advanced Research Projects Agency, 简称 DARPA）的一个关于知识共享的科研项目中，提出了一种构建智能系统方法的新思想，他们认为，构建的智能系统由两个部分组成，一个部分是“知识本体”（Ontology），一个部分是“问题求解方法”（Problem Solving Methods, 简称 PSMs）。知识本体涉及特定知识领域共有的知识和知识结构，它是静态的知识，而 PSMs 涉及在相应知识领域进行推理的知识，它是动态的知识，PSMs 使用知识本体中的静态知识进行动态的推理，就可以构建一个智能系统。这样的智能系统就是一个知识库，而知识本体是知识库的核心，这样，知识本体在计算机科学中就引起了学者们的极大关注。

1990 年，我国学者冯志伟提出，在机器翻译系统中，要把静态标记和动态标记结合起来，静态标记要表示存储在机器词典中的单词的词类特征和单词固有的语义特征，它们是与单词所在的上下文语境无关的，动态标记是使用静态标记经过计算机运算求出来的句法功能标记、语义关系标记、逻辑关系标记，它们是要根据单词的上下文语境来确定的。静态信息的制定要根据词类和语义系统的规范，动态标记的求解要根据产生式规则，产生式规则的基本形式是“条件-动作”偶对，因此，面向机器翻译的语言学研究要着重阐明规则的条件。冯志伟所说词类规范，实际上就是语法信息的规范，冯志伟所说的语义系统的规范，实际上

就是概念系统的规范，也就是“知识本体”。

冯志伟的构想可以表示为：

基于语法信息和知识本体的静态标记标注的机器词典 + 基于产生式规则的动态标记求解规则 = 机器翻译系统

尼彻斯的构想可以表示为：

静态的“知识本体” + 动态的“问题求解方法” = 知识库

通过比较可以看出：冯志伟关于静态标记与动态标记相结合的构想，与尼彻斯关于静态的“知识本体”与动态的“问题求解方法”相结合的构想是非常相似的。

在 20 世纪末和 21 世纪初，知识本体的研究开始成为计算机科学的一个重要领域。它主要任务是研究世界上的各种事物（例如，物理客体、事件等）以及代表这些事物的范畴（例如，概念、特征等）的形式特性和分类。计算机科学对于知识本体的研究当然是建立在上述的经典的知本体的研究的基础之上的，不过，有了很大的发展。因此，我们有必要重新给知识本体下定义。下面，我们介绍在计算机科学中对于知识本体的定义。

在人工智能研究中，格鲁伯（Gruber）在 1993 年给知识本体下的定义是：

“知识本体是概念体系的明确规范”（An ontology is an explicit specification of conceptualization）。

这个定义比较具体，也比较便于操作，在知识本体的研究中广为传布。

1997 年，波尔斯特（Borst）对格鲁伯的定义做了很小修改；提出了如下的定义：

“知识本体是可以共享的概念体系的形式规范”（Ontologies are defined as a formal specification of a shared conceptualization）。

1998 年，施图德（Studer）等在格鲁伯和波尔斯特的定义的基础上，对于知识本体给出了一个更加明确的解释：

“知识本体是对概念体系的明确的、形式化的、可共享的规范”（An ontology is a formal explicit specification of a shared conceptualization）。

在这个定义中，所谓“概念体系”是指所描述的客观世界的现象中有关概念的抽象模型；所谓“明确”是指对于所使用的概念的类型以及概念用法的约束都明确地加以定义；所谓“形式化”是指这个知识本体应该是机器可读的；所谓“共享”是指知识本体中所描述的知识不是个人专有的而是集体共有的。

具体地说，如果我们把每一个知识领域抽象成一个概念体系，再采用一个词表来表示这个概念体系，在这个词表中，要明确地描述词的涵义、词与词之间的关系、并在该领域的专家之间达成共识，使得大家能够共享这个词表，那么，这个词表就构成了该领域的一个知识本体。知识本体已经成为了提取、理解和处理领域知识的工具，它可以被应用于任何具体的学科和专业领域，知识本体经过严格的形式化之后，借助与计算机强大的处理能力，可以对于人类的全部知识进行整理和组织，使之成为一个有序的知识网络。

人们对于知识本体的认识可能存在差别，因此，有不同类型的知识本体。

• 通用知识本体（common ontology）常常从哲学的认识论出发，概念的根结点往往是很抽象的，例如，时间、空间、事件、状态、对象等。

• 领域知识本体（domain ontology）对领域的知识进行抽象，概念比较具体，容易进行形式化和共享。例如，我国学者最近研制的生物学领域知识本体（domain-specific ontology of botany）、考古学领域知识本体（domain-specific ontology of archeology）都是领域知识本体。

• 语言知识本体（language ontology）常常表现为一个词表，其中要描述单词和术语之间的概念关系，词网（WordNet）就是一个语言知识本体。如果语言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。

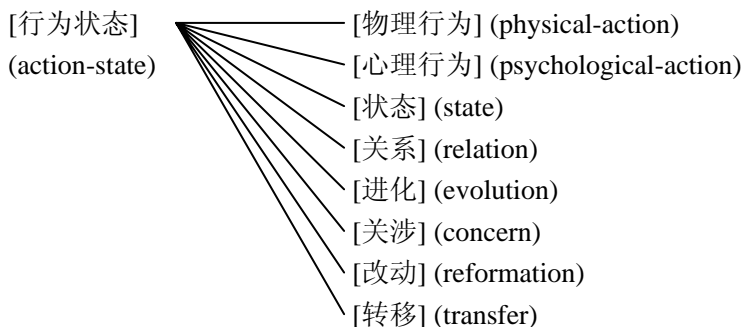
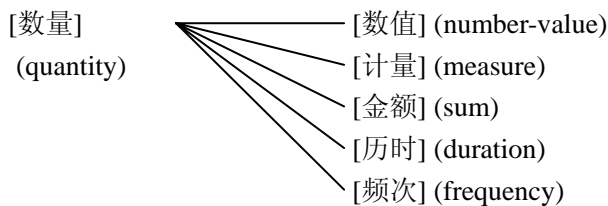
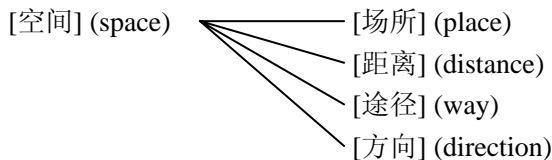
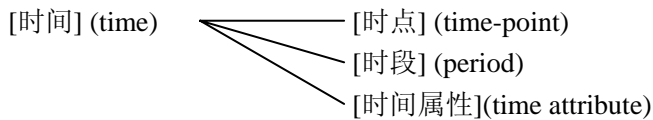
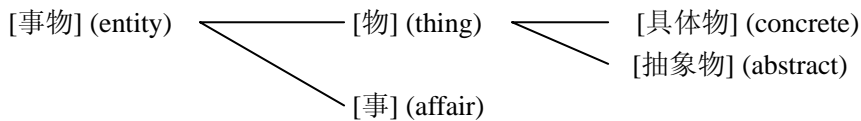
• 形式知识本体 (formal ontology) 对于概念和术语的分类很严格, 要按照一定的原则和标准, 明确地定义概念之间的显性和隐性关系, 明确概念的约束和逻辑联系。领域知识本体或术语知识本体经过进一步的抽象和提炼, 就可能发展成形式知识本体。

知识本体可以帮助我们对于领域知识进行系统的分析, 把领域知识形式化, 使之便于计算机处理。知识本体还可以实现人和人之间以及人和计算机之间知识的共享, 实现在一定领域中知识的重复使用。在自然语言处理的语义分析中, 知识本体可以给我们提供单词的各种信息, 帮助我们揭示单词之间的各种语义关系, 是语义分析的知识来源。

目前, 支持知识本体的开发工具已经有数十种, 功能各不相同, 对于知识本体语言的支持能力、表达能力各有差别, 可扩展性、灵活性、易用性也不一样。其中比较著名的有 Protégé-2000、OntoEdit、OilEd、Ontolingua 等。Protégé-2000 是使用比较广泛的知识本体工具, 是可以免费获得的开放软件, 它用 Java 语言开发, 通过各种插件支持多种知识本体格式。

本文作者冯志伟在日汉机器翻译的研究中, 设计了一个知识本体系统 ONTOL-MT。这个知识本体的初始概念有事物 (entity)、时间 (time)、空间 (space)、数量 (quantity)、行为状态 (action-state) 和属性 (attribute) 6 个。在这 6 个初始概念之下, 还有不同层次的下位概念。

ONTOL-MT 的基本结构如下:



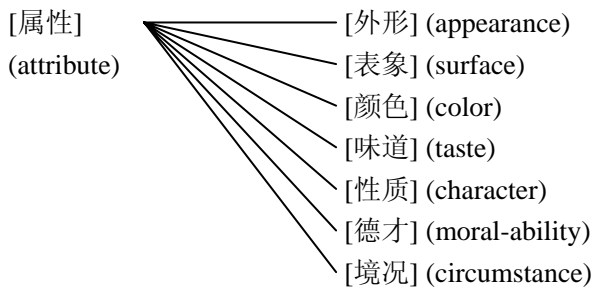


图 知识本体 ONTOL-MT

ONTOL-MT 中的上述主要概念的涵义定义如下：

- [事物] (entity) 在空间（包括思维空间）上和时间上延展的事物本体。
- [物] (thing) 主要在空间（包括思维空间）上延展的事物本体。
- [具体物] (concrete) 有形、有色、有质量的物。
- [抽象物] (abstract) 无形、无色、无质量的物。
- [事] (affair) 主要在时间上延展的事物本体。包括人类生活中的一切活动和所遇到的一切社会现象（政治、军事、法律、经济、文化、教育）或与人有关联的自然现象。
- [时间] (time) 由过去、现在和将来构成的连绵不断的系统，它是物质运动和变化的持续性的表现，是物质存在的一种客观形式。
- [时点] (time-point) 指时间里的某一点。
- [时段] (period) 指有起点和终点的一段时间。
- [时间属性] (time-attribute) 时间所具有的属性（年、月、日、小时、分、秒、毫秒等）。
- [空间] (space) 事物及其运动存在的另一种客观形式，它在不同的维度上延伸。
- [场所] (place) 由长度、宽度和高度表现出来的物质存在的一种客观形式，也就是活动的处所。
- [距离] (distance) 在空间或者时间上相隔。
- [途径] (way) 两地之间的通道。
- [方向] (direction) 指东、南、西、北、上、下等。
- [数量] (quantity) 事物的多少与计量。
- [数值] (number-value) 一个用数目表示出来的量的多少。
- [计量] (measure) 温度，长度，重量，使用量等。
- [金额] (sum) 钱数的多少。
- [历时] (duration) 时间在数量上的长短。
- [频次] (frequency) 事情发生的频繁程度的大小。
- [行为状态] (action-state) 人或事物表现出来的活动和形态。
- [物理行为] (physical-action) 人或事物在物理上表现出来的活动。
- [心理行为] (psychological-action) 人或动物在心理上表现出来的活动。
- [状态] (state) 人或事物表现出来的形态。
- [关系] (relation) 事物之间相互作用相互影响的状态。
- [进化] (evolution) 事物由简单到复杂、由低级到高级的变化。
- [关涉] (concern) 一事物关联或牵涉到另一事物。

[改动] (reformation) 使事物发生变化或者差别。

[转移] (transfer) 使事物从一方改动到另一方。

[属性] (attribute) 事物所具有的特性和关系。

[外形] (appearance) 人或事物的外部形体属性。

[表象] (surface) 从外表可以观察到的现象的属性。

[颜色] (color) 由物体发射、反射或者透过的光波通过视觉所产生的印象。

[味道] (taste) 能使舌头得到某种味觉的特性。

[性质] (character) 一个事物区别于另一个事物的属性。

[德才] (moral-ability) 人的道德和才能表现出来的属性。

[境况] (circumstance) 外界环境所具有的属性。

这里只是列出了 ONTOL-MT 中主要的上层概念，在这些概念的下层，还有很多其他的概念，限于篇幅，这里就不列出了。可以看出，ONTOL-MT 中的初始概念与亚里士多德的范畴系统中的范畴很接近，明显地受到了亚里士多德的范畴系统的影响。这个知识本体也反映了我们的世界观：万事万物都是在时间和空间中运动和存在的，它们都具有一定的属性和数量。所以，ONTOL-MT 虽然是为机器翻译的技术而设计的，但是，它继承了亚里士多德的范畴系统，反映了我们的世界观，具有鲜明的人文性。

ONTOL-MT 知识本体系统中的概念，实际上也就是单词本身所固有的语义特征，它们是独立于单词的上下文而存在的，因此，可以用这些概念来表示机器翻译词典中单词的固有语义特征。在日汉机器翻译中，我们利用单词固有的这些语义特征在机器翻译系统中进行日语分析中同形词的判别，效果良好。

在日语中，“きしゃ”是一个同形词，从机器翻译的角度看，它也是一个多义词，它可以有“记者、火车、回公司”等不同的涵义。在日语句子“きしゃはきしゃできしゃした”中有三个“きしゃ”，而且每一个“きしゃ”的含义各不相同，这里，为了表达上的方便，我们把第一个“きしゃ”记为“きしゃ1”，第二个“きしゃ”记为“きしゃ2”，第三个“きしゃ”记为“きしゃ3”；“は”是表示主语的提示助词，“で”是表示方式的助词，“した”是处于动词之后表示过去时态的助动词，这样，我们的句子可以记为如下的形式：

“きしゃ1 は きしゃ2 で きしゃ3 した”

在机器词典中，我们存储如下的信息：

如果“きしゃ”的语义特征是 [HUMAN]，则汉语译文为“记者”；

如果“きしゃ”的语义特征是 [VEHICLE]，则汉语译文为“火车”；

如果“きしゃ”的语义特征是 [MOVEMENT]，则其汉语译文为“回公司”，并且其有如下的语义框架：

“[HUMAN] は [VEHICLE] で [MOVEMENT]”。

这里的[HUMAN]、[VEHICLE]和[MOVEMENT]等语义特征都是 ONTOL-MT 知识本体中的概念。

在我们的句子中，助动词“した”在“きしゃ3”之后，所以“きしゃ3”必定是中心动词，它的语义范畴必定是 [MOVEMENT]，而它的汉语译文应该是“回公司”。

把我们的句子

“きしゃ1 は きしゃ2 で きしゃ3 した”

同“きしゃ3”的语义框架

“[HUMAN] は [VEHICLE] で [MOVEMENT]”

相比较，我们可以得到如下的认识：

——“きしゃ1”在は之前，它的语义范畴是 [HUMAN]，所以，它的汉语译文应该是“记者”，因而“きしゃは”应该是中心动词“きしゃ3”的主语。

——“きしゃ 2”在で之前,它的语义范畴是 [VEHICLE], 所以, 它的汉语译文应该是“火车”, 因而“きしゃ 2 で”应该是中心动词“きしゃ 3”的方式状语。

经过以上的分析我们可以得到三个“きしゃ”的正确的汉语译文, 再经过结构转换和汉语生成, 最后我们就可以得到这个句子的汉语译文: “记者乘火车回公司”。

由此可见, ONTOL-MT 知识本体中的语义特征对于区分同形词和辨别歧义是非常有用的。这样的语义特征信息还可以用在语音识别和语音机器翻译中。知识本体的研究和设计是自然语言处理的基础性工程之一, 我们应该给予足够的重视。

下面我们从知识本体的角度出发来介绍一个著名的语言知识本体——词网 (WordNet)。

3. 词网

词网 (WordNet) 是英语的词汇关系数据库, 从知识本体的角度来看, 词网是一个语言知识本体。词网是 1985 年由美国普林斯顿大学 (Princeton University) 的米勒 (G. A. Miller)、贝克威克 (R. C. Beckwick)、费尔鲍姆 (C. Fellbaum) 等研制的, 可以在因特网上访问, 网址: [http:// www.cogsci.princeton.edu/~wn/](http://www.cogsci.princeton.edu/~wn/)

3.1 词网的 3 个基本假设和规模

为了建造词网, 米勒等提出了如下 3 个基本假设:

——分离性假设 (separability hypothesis): 语言中的词汇成分可以从语言中分离出来, 单独地进行研究。

——模式化假设 (patterning hypothesis): 人们倾向于特别关注词语所表达的涵义之间的系统模式和关系。

——完全性假设 (comprehensiveness hypothesis): 系统需要尽可能地把人们的词语知识存储在词网中。

这意味着, 米勒等试图把词语从语言中分离出来用模式化的方法进行研究, 研究时, 尽量完全地收集词语的知识。

尽管词网中包含合成词、短语、惯用语和搭配关系描述, 但是, 词网的基本单位还是单词。词网包括动词、名词、形容词-副词 3 个数据库。词网中一个完全的涵义条目包含单词、同义词、定义以及一些使用实例。

在词网中不区分同形关系与多义关系, 同形词也就是多义词, 一个多义词可以有若干个不同的涵义。因此, 词网中涵义的数量比单词的数量大。

词网中单词及其涵义的数量是相当可观的。词网 1.6 (WordNet 1.6) 的规模如下:

| 范畴 | 单词数 | 涵义数 |
|-----|-------|--------|
| 名词 | 94474 | 116317 |
| 动词 | 10319 | 22066 |
| 形容词 | 20170 | 29881 |
| 副词 | 4546 | 5677 |

表

3.2 词网中的名词

由于存在多义关系, 词网中的 94 474 个名词可以表示 116 317 个涵义 (词汇化的概念)。

词网中的基本语义关系是同义关系。词网中同义词的基本原则与我们前面讲述是相同的。如果词网中的两个条目在某些上下文环境能够成功地进行替换, 则认为它们是同义词。同义词的集合构成了同义词集, 叫做 SYNSET。下面是 SYNSET 的一个例子:

{chump, fish, fool, gull, mark, patsy, fall guy, sucker, schlemiel, shlemiel, soft touch, mug}

这个 SYNSET 的定义是：“易受骗和易被利用的人”（a person who is gullible and easy to take advantage of）。因此，在这个 SYNSET 中的每个词条都可以在一些场景下表达这个概念。实际上，词网中许多条目的涵义都是由这类 SYNSET 组成的。具体地说，这样的 SYNSET 及其定义和例句，构成了 SYNSET 中所列条目的涵义。

从一个更理论化的观点看，每个 SYNSET 都可以表示语言中已经词汇化的一个概念。不过，词网不是用逻辑项来表示概念，而是通过把可用于表达概念的词典条目组成列表来表示概念。这种观点引出这样一个事实，正是 SYNSET，而不是词典条目或单个的涵义，参与了词网的名词中的大部分语义关系。这里我们讨论的各种语义关系，实际上都是 SYNSET 之间的关系。为了表达上的方便，我们一般只用 SYNSET 中的有代表性单词来表示 SYNSET。

下面我们讨论名词中的 3 种语义关系：上下位关系、部分—整体关系、反义关系。

——上下位关系：词网中的上下位关系与我们前面讨论过的上下位关系直接对应。特定性较强的单词叫做概括性较强的单词的下位词(hyponym)，概括性较强的单词叫做特定性较强的单词的上位词 (hypernym)。

例如，bird（鸟）是 robin（知更鸟）的上位词，robin 是 bird 的下位词。

根据上下位关系，我们可以把名词组织到一个词汇的层级体系中。

例如，根据词网中的定义，robin 是“一种会唱歌的候鸟，它的胸部为红色，背部和翅膀为灰黑色”（a migratory bird that has a clear melodious song and a reddish breast with gray or black upper plumage）。因此，robin 的上位词是 bird。

而 bird 在词网中的定义是：“一种热血的、会生蛋的动物，有羽毛，前肢变成了翅膀（a warm-blooded egg-laying animal might having feathers and forelimbs modified as wing）。因此，bird 的上位词是 animal（动物）。

词网中对 animal 的定义是：“能够主动地运动的生物体，有感觉器官，细胞壁不是纤维素的”（an organism capable of voluntary movement and possessing sense organs and cells with non-cellulose walls）。因此，animal 的上位词是 organism（生物体）。

词网中对 organism 的定义是：“有生命的物体”（a living entity）。

可以看出，在这样的上下位关系中，每一个单词代表了一个 SYNSET，每个 SYNSET 通过上位关系和下位关系与紧靠的更普遍化或更具体化的 SYNSET 相关联。为了找到一系列更普遍化或更具体化的 SYNSET，我们可以简单地跟随一个上位和下位关系的传递链往上查询或者往下查询。

应该注意的是，上下位关系表示的是单词所代表的某个特定的涵义之间的关系，它并不表示具体的单词形式（word form）之间的关系。例如，当我们说“tree”（树）是一种“plant”（植物）的时候，我们指的是涵义为“树”的 tree 和涵义为“植物”的 plant 之间的关系，并不是指 tree 的其他涵义和 plant 的其他涵义之间的关系，因此，我们说的并不是“树形图”（tree graph）和“工厂”（manufacturing plants）之间的关系。

因此，上下位关系是单词的特定涵义之间的关系，它代表的是词汇化的概念之间的关系。在词网中，上下位关系用指针“@->”把相应 SYNSET 联系起来表示。例如，我们有：

```
{robin, redbreast} @-> {bird} @-> {animal, animate_being} @-> organism, life_form, living_thing}
```

从数学上说，@ 是传递的，非对称的。它表示的语义关系可以读为“IS-A”或“IS-A-KIND-OF”。

“->”读为“指向”（to point upward）。

当由概括性较弱的涵义指向概括性较强的涵义时，叫做普遍化（generalization），也就是从特殊（specific）指向一般（generic），用“@->”表示，写为：Ss @-> Sg。

当由概括性较强的涵义指向概括性较弱的涵义时，叫做具体化（specification），也就是

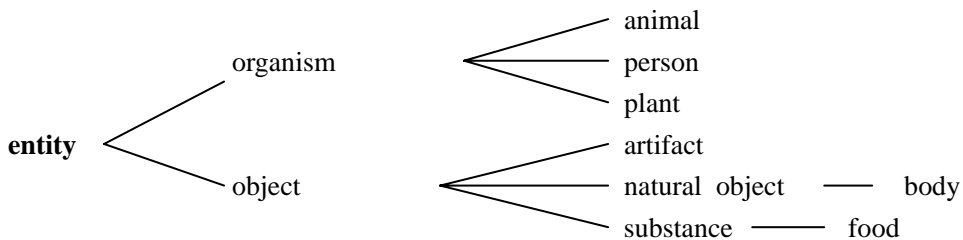
从一般 (generic) 指向特殊, 用 “~>” 表示, 写为: Sg ~> Ss。

在上下位关系中, 概念的特性可以继承 (inheritance), 因此, 我们就可以用上下位关系进行推理。例如, 如果 Rex 是一个 collie (牧羊犬), 那么, Rex 就是一个 dog (狗); 如果 Rex 是一个 dog, 那么, Rex 就是一个 animal (动物); 如果 Rex 是一个动物, 那么, Rex 就能够主动地运动 (capable of voluntary movement)。这样, 上下位关系可以形成传递链, 一步一步地把概念普遍化。当到达最普遍的概念的时候, 这样的概念就是语义的基元 (Primitive semantic component), 在词网中叫做初始概念 (Unique Beginners)。

词网的名词数据库中使用了 25 个初始概念。它们是:

- {act, activity} (活动)
- {animal, fauna} (动物, 动物群)
- {artifact} (人工物)
- {attribute} (属性)
- {body} (躯体)
- {cognition, knowledge} (认知, 知识)
- {communication} (交际)
- {event, happening} (事件)
- {feeling, emotion} (感觉, 情感)
- {food} (食物)
- {group, grouping} (集体)
- {location} (位置)
- {motivation, motive} (动机)
- {natural object} (自然物)
- {natural phenomenon} (自然现象)
- {person, human being} (人, 人类)
- {plant flora} (植物, 植物群)
- {possession} (所属)
- {process} (过程)
- {quantity, amount} (数量)
- {relation} (关系)
- {shape} (外形)
- {state} (状态)
- {substance} (实体)
- {time} (时间)

后来, 词网又对这 25 个初始概念进行归纳和整理, 形成了如下的 11 个初始概念 (用粗体字表示):



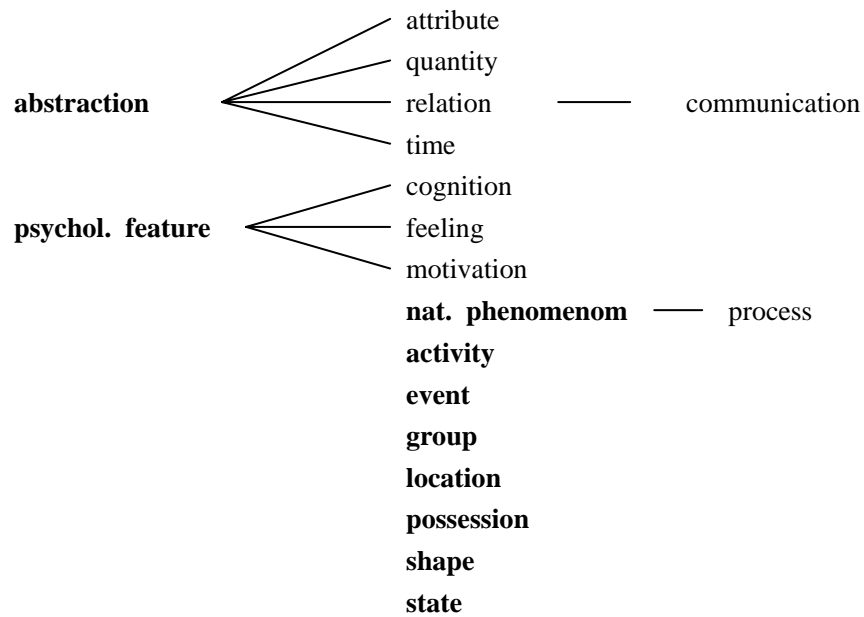


图 词网的初始概念

经过整理之后的 11 个初始概念是：entity（实体），abstraction（抽象），psychological feature（心理特征），natural phenomenon（自然现象），activity（活动），event（事件），group（集体），location（位置），possession（所属），shape（外形），state（状态）。

“bass”这个词位有两个不同涵义：“涵义 3”和“涵义 7”。下面是“bass”的这两个涵义上下位关系的传递链。注意，这两个传递链是完全分别开来的，但是它们在初始概念“实体”（entity）处汇集在一起了。

Sense 3（涵义 3）

bass, basso

(an adult male singer with the lowest voice)（成年的男低音歌唱家）

=> singer, vocalist（歌唱家）

=> musician, instrumentalist, player（音乐家，演奏家）

=> performer, performing artist（表演艺术家）

=> entertainer（演艺人员）

=> person, individual, someone（人）

=> life form, organism, being（生物体）

=> **entity**, something（实体）

=> causal agent, cause, causal agency（作为导因的人或事物）

=> **entity**, something（实体）

Sense 7（涵义 7）

bass –

(the number with the lowest range of a family of musical instruments)（低音乐器）

=> musical instrument（乐器）

=> instrument（工具）

=> device（设备）

=> instrumentality, instrumentation (设施)
=> artifact, artefact (人工物)
=> object, physical object (物理客体)
=> **entity**, something (实体)

在第一个传递链中，链的开始是“男低音歌唱家”，它的上位词是“歌唱家”这个更为一般的概念，再上位的概念顺次是“音乐家”、“表演艺术家”、“演艺人员”、“人”、“生物体”、“实体”。第二个传递链从“低音乐器”开始，顺着完全不同的链，顺次经过“乐器”、“工具”、“设备”、“设施”、“人工物”、“物理客体”等概念，最后也到达初始概念“实体”。这两个传递链顺着不同的路径殊途同归。在概念的层级系统中，“实体”处于最顶端的位置，它是词网的 11 个初始概念之一。

——部分—整体关系

词网名词数据库中的部分和整体之间构成的关系叫做部分—整体关系(meronymy)。“meros”来自希腊语，它的意思是“部分”。

在部分—整体关系中，表示“部分”(part)的词叫做“部分词”(meronym)，记为 S_m ，表示“整体”(whole)的词叫做“整体词”(holonym)，记为 S_h 。显而易见，如果 S_m 是 S_h 的部分词，那么， S_h 就是 S_m 的整体词。

在词网中，也用 W_m 和 W_h 分别表示部分词和整体词，用“is a part of”(“是一部分”)和“has a”(“有...作为一部分”)来描述部分—整体关系的语义。如果' W_m is a part of W_h ' (如果 W_m 是 W_h 的一部分)是可接受的，那么我们就说' W_m is a meronym of W_h ' (W_m 是 W_h 的部分词); 如果' W_h has a W_m (as a part)' (W_h 有 W_m 作为一部分)是可接受的，那么我们就说' W_h is a holonym of W_m ' (W_h 是 W_m 的整体词)。

部分—整体关系与上下位关系在数学方面的特性很相似：它们都是可传递的，非对称的。例如，finger(指头)是hand(手)的一部分，hand(手)是arm(胳膊)的一部分，arm(胳膊)是body(躯体)的一部分。

根据温斯顿(Winston)和切芬(Chaffin)在1987年的研究，部分—整体关系可以分为如下6种类型：

组成成分—客体(component-object): 例如，branch(树枝) - tree(树)。

成员—集体(member-collection): 例如，tree(树) - forest(森林)。

局部—物质(portion-mass): 例如，slice(一片蛋糕) - cake(蛋糕)。

材料—客体(stuff-object): 例如，aluminum(铝) - airplane(飞机)。

特征—活动(feature-activity): 例如，paying(支付) - shopping(购物)。

地点—地域(place-area): 例如，Princeton(普林斯顿) - New Jersey(新泽西州)

在词网中，仅仅使用了3种关系：组成成分—客体关系、成员—集体关系、材料—客体关系，分别用#p->、#m->、#s->来表示。具体地说，

“ W_m #p-> W_h ”表示“ W_m 是 W_h 的组成成分”；

“ W_m #m-> W_h ”表示“ W_m 是 W_h 的成员”；

“ W_m #s-> W_h ”表示“ W_m 是制造 W_h 的材料”。

——反义关系

相反或对立的词之间的关系，叫做反义关系(Antonymy)。涵义彼此相反并不是名词之间的一种基本的意义组织方式，可是，反义关系在词网中还是存在的，所以，我们也要说明一下它的表示方法。

反义关系用“!->”表示。例如，

[{man} !-> {woman}]

表示 man 是 woman 的反义词；

[{woman} !-> {man}]

表示 woman 是 man 的反义词。

具有反义关系的名词的上位词往往是相同的，它们通常具有一个直接上位词。例如，man 和 woman 的直接上位词是 human（人）。

3.3 词网中的形容词

词网中的 20 170 个形容词组织到 29 881 个涵义（词汇化的概念）中。

在词网中凡是修饰名词的词都看成形容词。因此，除了通常的形容词之外，名词、现在分词、过去分词、介词短语、小句（clause）都算形容词。例如，短语“a *large chair*, a *comfortable chair*”中的 large 和 comfortable 是形容词，在词网中，当然也算形容词。但是，在下面例子中的用斜体字标出的词、短语或小句，在词网中也都算为形容词。

kitchen chair, *barber chair* (原来是名词)

The *creaking chair* (原来是现在分词)

The *overstuffed chair* (原来是过去分词)

Chair *by the window* (原来是介词短语)

The chair *that you bought at the auction* (原来是小句)

词网的 16 428 个形容词 SYNSET 中包含了很多的形容词、分词和介词短语。

形容词可以分为描写形容词（descriptive adjective）和关系形容词（relational adjective）两种。

• 描写形容词：例如，big, beautiful, interesting, possible, married 等。

描写形容词可以给被它修饰的名词赋上一个属性值。“X is Adj”意味着，存在着一个属性 A 使得 A(X) = Adj。例如，“the package is heavy”意味着，存在着一个属性 WEIGHT（重量）使得 WEIGHT(package) = heavy。“heavy”或“light”是属性 WEIGHT 的值。词网中使用一个指针把描写形容词与它所修饰的名词联系起来。

• 关系形容词：例如，electrical。

关系形容词是由名词派生而来的，因此，关系形容词和派生它的名词之间是有联系的。例如，关系形容词 electrical 与名词 electricity 有联系。

描写形容词之间的基本语义关系是反义关系（antonymy）。例如，

good—bad

描写形容词有两个显著的特征：一个显著特征是属性的两极性（bipolar），另一个显著特征是属性的分级性（gradeness）。分述如下：

——两极性：描写形容词的属性具有两极化的倾向。

反义形容词表示的属性是彼此对立的。例如，heavy 的反义词是 light，它们表示 WEIGHT（重量）这个属性的彼此对立的两极的值。

在词网中，这种两极对立用符号“!->”表示，它的含义是“IS-ANTONYMOUS-TO”。例如，“heavy (vs. light)”和“light (vs. heavy)”可以分别表示为：

heavy !-> light

light !-> heavy

如果一个单词具有两个不同的涵义，就把它作为两个不同的词形（word form）来处理。在词网中，同一个单词的不同的词形标以不同的数字。例如，hard 有“坚硬”和“困难”两个不同的涵义，涵义为“坚硬”的 hard 写为 hard1，涵义为“困难”的 hard 写为 hard2，ha

rd1 的反义词是“soft”（柔软），hard2 的反义词是“easy”（容易）。

在英语中，如像“heavy/light”，“weighty/weightless”这样直接对立的反义词叫做直接反义词（antonym）。此外还存在着间接反义词（indirect antonym）。例如，“ponderous”（笨重）这个词很难说出它的反义词是什么，但是，“ponderous”的涵义与“heavy”的涵义很近似，所谓“涵义近似”，是因为凡是能够被“heavy”修饰的名词，也能够被“ponderous”修饰；而“heavy”的反义词是“light”，所以，我们可以通过“heavy”的中介，近似地把“light”看成是“ponderous”的反义词，可见，“ponderous/light”这一对概念的对立是通过“heavy”的中间而建立起来的，它们不是直接反义词，它们之间的反义关系是间接的，所以，我们把“ponderous/light”叫做间接反义词。从词汇的角度说，“ponderous/light”不是对立的词汇偶对，但是通过“heavy”的中介，我们可以在概念上给它们建立反义关系。在词网中，“涵义近似”的意思是“IS SIMILIAR TO”，用指针“&->”表示，这样，间接反义词的推理过程是：

由于 heavy !-> light 而且 ponderous &-> heavy.

所以我们有：ponderous !-> light

按照这样的办法，我们就可能给英语中所有的描写形容词都找到反义词，对于那些很难判定反义词是什么的形容词，我们也可以给它们找到间接反义词。

这样一来，我们就有可能把直接反义词和间接反义词组织到“两极聚类”（bipolar cluster）中。下面是一个两极聚类：

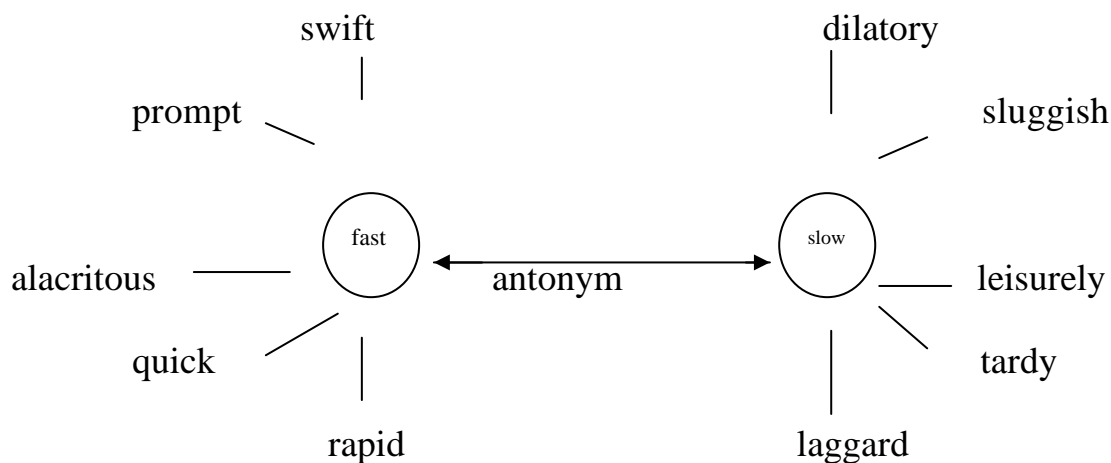


图 两极聚类

在这个两极聚类中，“中心词 SYNSET”（head SYNSET）是“fast/slow”，共包含两半聚类，一半聚类以 fast 为中心词，另一半聚类以 slow 为中心词，在中心词周围是“涵义近似”的单词，它们构成了“卫星词 SYNSET”（satellite SYNSET），fast 的卫星词是 swift, prompt, alacritous, quick, rapid，它们都具有“快”的涵义，slow 的卫星词是 dilatory, sluggish, leisurely, tardy, laggard，它们都具有“慢”的涵义。这个两极聚类确定了“SPEED”（速度）这个属性。

在两极聚类中的反义词偶对表示相同的涵义或紧密相关的涵义，它们可以代表一个属性值。例如，“large/small”和“big/little”这样的反义词偶对确定了“SIZE”（大小）这个属性；在词网中，这个两极聚类的一半表示为 large (vs. small), big (vs. little)，另一半表示为 small (vs. large), little (vs. big)。

词网的形容词数据库中包含 1 732 个这样的两极聚类, 每一个两极聚类的两侧的单词都具有反义关系, 也就是说, 每一个单词都有其相应的反义词, 而两极聚类的每一侧有 1 732 个“近似涵义”, 如果考虑到两极聚类两侧的所有的“近似涵义”, 那么, 词网形容词数据库中“近似涵义”聚类的总数就应该是 3 464 个。这些“近似涵义”聚类的数目也就是词网中形容词涵义的大致数目。

——分级性 (gradeness): 词网中的形容词可以按不同的属性进行分级。例如, 可以按 SIZE (大小), LIGHTNESS (亮度), QUALITY (质量), BODY-WEIGHT (体重) 和 TEMPERATURE (温度) 对形容词进行如下的分级:

| SIZE | LIGHTNESS | QUALITY | BODY-WEIGHT | TEMPERATURE |
|---------------|-------------|-----------|-------------|-------------|
| Astronomical | snowy | superb | obese | torrid |
| huge | white | great | fat | hot |
| large | ash-gray | good | plump | warm |
| ... | gray | mediocre | ... | tepid |
| small | charcoal | bad | slim | cool |
| tiny | black | awful | thin | cold |
| infinitesimal | pitch-black | atrocious | gaunt | frigid |




图 形容词的分级性

根据这样的分级, 可以看出形容词涵义近似的程度, 而形容词表示的属性也就会因此而显示出方向性, 而方向性也就是维度 (dimension), 所以, 我们可以把词网中的形容词想象成一个具有多个维度的超空间 (hyperspace), 其中, 每一个维度的一端紧紧地嵌在这个多维超空间的一个原点上。

关系形容词在语义上或形态上与名词有联系, 尽管关系形容词与名词形态上的联系还不是很直接的。

例如, “musical” (音乐的) 与名词 “music” (音乐) 有关; “dental” (牙科的) 与名词 “tooth” 有关。

因此, 名词常常可以用关系形容词或者该关系形容词所由派生的名词来修饰。例如, 关系形容词 + 名词 : 名词 + 名词

‘atomic bomb’ : ‘atom bomb’

‘dental hygiene’ : ‘tooth hygiene’.

关系形容词与描写形容词的区别之处在于:

——关系形容词不涉及它们所修饰的名词的性质, 因此, 与属性无关。

——关系形容词不能分级。不能说* “the very atomic bomb”。

——大多数关系形容词没有直接反义词。

因此, 关系形容词不能包括到聚类中, 它们也没有两极性。在词网中, 关系形容词的文档包含 2 823 个 SYNSET。每一个关系形容词有一个指针指向相应的名词。

例如, 关系形容词 stellar, astral (星的, 星形的)

⇒ star (星)

⇒ celestial body, heavenly body (天体)

3.4. 词网中的副词

词网中有 4 546 个副词形式, 它们被组织为 5 677 个涵义 (词汇化的概念)。

大多数副词是从形容词通过加后缀“-ly”的方法派生而成的。例如，“beautifully, oddly, quickly, interestingly, hurriedly”等副词分别来自形容词“beautiful, odd, quick, interesting, hurried”。其他的副词是通过加后缀“-ward, -wise, -ways”的方法派生而成的，例如，“northward, crosswise, sideways”。

在词网中，这些派生出来的副词都通过一个意思为“DERIVED-FROM”的指针与相应的形容词联系起来。

3.5. 词网中的动词

词网中的 10 319 个动词被组织到 22 066 个涵义（词汇化的概念）中。

词网的动词数据库中的语义领域有 14 个：motion（运动），perception（感知），contact（接触），communication（交际），competition（竞争），change（变化），cognition（认知），consumption（消耗），creation（创造），emotion（情绪），possession（占有），body care and function（身体保健和功能），social behavior（社会行为），interaction（交互）。普尔曼（Pulman, 1983）建议使用“be”和“do”作为概念系统中一切动词的根结点，用动词“be”表示静态动词，用“do”表示行为动词。但是，这两个动词都是多义的，如果用来作为一切动词的根结点，实际上很不方便，因此，词网没有采用普尔曼的这个建议。词网中的“be”和“do”各有 12 个涵义，例如，在“to be or not be, that is the question”和“Let him be, I tell you”中的 be，涵义各不相同，在“do my hair”，和“do my room in blue”中的 do，涵义也各不相同。显然不能选用 be 和 do 作为一切动词的根结点。

词网 1.5 版中，共有 11 500 个动词的 SYNSET。

在一个单独的语义领域内，很难把所有的动词归属到一个单独的初始概念之下。有些语义领域需要使用若干个独立的树形结构来表示。

例如，表示 motion（运动）的动词要分为 move1 和 move2。move1 表示有位移的运动，move2 表示没有位移的运动。

表示 possession（所属）的动词向上归属时，要归属到 3 个不同的概念，用 3 个不同的 SYNSET 分别表示为{give, transfer}, {take, receive}, {have, hold}。

表示 communication（交际）的动词要分为 verbal communication（口头交际）和 nonverbal communication（非口头交际，如使用手势进行交际）。

词网中使用“承袭”（entailment）来描述两个动词之间的关系。两个动词 V1 和 V2，如果句子“someone V1”表示的行为合乎逻辑地承袭了句子“some V2”表示的行为，那么，我们就说，V1 承袭了 V2。

例如，句子“He is snoring.”表示的行为“他打鼾”，承袭了句子“He is sleeping.”表示的行为“他睡觉”，我们就说，动词“snore”（打鼾）承袭了动词“sleep”（睡觉）。从逻辑上说，如果第一个句子成立，那么，第二个句子也成立。

动词之间的承袭关系具有如下的性质：

- 单词的承袭关系是单向关系：如果动词 V1 承袭了动词 V2，而且它们不是同义词，那么，动词 V2 不能承袭动词 V1。
- 如果两个动词彼此承袭，那么，它们必定是同义词，也就是说，它们具有相同的涵义。
- 否定可以改变承袭的方向：“not sleeping”承袭“not snoring”，但是，“not snoring”不承袭“not sleeping”。
- 否定承袭的一方会造成矛盾：如果句子“he is snoring”承袭“He is sleeping”，那么，句子“he is snoring”与句子“he is not sleeping”矛盾。
- 具有承袭关系的动词在时间上存在着联系：例如，“drive”（驾驶）和“ride”（乘

车)在时间上是相互联系的。如果你“drive”,那么,你一定也同时在“ride”。

•承袭关系在时间上的包含关系:“Snoring”和“sleeping”在时间上是同时存在的,你用来 snoring 的时间是你 sleeping 的时间的一个部分。Snoring 的时间包含在 sleeping 的时间之中,但不一定总是同时的。如果你停止 sleeping,那么,你一定也必须停止 snoring(不过你可以继续 sleeping 而不再继续 snoring)。这就是说,由承袭关系联系起来的两个动词中,一个动词在时间上包含在另一个动词之中。如果在动词 V1 和 V2 发生的时间片段中,动词 V1 发生而动词 V2 不发生,那么,我们就说,动词 V2 发生的时间真正包含在动词 V1 发生的时间之中。

在词网中动词涵义的分布情况可以用坐标来表示。在一个直角坐标系中,如果我们用 y 轴表示词位的涵义数,用 x 轴表示多义词的数目,那么,动词涵义的分布情况如图 10-19 所示:

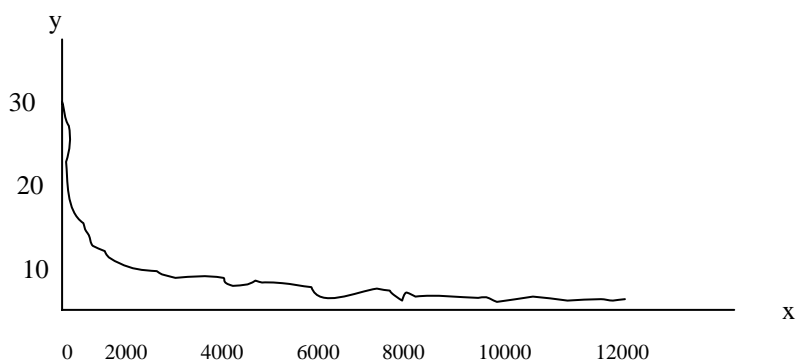


图 词网中动词涵义的分布

从图中可以看出,在词网中,多义程度很高的动词的数量相对地小,大多数动词只有一个涵义。

词网实际上是一个语言知识本体,它给我们提供了极为丰富的词汇语义信息。这些信息对于自然语言处理的语义分析是大有用处的。

附录: ONTOL-MT 中的概念层次结构简表

0 [语义特征] (semantic feature)

1 [事物] (entity) 在空间(包括思维空间)上和时间上延展的事物本体。

1.1 [物] (thing) 主要在空间(包括思维空间)上延展的事物本体。

1.1.1 [具体物] (concrete) 有形、有色、有质量的物。

1.1.1.1 [动作主体] (agentive) 可以发出动作的具体物。

1.1.1.1.1 [人] (human) 人类的个体或群体,[人]有时也包括[组织][集体]。

例如:孩子,女儿,敌人,候选人,医生,作曲家,电话接线员,奔跑者,说话人,学生,工人,农民。

1.1.1.1.1.1 [职务] ([官职],[专职]) (title) 工作中所规定担任的事情。

例如:校长,经理,主任,办事员,秘书。

1.1.1.1.1.2 [作用者] (influential) 对事物产生某种影响的人。

例如:护卫,顾客,委员长。

1.1.1.1.2 [动物] ([活物]) (animate) 多以有机物为食料,有神经,有感觉,能运动的生物。

例如：鸟，鱼，虫，家畜，家禽，兽类，禽类，胎儿，卵。

1.1.1.1.3 [身体] (body) 一个人或者一个动物的生理组织的整体，有时专指躯干和四肢。

例如：全身，母胎。

1.1.1.1.3.1 [身体部位] (body-part) 组成身体的各个部位。

例如：头，脸，眼睛，嘴，耳朵，眉，脸庞，牙齿，下颚，头发，手，手臂，手指，腿（脚），胸，心脏，皮肤，关节，肘，腹，喉咙，肩，脖子，嘴唇，乳房，骨，爪，皮肤，角，翅膀，羽毛，皱纹，感觉器官。

1.1.1.1.4 [集体]（[组织]）(organization) 许多人合起来的有组织的整体，跟“个体”相对。

例如：队伍，流派，家庭，家族，民族，部落，公司，国家，学校，医院，机关，金融机构，新闻机构，交通机关，法庭，新闻媒介，银行，店铺，批发商，零售商。

1.1.1.2 [非动作主体] (non-agentive) 不能发出动作的具体物。

1.1.1.2.1 [植物] (plant) 多以无机物为食料，细胞多具有细胞壁，一般含有叶绿素，没有神经，没有感觉的生物。

例如：花，草，叶子，根，芽，花蕾，树枝，果实，水果，蔬菜。

1.1.1.2.2 [无生物] ([物]) (inanimate) 由物质构成的、占有空间空间的个体。

例如：器具，设备，桥，电话，煤气，上下水道，广播设备。

1.1.1.2.2.1 [天体] (celestial) 恒星、行星、卫星以及彗星、流星、宇宙尘、星云、星团的总称。

例如：太阳，月亮，星，流星，天。

1.1.1.2.2.2 [固体] (solid) 有一定体积，有一定形状，不能流动的物体。

例如：沙石，矿物。

1.1.1.2.2.3 [液体] (liquid) 有一定的体积、没有一定的形状、可以流动的物体。

例如：水，汞，石油，酒。

1.1.1.2.2.4 [气体] (gas) 没有一定的形状、也没有一定的体积、可以流动的物体。

例如：蒸汽，空气，沼气，云，雾，烟。

1.1.1.2.2.5 [信息物] (informative-carrier) 负荷信息的实体。

1.1.1.2.2.5.1 [语言] (language) 人类用来表达意思、交流思想的工具，一般包括它的书面形式，但在与“文字”并举时只指口语。

1.1.1.2.2.5.1.1 [话语] (speech) 说出来的话。

例如：话，流言。

1.1.1.2.2.5.1.2 [书面语] ([语言作品]) (writing) 用语言写出来的书面成品。

例如：文件（包括信件，文书等），诗歌，书籍，出版物，词典，经文，故事。

1.1.1.2.2.5.1.2.1 [通知] (notice) 把事情告诉人知道。

例如：报道，海报。

1.1.1.2.2.5.1.2.2 [记录] (record) 把听到的话语或者发生的事情写下来。

例如：录音，速记。

1.1.1.2.2.5.2 [符号] ([记号]) (symbol) 为引起注意，帮助识别、记忆而做成的标记。

例如：信号，文字，名字，名称，手势，HCL（盐酸的化学符号）。

1.1.1.2.2.6 [工具] (tool) 工作时所需要的器具。

1.1.1.2.2.6.1 [交通工具] (vehicle) 运输用的车辆、船只和飞机等。

例如：汽车，飞机，轮船。

1.1.1.2.2.6.2 [机器] (machine) 由零件构成的、能运转、能变换能量或信息、或者产生有用功的装置。

- 例如：车床，磨床，铣床，计算机，复印机。
- 1.1.1.2.2.6.3 [容器] (container) 盛物品用的器具。
例如：试管，烧瓶，水缸。
- 1.1.1.2.2.6.4 [用具] (apparatus) 日常生活或生产中所使用的小工具。
例如：锥子，锤子，刨子，锯。
- 1.1.1.2.2.6.5 [乐器] (music instrument) 可以发出乐音，供演奏音乐使用的器具。
例如：钢琴，小提琴，锣，鼓。
- 1.1.1.2.2.6.6 [体育用品] (sports goods) 体育运动中使用的物品。
例如：篮球，排球，足球，网球。
- 1.1.1.2.2.6.7 [文化用品] (cultural goods) 运用文字时所使用的物品。
例如：钢笔，圆珠笔，钉书机，文具盒。
- 1.1.1.2.2.6 [产品] (product)
- 1.1.1.2.2.6.1 [物质产品] ([建筑物]) (material-product) 建筑而成的房屋、桥梁、隧道、水坝等。
- 1.1.1.2.2.6.2 [精神产品] ([作品]) (mental-product) 文学或者艺术方面的成品。
例如：音乐，乐曲，歌。
- 1.1.1.2.2.7 [材料] (material) 可以直接制造成品的东西。
例如：水泥，金属，钢材。
- 1.1.1.2.2.8 [金钱] (money) 充当一切商品的等价物的特殊商品，也就是货币。
例如：收入，财产，货币，经费，家庭经济，生计。
- 1.1.1.2.3 [自然现象] (natural-phenomena) 存在于自然界（无机界、有机界）的各种现象。
- 1.1.1.2.3.1 [天文地理] (astro-geo) 日月星辰等天体或者山川、气候等地物分布、运行的现象。
例如：自然，火，雨，雪，风，波浪，光，热，露水，火山，山，山脉，河，天空，土，雷，闪电，天气，阳光，日，光亮，虹，气压，潮汐
- 1.1.1.2.3.2 [物理现象] (physical-phenomena) 只涉及物质的形态变化而不改变物质的化学成分的自然现象。
例如：电，电波，声音，颜色，气味，味，影。
- 1.1.1.2.3.3 [生理现象] (physiological-phenomena) 有关机体的生命活动和体内各器官的机能的现象。
例如：表情，视线，眼泪，刺激，麻醉，疲劳，醉，口水，汗，脉，便（排泄物）。
- 1.1.1.2.3.3.1 [病] (illness) 生理上或者心理上发生的不正常的状态。
例如：感冒，失眠、受伤，伤。
- 1.1.1.2.4 [力能] (energy) 能源和能量
- 1.1.1.2.4.1 [能源] (energy-resource) 可以产生能量的物质。
例如：石油，煤炭，燃料。
- 1.1.1.2.4.2 [能量] (energy-form) 物体做功的能力。
例如：动能，势能，原子能。
- 1.1.2 [抽象物] (abstract) 无形、无色、无质量的物。
- 1.1.2.1 [学问] (theory-principle) 正确反映客观事物的系统知识。
例如：学科，物理学，语言学，相对论，牛顿定律，欧拉公式。
- 1.1.2.2 [概念] (concept) 反映事物本质特征的思维的基本形式。
例如：加速度，圆周率，变量。

- 1.1.2.3 [事理](thing-logic) 事物的原因、结果、目的、方式等。
 例如：原因，结果，方式，目标。
- 1.1.2.4 [伦理](ethics) 人与人相处的各种道德和行为的准则。
 例如：责任，权利，道德，规范，功绩，名声，品行，作风，伦常，贞操，胆量，气量，气概，才能，智慧。
- 1.1.2.5 [力量](force) 人或者动物的力气、能力或效力。
 例如：帮助，能力，势力（“自信，精力”也包含在[力量]之内）。
- 1.1.2.6 [心理](mentality) 人的思想、感情等内心活动。
- 1.1.2.6.1 [知觉]([感觉]): (perception) 肤觉、味觉、嗅觉、听觉、视觉。
 例如：痛感，失聪，失明。
- 1.1.2.6.2 [感情] (feeling) 对于外界刺激的比较强烈的心理反应。
 例如：精神，性情，平常心，良心，愤怒，迷惑，怀疑，关心，好奇心，注目，恐惧，嫉妒，希望，不安，生存价值，恩情，热情，紧张，自暴自弃。
- 1.1.2.6.3 [想法](thought) 思索所得的结果或意见。
 例如：判断，推理。
- 1.1.2.7 [意识] (consciousness) 对于外界世界的比较理智反应。
 例如：意图，愿望，志气，意志，干劲，想法，主意，见解，理想，幻想，兴趣，情致。
- 1.2 [事] ([事情]) (affair) 主要在时间上延展的事物本体。包括人类生活中的一切活动和所遇到的一切社会现象（政治、军事、法律、经济、文化、教育）或与人有关联的自然现象。
 例如：状况，关系，基准，决定，法律，规则，案件，顺序，地位，身份，程度，速度，范围，灾难，灾害，势态，效果，证据，声望，评价，预测，期待，批评，认识。
- 1.2.1 [活动] (activity) 为达到某种目的而采取的行动。
 例如：革命，解放，抗议，游行，排练。
- 1.2.2 [形态] (shape) 人或者事物表现出来的状态。
 例如：巩固，增强，好转，新生，衰落。
- 1.2.3 [争斗] ([战斗]) (struggle) 相对立的一方力求克服另一方的活动。
 例如：比赛，游戏，争吵，围棋，赌博，抓阄，抽签。
- 1.2.4 [集会] (assembly) 许多分散的人聚在一起。
 例如：会议，婚礼，晚会。
- 2 [时间] (time) 由过去、现在和将来构成的连绵不断的系统，它是物质运动和变化的持续性的表现，是物质存在的一种客观形式。
 例如：时代，空闲，途中。
- 2.1 [时点] (time-point) 指时间里的某一点。
 例如：时间的某一点，今天，明天，时刻，节日，假日，生日。
- 2.2 [时段] (period) 指有起点和终点的一段时间。
 例如：时期，期限，年，季节，朝代，年代，干支时。
- 2.2 [时间属性] (time-attribute) 时间所具有的属性（年、月、日、小时、分、秒、毫秒等）。
 例如：时间量（3小时，1个月）。
- 3 [空间] (space) 事物及其运动存在的另一种客观形式，它在不同的维度上延伸。
- 3.1 [场所] (place) 由长度、宽度和高度表现出来的物质存在的一种客观形式，也就是活动的处所。

例如：位置，空间，沟，洞，饭桌，都市，社会，人世，神社，天下，寺庙，墓，井，水池，焦点，水田，旱田，旅馆。

3.2 [距离] (distance) 在空间或者时间上相隔。

例如：差距。

3.3 [途径] ([道]) (way) 两地之间的通道。

例如：出路，路线。

3.4 [方向] (direction) 指东、南、西、北、上、下等。

例如：方位，前后，左右，西。

4 [数量] (quantity) 事物的多少与计量。

4.1 [数值] (number-value) 一个量用数目表示出来的多少。

例如：“10，3人，80公斤”中数值表现。

4.2 [计量] (measure) 温度，长度，重量，使用量等。

例如：度、升，枚，册，匹，羽。

4.3 [金额] (sum) 钱数的大小。

例如：5元。

5 [行为状态] (action-state) 人或事物表现出来的活动和形态。

5.1 [物理行为] (physical-action) 人或事物在物理上表现出来的活动。

5.1.1 [物理移动] (movement) 物理位置的改换。

例如：上升，下降，滚，滑。

5.1.2 [所有权转移] (ownership-transfer) 所有权隶属关系的改换。

例如：给，交，送，赠。

5.1.3 [属性变化] (property-change) 事物属性的改换。

例如：减肥，扩大，伸长，凉干。

5.1.4 [身体动作] (bodily-action) 人或动物身体的动作。

5.1.4.1 [头部动作] (head-action) 头部的动作。

例如：点头，摇头。

5.1.4.2 [五官动作] (sense-action) 眼、耳、口、鼻舌的动作。

例如：看，听，闻。

5.1.4.3 [徒手动作] (hand-action) 不持工具的手部动作。

例如：摸，握手。

5.1.4.4 [持工具动作] (tool-action) 持工具的手部动作。

例如：挥动，敲打。

5.1.4.5 [脚部动作] (foot-action) 脚部的动作。

例如：踢，走，跑。

5.1.4.6 [全身动作] (body-action) 全身的动作。

例如：蠕动，跌倒。

5.1.5 [生产活动] (production-activity) 人使用工具创造各种生产资料或生活资料的活动。

例如：装配，架设，粉刷，冶炼，垦荒，施肥，打猎。

5.2 [心理行为] (psychological-action) 人或动物在心理上表现出来的活动。

5.2.1 [知识传达] (knowledge-transfer) 知识的转移和传递。

例如：感到，传播，说明，通知。

5.2.2 [知觉作用] (perceptive action) 反映客观事物的整体形象和表面联系的心理行为。

例如：明白，懂得，疏忽，醒悟，听见，看到。

- 5.2.3 [感情作用] (feel) 对于外界刺激的比较强烈的心理行为, 或对于人或事物的关切喜爱的心情。
- 例如: 热爱, 憎恨, 惭愧, 沉迷, 感激。
- 5.2.4 [思考活动] (think) 比较深刻、周到的思维活动。
- 例如: 分析, 归纳, 推敲, 考证, 研究。
- 5.2.5 [情态能愿] (will) 表示可能性或必然性的心理行为。
- 例如: 能够, 可以, 应该, 必须。
- 5.3 [状态] (state) 人或事物表现出来的形态。
- 5.3.1 [存在] (existence) 人或事物占有或者未占有时间和空间的状态。
- 例如: 有, 在, 出现, 消失。
- 5.3.2 [心态] ([表情]) (mental-state) 内心的和思想的状态与表情。
- 例如: 害臊, 撒娇, 发呆, 瞪眼。
- 5.3.3 [物态] (physical-state) 物体的状态。
- 例如: 凹陷, 裸露, 倒塌, 裂开。
- 5.3.4 [体态] (body-state) 人或动物身体的状态。
- 例如: 卧, 躺, 坐, 站。
- 5.3.5 [事态] (state of affairs) 事情发展过程中的状态。
- 例如: 流行, 积压, 生效, 失效。
- 5.3.6 [境遇] (circumstance-state) 状况和遭遇。
- 例如: 享福, 吃苦, 沾光, 上当。
- 5.3.7 [始末] (begin-end) 开始和结束。
- 例如: 开始, 结束, 中断, 继续。
- 5.3.8 [变化] (change) 事物在形态上或者本质上产生新的状况。
- 例如: 成为, 改变, 改换, 添补。
- 5.3.9 [类属] (isa) 所属关系表现出来的形态。
- 例如: 是, 成为, 当作, 如同。
- 6 [属性] (attribute) 事物所具有的特性和关系。
- 6.1 [外形] ([仪容]) (appearance) 人或事物的外部形体属性。
- 例如: 宽敞, 广阔, 狭窄, 平坦, 漂亮。
- 6.2 [表象] (surface) 从外表可以观察到的现象的属性。
- 例如: 稀少, 丰富, 齐全, 沙哑。
- 6.3 [颜色] ([光亮]) (color) 由物体发射、反射或者透过的光波通过视觉所产生的印象。
- 例如: 红, 白, 蓝, 绿, 鲜艳, 五彩。
- 6.4 [味道] (taste) 能使舌头得到某种味觉的特性。
- 例如: 香, 臭, 苦, 辣, 甜, 酸, 油腻, 清淡。
- 6.5 [性质] (character) 一个事物区别于另一个事物的属性。
- 例如: 好, 坏, 真实, 虚假。
- 6.6 [德才] (moral-ability) 人的道德和才能表现出来的属性。
- 例如: 善良, 仁慈, 机智, 能干。
- 6.7 [境况] (circumstance) 外界环境所具有的属性。
- 例如: 太平, 安定, 舒服, 艰苦。
- 6.8 [关系] (relationship) 事物之间相互作用与相互影响时具有的属性。
- 例如: 相同, 相等, 矛盾, 连贯。

参考文献

1. 冯志伟, 应用语言学综论, 广东教育出版社, 1999年12月, 广州。
2. 冯志伟, 计算语言学基础, 商务印书馆, 2001年8月, 北京。
3. 冯志伟, 计算语言学探索, 黑龙江教育出版社, 2001年9月, 哈尔滨。
4. 冯志伟, 应用语言学新论, 当代世界出版社, 2003年10月, 北京。
5. 冯志伟, 机器翻译研究, 中国对外翻译出版公司, 2004年12月, 北京。
6. Fang Gu et al., Domain-specific ontology of botany, *Journal of computer science & technology*, March 2004, Vol.19 No.2, pp.238-248.
7. Chunxia Zhang, Domain-specific formal ontology of archeology and its application in knowledge acquisition and analysis, *Journal of computer science & technology*, May 2004, Vol.19 No.3, pp. 290-301.
8. Asuncion Gomez-Perez, *Ontological Engineering with examples from the areas of Knowledge Management, e-Commerce and Semantic Web*, Springer, 2004.
9. T. R. Gruber. A translation approach to portable ontologies. *Knowledge Acquisition*, 5(2):199-220, 1993.
10. G. Miller, R. Beckwith, C. Fellbaum, D. Gross, K. Miller, Introduction to WordNet: A on-line lexical database, *International Journal of lexicography*, 3(4), 235-244.,1990.
11. G. Miller, WordNet: a lexical database for English. *Communication of the ACM*, 38(1), 39-41, 1995.
12. W. N. Borst, *Construction of engineering ontologies*. Centre for Telematica and information technology, University of Twente. Enschede, The Netherlands, 1997.
13. R. Studer, V. R. Benjamins, D. Fensel, *Knowledge Engineering: Principle and Methods*, 1998.
14. D. Jurafsky & J. Martin, *Speech and Language Processing*, Prentice Hall, 2000, New Jersey.