载《语言文字应用》，2005 年，第 4 期

On humanity spirit of NLP from the viewpoint of ontology

Feng Zhiwei

zwfengde@public.bta.net.cn

The ontology is a formal explicit specification of a shared conceptualization. The ONTOL-MT system originally is an ontology for Japanese-Chinese machine translation. This ONTOL-MT system is based on the system of categories proposed by Aristotle. It is useful not only for the homonym disambiguation in Japanese parsing, but also for syntactic function disambiguation in Chinese parsing. This paper also pointes out some shortcomings in the unique beginner of Wordnet.

从知识本体谈自然语言处理的人文性

冯志伟

教育部语言文字应用研究所

zwfengde@public.bta.net.cn

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”。

远在古希腊时代，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，实际上也就是最早的知识本体。亚里士多德以他卓越的学识和深刻的洞察力，抓住了人类认识中最关键的概念。

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”紧密地联系起来了。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 unity（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：unity, 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。因此，康德的范畴框架是帮助我们捕捉外在世界的有力手段。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。它主的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哲学中经典的知识本体研究的基础之上的，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植物学领域知识本体（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。

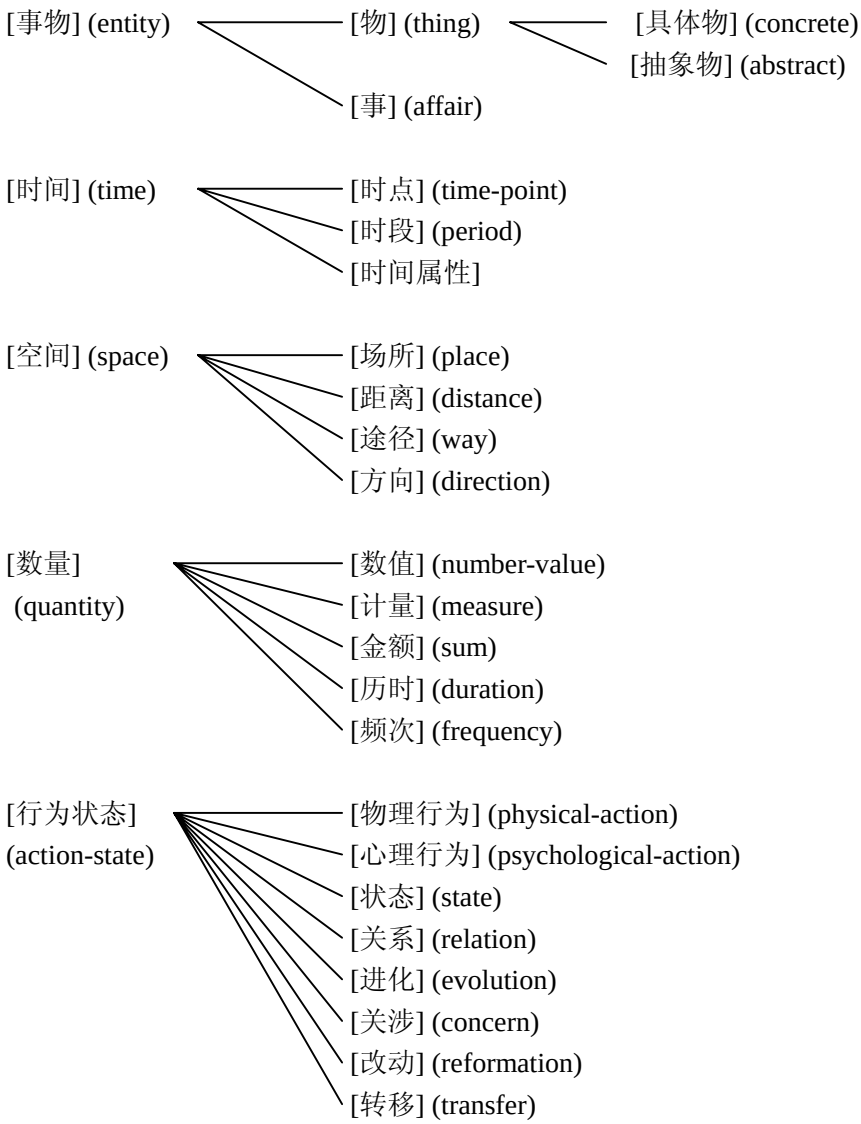
- 形式知识本体（formal ontology）对于概念和术语的分类很严格，要按照一定的原则和标准，明确地定义概念之间的显性和隐性关系，明确概念的约束和逻辑联系。领域知识本体或术语知识本体经过进一步的抽象和提炼，就可能发展成形式知识本体。

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。在机器翻译的语义分析中，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。

我们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。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：



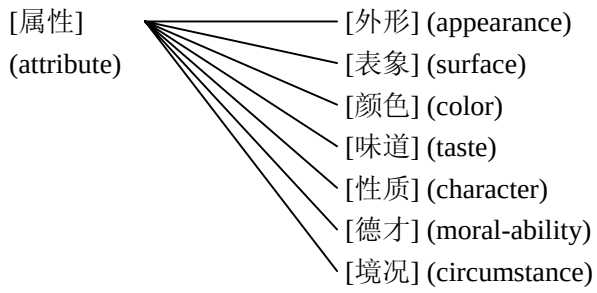


图 1　知识本体 ONTOL-MT

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)　一个量用数目表示出来的多少。

[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，这里就不一一列出了。详细的描述请参阅冯志伟《计算语言学基础》一书第 4 章第 2 节。

可以看出，ONTOL-MT 中的初始概念基本上是采用了亚里士多德的范畴系统。因为我们认为，世界上的一切事物都是在时间和空间中运动和存在的，它们要表现出一定的行为和状态，并且具有一定的属性和数量。亚里士多德的这个范畴系统是从哲学的角度出发的，充满了理性的色彩，闪耀出智慧的光芒。康德的范畴框架是帮助人们捕捉外在世界的有力手段，但是，这个范畴框架还不是一个完整的概念系统，我们认真研究了亚里士多德的范畴系统和康德的范畴框架，经过反复的比较和权衡，最后我们决定基本上采用亚里士多德的范畴系统，同时参考康德的范畴框架。我们认为，建立知识本体系统固然有很多技术性的问题需要解决，但是，首先我们必须考虑的是它的人文性，知识本体首先必须合乎人们的理性。

自然语言处理是技术性和实践性很强的学科，自然语言处理的研究不仅要说得通，还要做得通，每一个研究自然语言处理的人都不应该纸上谈兵，我们应该重视自然语言处理的技术性和实践性，如果自然语言处理的理论在计算机上做不通，那决不是好的理论。但是，自然语言处理的对象终究是人类的自然语言，它与物理学、化学、生物学的研究对象的最大的区别在于它与人密切相关，在于它的人文性。我们在自然语言处理的研究中，在重视它的技术性和实践性的同时，千万不要忘记了自然语言的人文性。正是处于这样的原因，我们在建立 ONTOL-MT 知识本体的时候，我们从亚里士多德的范畴系统中找到了人文性的依据。

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。例如，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，机器翻译时会出现很大的困难。我们根据 ONTOL-MT，在机器翻译的词典中，给这三个“きしゃ”分别标上不同的固有语义特征，给第一个“きしゃ”标上语义特征“人”，第二个“きしゃ”标上语义特征“交通工具”，给第三个“きしゃ”标上语义特征“移动”，这样就可以把它们区分开来，计算机就可以得到三个“きしゃ”的正确的汉语译文，再经过结构转换和汉语生成，最后计算机得到的这个句子的汉语译文是“记者乘火车回公司”。显然，这是正确的汉语译文。

我的博士生杨泉使用 ONTOL-MT 研究现代汉语中的同类词短语的句法功能歧义的消解问题，效果良好。所谓“同类词短语”是指词性相同的短语，如 N+N（名词+名词），V+V（动词+动词），　A+A（形容词+形容词）等。这些短语格式都具有句法功能的潜在歧义。例如，N+N 的句法功能关系可以是并列关系（如“爷爷奶奶”），也可以是复指关系（如“胡

锦涛主席”），也可以是定中关系（如“金色魔杖”），也可以是主谓关系（如“今天星期一”），这时，我们就说 N+N 具有潜在的句法功能歧义，在自然语言处理中分析 N+N 结构的时候，就有必要进行歧义消解。为了表述方便，我们把 N+N 改写为 N1+N2，这样，计算机就可以根据 N1+N2 中每个名词在 ONTOL-MT 中的概念特征来辨别它们之间不同的句法功能关系，从而达到歧义消解的目的。例如，当 N1 和 N2 都具有概念特征“亲属”时，计算机就判断它们之间的关系是并列关系；当 N1 的概念特征是“专名”，N2 的概念特征是“职务”时，计算机就判断它们之间是复指关系；当 N1 的概念特征是“颜色”，N2 的概念特征是“具体物”时，计算机就判断它们之间是定中关系；当 N1 的概念特征是“时点”，N2 的概念特征是“时间属性”时，计算机就判断它们之间是主谓关系。

由此可见，ONTOL-MT 知识本体中的语义特征对于在机器翻译中区分同形词以及在自动结构分析中辨别句法功能歧义是非常有用的。在这些工作中，由于我们充分地考虑了 ONTOL-MT 知识本体的人文性，整个系统的结构很合理，保证了这个知识本体成为“概念体系的明确规范”。

在建立 ONTOL-MT 知识本体的过程中，我们也学习了词网 WordNet。词网是一个被广泛使用的知识本体，在自然语言处理中有很大的影响。但是，我们发现，词网的初始概念与亚里士多德的范畴系统差别太大。

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）：

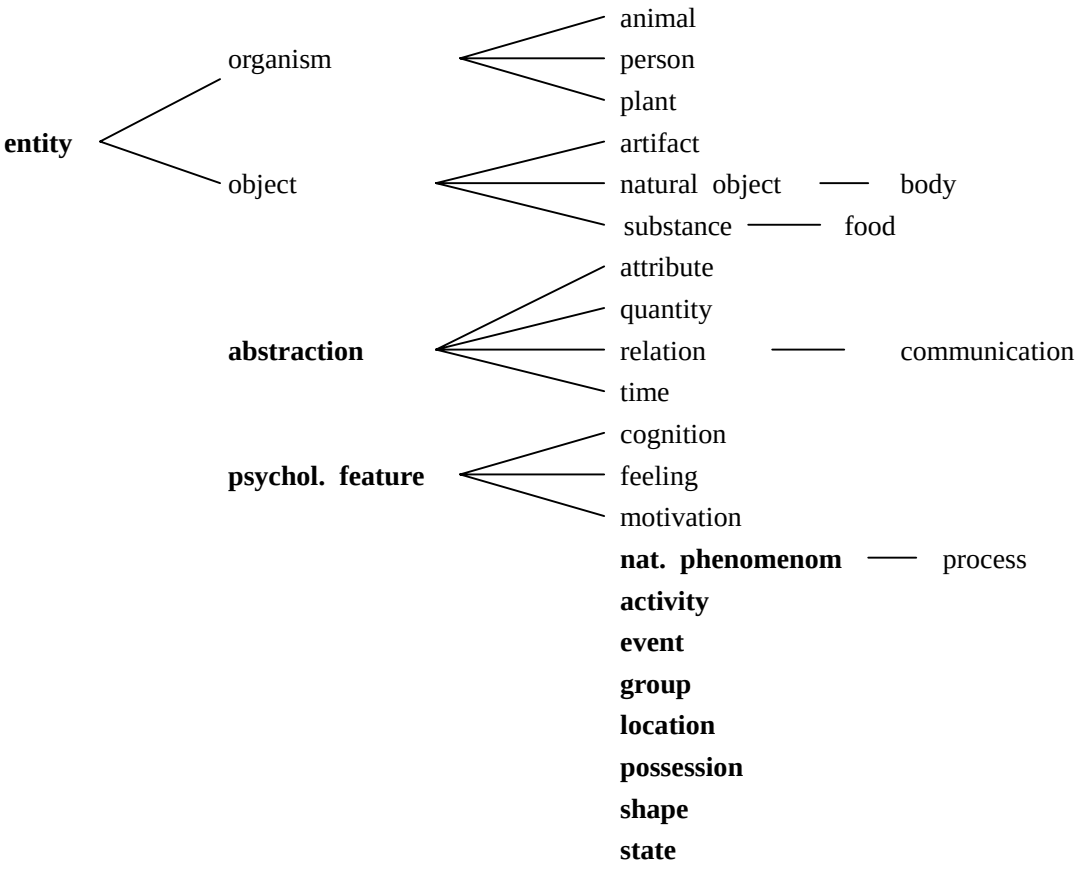


图 2　词网的初始概念

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）。

可以看出，亚里士多德范畴系统中的初始概念“时间”（time）、“数量”（quantity）和“关系”（relation）在词网中都归并到初始概念“抽象”（abstraction）中去了，成为了“抽象”的下位概念，而“外形”(shape)和“所属”（possession）这样的显然不属于基本范畴的概念，在词网中却是初始概念。词网中的初始概念与亚里士多德的范畴系统相距甚远，因此我们使用起这些初始概念来，总觉得词网的这些初始概念有些古里古怪的，不好理解，这说明，词网的初始概念的设计缺乏人文性。

自然语言处理是一个技术性很强的学科，我们当然应当重视它的技术性，但是，自然语言处理的研究对象终究还是人的语言，在研究语言知识本体的时候，我们除了考虑技术性的因素之外，也应当考虑人文性的因素，我们应当充分注意人文科学中在知识本体方面的已经取得的宝贵成果，来充实我们的研究工作，我们要呼唤人文性！

参考文献

1. 冯志伟，计算语言学探索，黑龙江教育出版社，2001 年。
2. 冯志伟，计算语言学基础，商务印书馆，2001 年。
3. 冯志伟，机器翻译研究，中国对外翻译出版公司，2004 年。
4. 杨泉，面向信息处理的现代汉语同类词短语句法功能研究，中国传媒大学博士论文，2005

年。
5. D. Jurafsky, J. Martin（冯志伟 孙乐 译），自然语言处理综论，电子工业出版社，2005年。
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, 38911), 39-41, 1995.
12. W. N. Borst, Construction of engineering ontologoes. Centre for Telemetica and information technology, University of Tweenty. Enschede, The Netherlands, 1997.
13. R. Studer, V. R. Benjiamins, D. Fensel, Knowledge Engineering: Principle and Methods, 1998.