

第六章 本体

戴洪良

计算机科学与技术学院/人工智能学院

hongldai@nuaa.edu.cn



概念

- 本体 (Ontology) 一词来自哲学中的本体论 (Ontology)
 - 本体论研究事物存在的本质；研究 “什么是本质” 、 “所有事物的一般特征是什么” 等一些基本的问题。
 - “ontology is the philosophical study of being.”
 - “It investigates what types of entities exist, how they are grouped into categories, and how they are related to one another on the most fundamental level (and whether there even is a fundamental level).”

概念

- 计算机科学/信息科学中，对本体这一概念的定义和描述
 - 1991/Neches等：本体定义了“给出构成相关领域词汇的基本术语和关系，以及用于扩展这些词汇的组合术语和关系的规则”
 - 1993/Gruber等：本体是“概念模型的明确的规范说明”（“an explicit specification of a conceptualization.”）。该定义得到人们的广泛人认可
 - 2001/Lassila等：有限且受控的词汇表，对类和术语间关系的无歧义的解释，和类之间子类关系的严格层次化结构

概念

- 1998/Studer 对本体的定义包含了4 层含义：概念模型、明确化、形式化和共享
 - 概念模型 (conceptualization) : “A conceptualization is an abstract, simplified view of the world that we wish to represent for some purpose”
 - 明确化 (explicit) : 所使用的概念及使用这些概念的约束都有明确的定义
 - 形式化 (formal) : 本体是计算机可读的
 - 共享 (share) : 本体中体现的是共同认可的知识, 反映的是相关领域中公认的概念集, 它所针对的是团体而不是个体

概念

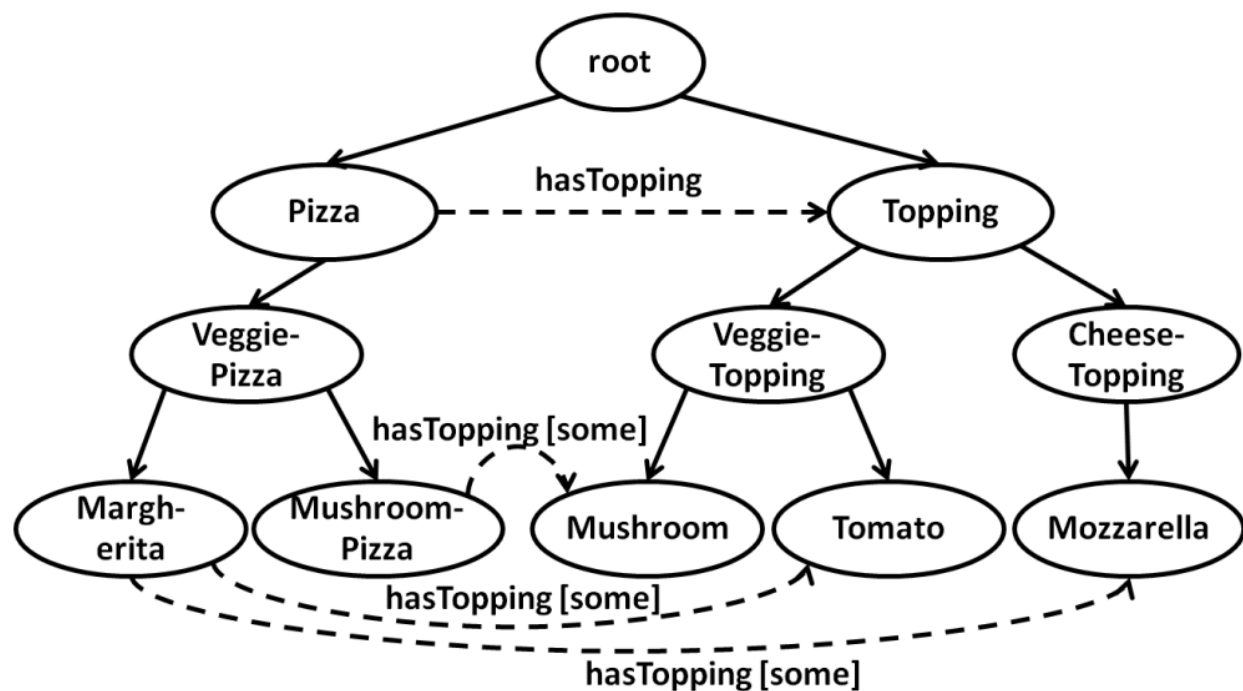
- 从以上不同研究者的定义，可以看出本体涉及到的概念有：术语（词汇）、概念化、术语关系、规则、形式化、领域知识、共享。
- 本体通过对于概念、术语及其相互关系的规范化描述，勾画出某一领域的基本知识体系和描述语言
- 很多人工产品都与本体相关，比如：术语表、字典、百科全书、知识库

本体的组成

- 本体的常见组成部分有：
 - 类/概念 (classes/concepts) : 概念的含义很广泛, 可以指任何事物, 如“人”、“动物”、“疾病”等等。
 - 属性 (attributes) : 可以给类或概念赋予属性, 表示其个体的特征。如老虎有四只脚。
 - 关系 (relations) : 不同类/概念之间的关联。如老虎是动物的子类。
 - 公理 (axioms) : 永真的断言。可以是限制、规则等。如: 男人和女人是两个不相交的类。
- 此外还可以有实例 (instances)、事件 (events, the changing of attributes or relations) 等

本体 – 例

- 关于披萨和配料的本体样例



本体 – 例

- Disease Ontology
 - <https://disease-ontology.org/>

The screenshot displays the Disease Ontology website interface. On the left, a 'Navigation' panel shows an 'OBO tree' with a hierarchical list of disease categories. The 'chordoid glioma' entry is highlighted. On the right, a 'Chordoid Glioma' details panel is shown, featuring a 'Metadata' table with various identifiers and definitions, and a 'Visualize' button.

Navigation

OBO tree View OWL tree

- disease
 - disease by infectious agent
 - disease of anatomical entity
 - disease of cellular proliferation
 - benign neoplasm
 - cancer
 - cell type cancer
 - aggressive NK-cell leukemia
 - blastoma
 - carcinoma
 - germ cell cancer
 - high grade glioma
 - brain stem glioma
 - cerebellar astrocytoma
 - chordoid glioma**
 - diencephalic astrocytoma
 - diffuse pediatric-type high-grade glioma
 - high grade ependymoma
 - histone mutated tumor
 - malignant astrocytoma
 - mixed glioma
 - oligodendroglioma
 - optic nerve astrocytoma
 - pineal gland astrocytoma
 - spinal cord glioma
 - malignant adenoma
 - malignant mesothelioma
 - mast cell neoplasm
 - melanoma

Chordoid Glioma

Metadata

ID	DOID:3774
Name	chordoid glioma
Definition	A high grade glioma that is characterized by the presence of epithelioid cells which express GFAP, and mucinous stroma which contains lymphoplasmacytic infiltrates. https://pubmed.ncbi.nlm.nih.gov/28315998/ , https://www.frontiersin.org/articles/10.3389/fonc.2020.00502/full
Xrefs	ICDO:9444/1 NCI:C5592 ORDO:251674 SNOMEDCT_US_2023_03_01:128789002 UMLS CUI:C1322252
Alternate IDs	DOID:3773
Subsets	DO_cancer_slim DO_rare_slim NCItthesaurus
Synonyms	Chordoid glioma of 3rd Ventricle [EXACT] Chordoid glioma of third ventricle [EXACT] third ventricle chordoid glioma [EXACT]
Parent	is_a cerebral ventricle cancer
Relationships	is_a high grade glioma

[Submit Comment](#) [Visualize](#)

为什么研究本体？

- 本体的目标是表示相关的领域的知识，提供对该领域知识的共同理解，确定该领域内共同认可的词汇（术语），对它们描述，并给出这些词汇之间相互关系的明确定义
- 支持数据、信息和知识的交换、重用和共享
 - 本体是一个正式的词汇表，将概念和相互间的关系进行较为精确的定义。在其支持下进行知识搜索、知识积累、知识共享的效率将大大提高
- 本体可以为知识库的构建提供一个基本的结构
- 本体的使用可以帮助我们清楚地理解特定领域的相关元素、关系和概念，让知识表达更加准确便捷

本体表示

- 具体描述本体的方法很多
 - 自然语言、框架、语义网络或逻辑语言等都可以用来描述本体
- 本体建模语言大致可分为两类：传统的本体建模语言和面向Web的本体建模语言
 - 主要区别在于面向Web的建模语言语法一般采用XML作为语法基础，常用于表达Web信息的语义
 - 传统的本体建模语言有KIF、Cycl、OKBC、OCML、Frame Logic和LOOM等
 - 面向Web的建模语言有OWL、XOL、SHOE、OML等

本体表示

- CycL语言

Specialization and generalization [\[edit \]](#)

The most important predicates are `#$isa` and `#$genls`. The first one (`#$isa`) describes that one item is an instance of some collection (i.e.: specialization), the second one (`#$genls`) that one collection is a subcollection of another one (i.e.: generalization). Facts about concepts are asserted using certain CycL *sentences*. Predicates are written before their arguments, in parentheses:

For example:

```
(#$isa #$BillClinton #$UnitedStatesPresident) \;
```

"Bill Clinton belongs to the collection of U.S. presidents" and

```
(#$genls #$Tree-ThePlant #$Plant) \;
```

"All trees are plants".

```
(#$capitalCity #$France #$Paris) \;
```

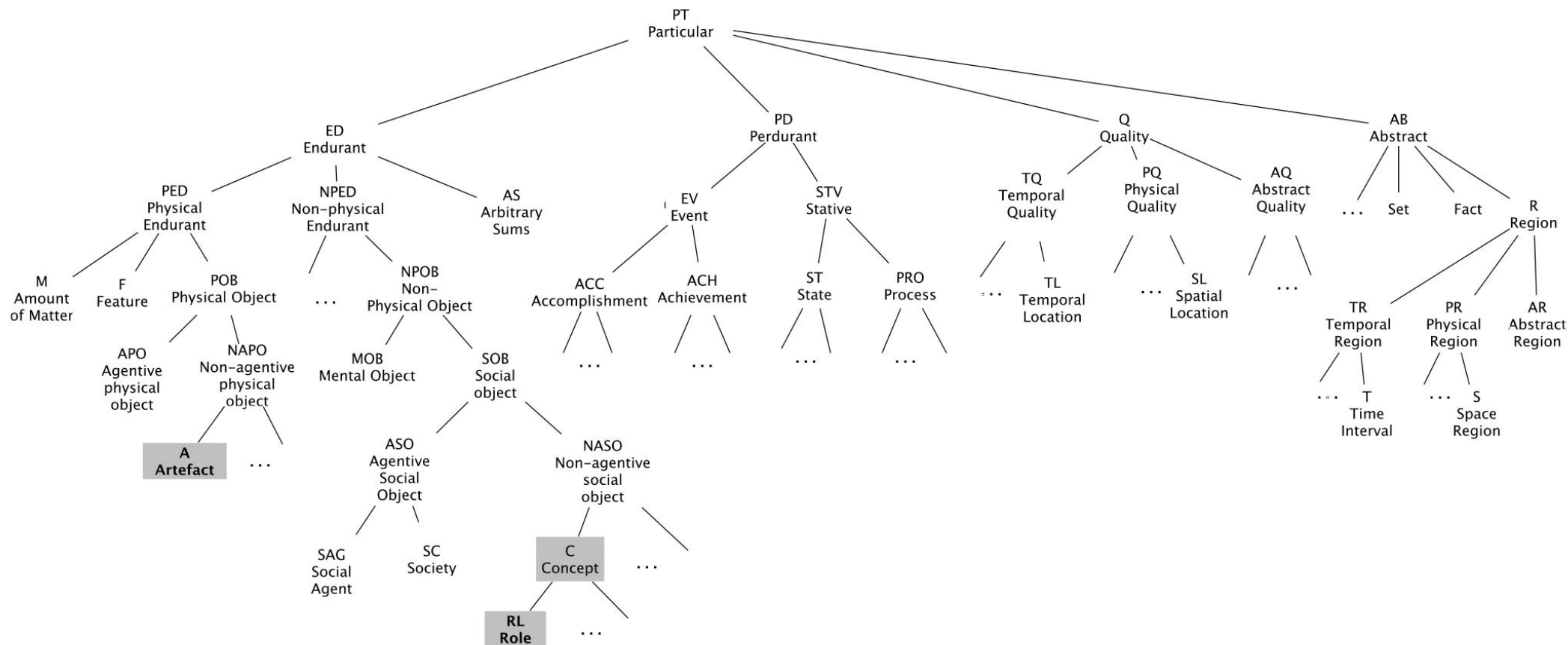
"Paris is the capital of France."

本体类别

- 顶层本体 (Upper Ontology) 描述的是最普通的概念及概念之间的关系, 如空间、时间、事件、行为等等, 与具体的应用无关, 其他种类的Ontologies 都是该类Ontologies 的特例。
- 领域本体 (Domain Ontology) 描述的是特定领域(医药、汽车等) 中的概念及概念之间的关系。

顶层本体 – 例

- DOLCE的taxonomy



Taxonomy is the practice and science of categorization or classification.

本体的构建

- 本体构建，是从某个领域中抽取知识，形成描述该领域数据的语义概念和其间的关系。
- 1995年Gruber提出的5条规则较有影响：
 - (1) **明确性和客观性**：即本体应该对所定义术语给出明确的、客观的语义定义。
 - (2) **完全性**：即所给出的定义是完整的，完全能表达所描述术语的含义。
 - (3) **一致性**：即由术语得出的推论与术语本身的含义是相容的，不会产生矛盾。
 - (4) **可扩展性**：即向本体中添加通用或专用的术语时，不需要修改其已有的内容。
 - (5) **最小承诺**：即对待建模对象给出尽可能少的约束。

本体的构建

- 本体构建的一些方法论：
 - Uschold的“骨架”法 [Uschold, 1998]
 - 确定本体的目的和范围；创建本体；本体评估；建立文档
 - Berneras方法
 - 基于应用的方法，每创建一个应用，就要创建一个针对它所需知识的本体，创建时可重用其他本体。
 - Methontology方法 [Gómez-Pérez, 1998]
 - 本体生命周期的概念来管理整个本体的开发过程，使本体的开发过程更接近于软件工程中的软件开发过程。具体分为三个阶段：管理阶段、开发阶段和维护阶段。

本体的构建

- 目前的本体构建方法论还未能像软件工程那样成为“科学”或“工程过程”的完整方法论。
- 因此，只有总结和发展现有的各种方法论，结合具体应用，再配合领域专家的支持，才能提出适合具体项目的本体构建方法。

本体工具

- 包括编辑工具、标注工具和集成工具等。

(1) 本体编辑工具：

- 本体编辑是一项比较庞大的复杂反复的系统工程，包括：问题说明、领域知识的获取和分析、概念的设计、迭代建设及测试等一系列环节。
- 常用的编辑工具有Protégé、NeOn Toolkit、Onto4ALL等。

本体工具

(2) 本体标注工具:

- 本体标注工具可以在Web页面及其他文档中自动或半自动插入本体标记，将非结构化、半结构化信息与本体联系起来。
- 如：AeroDAML、COHSE和SMORE。

(3) 本体集成工具:

- 本体集成工具用于解决同一领域内本体的融合和集成问题。如：OntoMerge、PROMPT和MAFRA等。

(4) 其他工具:

- 除了上述本体编辑工具、本体标注工具和本体集成工具外，还有本体存储查询工具和学习工具等。

本体的构建-例子

- 骨架法在所有方法中最具参考性，它提供了一个本体构建的方法学框架。骨架法的步骤：
 1. 确定构建本体的目的和范围
 2. 构建本体
 - Top-down
 - Bottom-up
 - Middle-out
 3. 本体评估
 4. 文档化

本体的构建-例子

- 例：构建旅游信息资源本体

过程：

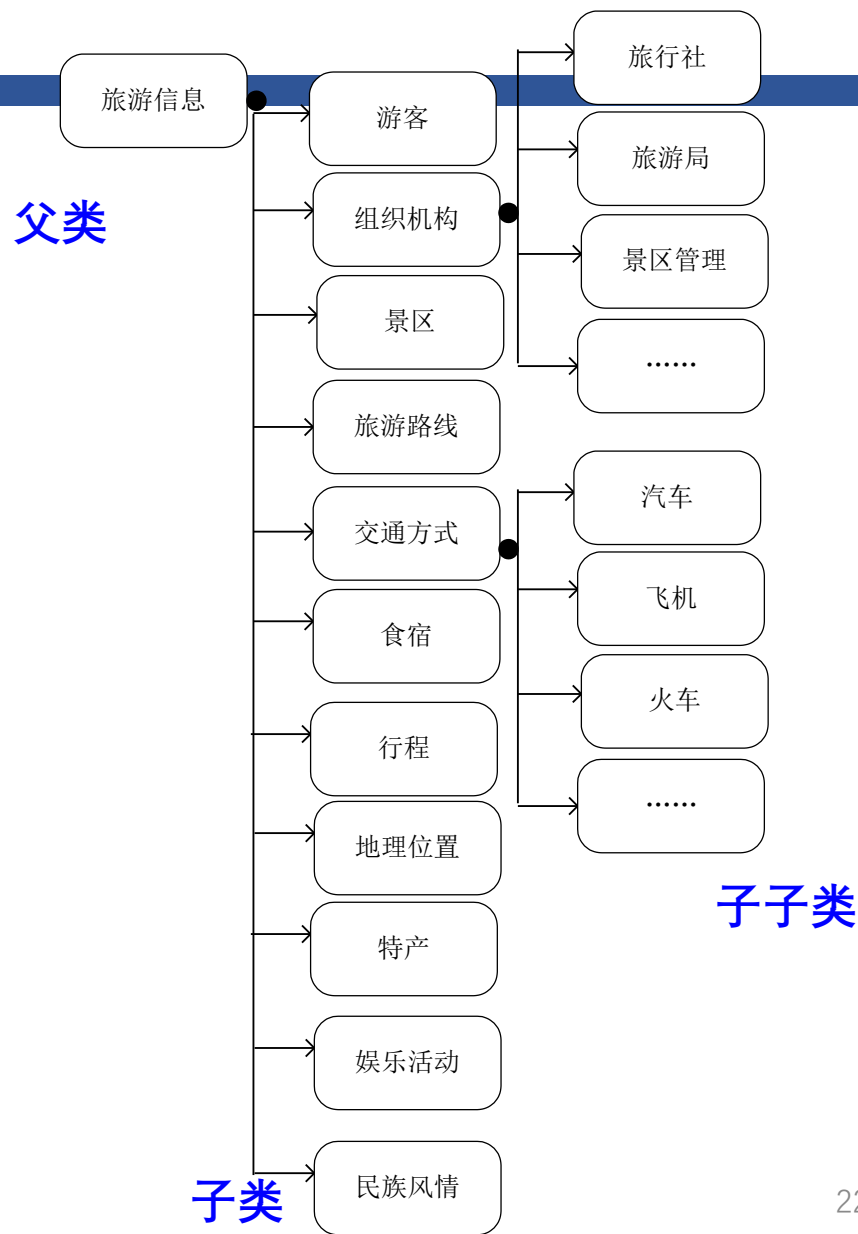
1. 确定本体领域和范畴
2. 列举旅游信息资源本体中的重要术语、概念
3. 定义类和类的层次体系
4. 定义类的属性及其取值类型
 - 数据属性（与值关联，如姓名、身高、颜色等）
 - 对象属性（与其他类或个体关联，如整体部分关系、其他不同概念/个体间关系）
5. 对领域本体编码、形式化

本体的构建-例子

- 旅游信息本体中重要术语与概念为：
 - 人、组织机构、景区、旅游路线、交通方式、食宿、行程、地理位置、特产、娱乐活动、民族风情、旅行社、景区管理机构、交通运输企业、食宿企业、旅游局、保险公司、特产企业、娱乐企业、水文景观、地文景观、人文景观、历史遗产、国家非物质文化遗产、全程路线、地接线路.....

本体的构建-例子

- 定义类和类的层次结构：类用于描述抽象的实体对象，代表着一类具有共性的实例对象；类具有继承性并以层次结构的形式组织。定义类的层次采用自顶向下的方法，其中顶为父类。



本体的构建-例子

- 定义类的属性

- 由于每个类的属性较多，原则是根据需求来定义该领域类的属性。如在旅游信息资源本体中，游客及景区的属性表示为：

- 如，数据属性：

- 游客（姓名，性别，身份证，年龄，旅游类别，爱好，电话，邮箱）
- 景区（名称，景点等级，管理机构，景点类别，地址，容纳人数，服务电话）

- 对象属性：

- 游客 使用 交通方式；游客 参加 娱乐活动；景区 有 娱乐活动

- 生成本体

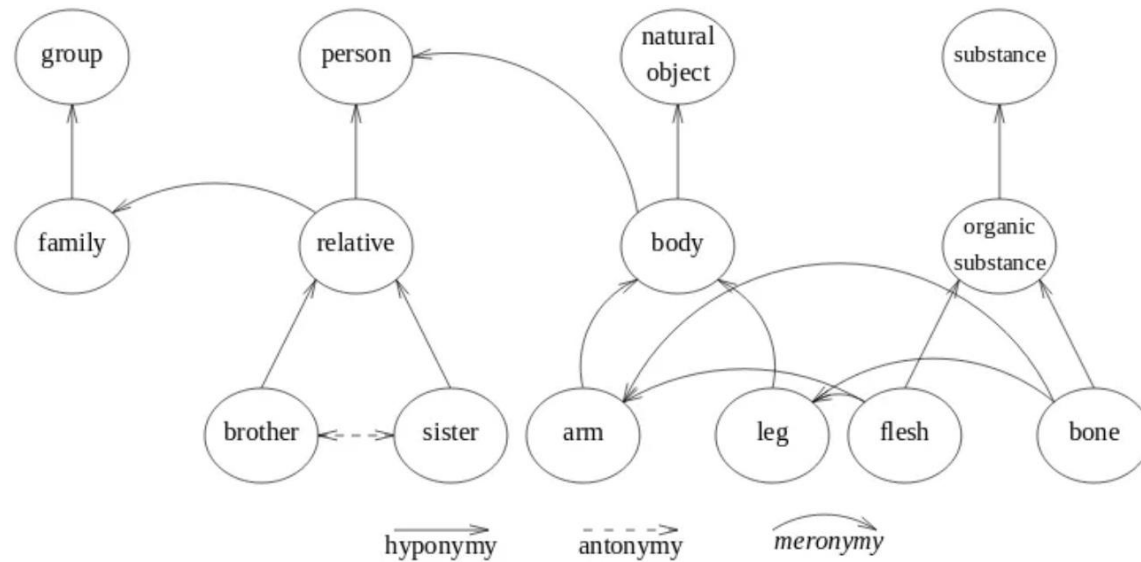
- 通过利用上述信息，结合本体构建的工具，就可以建立起一个旅游信息资源本体库

本体的例子

Published examples [\[edit \]](#)

- [Arabic Ontology](#), a linguistic ontology for Arabic, which can be used as an Arabic Wordnet but with ontologically-clean content.^[36]
- AURUM - Information Security Ontology,^[37] An ontology for [information security](#) knowledge sharing, enabling users to collaboratively understand and extend the domain knowledge body. It may serve as a basis for automated information security risk and compliance management.
- [BabelNet](#), a very large multilingual semantic network and ontology, lexicalized in many languages
- [Basic Formal Ontology](#),^[38] a formal upper ontology designed to support scientific research
- BioPAX,^[39] an ontology for the exchange and interoperability of biological pathway (cellular processes) data
- BMO,^[40] an e-Business Model Ontology based on a review of enterprise ontologies and business model literature
- SSBMO,^[41] a Strongly Sustainable Business Model Ontology based on a review of the systems based natural and social science literature (including business). Includes critique of and significant extensions to the Business Model Ontology (BMO).
- CCO and GexKB,^[42] Application Ontologies (APO) that integrate diverse types of knowledge with the Cell Cycle Ontology (CCO) and the Gene Expression Knowledge Base (GexKB)
- CContology (Customer Complaint Ontology),^[43] an e-business ontology to support online customer complaint management
- [CIDOC Conceptual Reference Model](#), an ontology for [cultural heritage](#)^[44]
- COSMO,^[45] a Foundation Ontology (current version in OWL) that is designed to contain representations of all of the primitive concepts needed to logically specify the meanings of any domain entity. It is intended to serve as a basic ontology that can be used to translate among the representations in other ontologies or databases. It started as a merger of the basic elements of the OpenCyc and SUMO ontologies, and has been supplemented with other ontology elements (types, relations) so as to include representations of all of the words in the [Longman dictionary defining vocabulary](#).
- [Computer Science Ontology](#), an automatically generated ontology of research topics in the field of [computer science](#)
- [Cyc](#), a large Foundation Ontology for formal representation of the universe of discourse
- [Disease Ontology](#),^[46] designed to facilitate the mapping of diseases and associated conditions to particular medical codes
- [DOLCE](#), a Descriptive Ontology for Linguistic and Cognitive Engineering^{[23][24]}
- Drammar, ontology of drama^[47]^[citation needed]
- [Dublin Core](#), a simple ontology for documents and publishing
- Financial Industry Business Ontology (FIBO), a business conceptual ontology for the financial industry^[48]
- Foundational, Core and Linguistic Ontologies^[49]
- [Foundational Model of Anatomy](#),^[50] an ontology for human anatomy
- [Friend of a Friend](#), an ontology for describing persons, their activities and their relations to other people and objects

WordNet



<https://wordnet.princeton.edu/>

WordNet历史

- WordNet由普林斯顿大学心理学教授George Armitage Miller于1985年创建
- 之后由Christiane Fellbaum管理
- 后者创立了Global WordNet Association, 提供了一个讨论、分享、连接世界上不同语言的wordnet的平台

Synsets

- WordNet将词组织成synsets (同义词集合)
 - 每个synset是一个同义词的集合，对应了一种抽象的概念

例：

{dog, domestic_dog, Canis_familiaris}

a member of the genus Canis (probably descended from the common wolf) that has been domesticated by man since prehistoric times; occurs in many breeds; "the dog barked all night"

{pooch, doggie, doggy, barker, bow-wow}

informal terms for dogs

WordNet的规模与版本

- 1989年4月, WordNet中有37409个同义词集合, 没有注释
- 1991年7月, WordNet 1.0版, 包含44983个同义词集合13688个注释(30%)
- 1991年8月, WordNet 1.1版
- 1992年1月, WordNet包含49771个同义词集合, 19382个注释(39%)
- 1992年4月, WordNet 1.2版
- 1992年12月, WordNet 1.3版
- 1993年1月, WordNet包含61023个同义词集合, 36880个注释(60%)
- 1993年8月, WordNet 1.4版

WordNet的规模与版本

- 1994年1月, WordNet中包含79542个同义词集合, 58705个注释(74%)
- 1995年1月, WordNet包括了91050个同义词集合, 同时包含了75389个注释 (占同义词集合数量的83%)
- 1995年3月, WordNet 1.5版
- 1997年, WordNet 1.6版
- 2001年, WordNet 1.7版
- 2001年, WordNet 1.7.1版
-
- 目前 (2024年) 版本: WordNet 3.1

WordNet规模

- WordNet 2.0

POS	Unique	Synsets	Total
	Strings		Word-Sense Pairs
Noun	114648	79689	141690
Verb	11306	13508	24632
Adjective	21436	18563	31015
Adverb	4669	3664	5808
Totals	152059	115424	203145

WordNet词汇的来源

- 利用的已有词表
 - Laurence Urdang (1978)的《同义反义小词典》
 - Urdang (1978)修订的《Rodale同义词词典》
 - Robert Chapmand (1977)的第4版《罗杰斯同义词词林》
 - 美国海军研究与发展中心的Fred Chang的词表, 与WordNet原有词表只有15%的重合词语 (1986)
 - Ralph Grishman和他在纽约大学的同事的一个词表, 包含39143个词, 这个词表实际上包含在著名的COMLEX词典中。WordNet当时词表与该词表重合率为74%(1993年)

Synsets间的关系

- WordNet表示了各种类型的Synsets间关系，如：
- 上下位词关系 (hypernym/hyponym) :
 - {robin, redbreast} @-> {animal, animate_being} @-> {organism, life_form, living_thing}
- 整体与部分关系 (meronym/holonym) :
 - A是B的组成部分: {wing} #p-> {bird}
 - A是B的成员; {tree} #m-> {forest}
 - A是B的构成材料。{aluminum, aluminium, Al} #s-> {aluminum_foil, aluminium_foil}

WordNet的关系指针及标记符号

• 名词

- ! Antonym
- @ Hypernym
- @i Instance Hypernym
- ~ Hyponym
- ~i Instance Hyponym
- #m Member holonym
- #s Substance holonym
- #p Part holonym
- %m Member meronym
- %s Substance meronym
- %p Part meronym
- = Attribute
- + Derivationally related form
- ;c Domain of synset - TOPIC
- -c Member of this domain - TOPIC
- ;r Domain of synset - REGION
- -r Member of this domain - REGION
- ;u Domain of synset - USAGE
- -u Member of this domain - USAGE

WordNet的关系指针及标记符号

• 动词

- ! Antonym
- @ Hypernym
- ~ Hyponym
- * Entailment
- > Cause
- ^ Also see
- \$ Verb Group
- + Derivationally related form
- ;c Domain of synset - TOPIC
- ;r Domain of synset - REGION
- ;u Domain of synset - USAGE

• 形容词

- ! Antonym
- & Similar to
- < Participle of verb
- \ Pertainym (pertains to noun)
- = Attribute
- ^ Also see
- ;c Domain of synset - TOPIC
- ;r Domain of synset - REGION
- ;u Domain of synset - USAGE

• 副词

- ! Antonym
- \ Derived from adjective
- ;c Domain of synset - TOPIC
- ;r Domain of synset - REGION
- ;u Domain of synset - USAGE

动词的关系

- 上下位
 - {pick, pluck} @-> {gather, collect, ...}
- 蕴含 (Entailment)
 - {look} *-> {see}, {buy} *-> {pay}

形容词关系

- 反义词 (Antonym)
 - {wet} !-> {dry}
- 相似 (similar to)
 - {wet} &-> {damp, dampish, moist}

WordNet的应用

- 词义消歧 (Word Sense Disambiguation, WSD)
- 机器翻译
- 信息检索
- 文本分类
- 等等

WordNet的应用

WordNet: a lexical database for English

GA Miller

Communications of the ACM, 1995 • dl.acm.org

Because meaningful sentences are composed of meaningful words, any system that hopes to process natural languages as people do must have information about words and their meanings. This information is traditionally provided through dictionaries, and machine-readable dictionaries are now widely available. But dictionary entries evolved for the convenience of human readers, not for machines. WordNet¹ provides a more effective combination of traditional lexicographic information and modern computing. WordNet is an

SHOW MORE ▾

☆ Save  Cite Cited by 20503 Related articles All 24 versions Web of Science: 6927

WordNet的应用

1. [arXiv:2403.09207](#) [pdf, other] [cs.CL](#)

TaxoLLaMA: WordNet-based Model for Solving Multiple Lexical Semantic Tasks

Authors: Viktor Moskvoretskii, Ekaterina Neminova, Alina Lobanova, Alexander Panchenko, Irina Nikishina

Abstract: In this paper, we explore the capabilities of LLMs in capturing lexical-semantic knowledge from WordNet on the example of the LLaMA-2-7b model and test it on multiple lexical semantic tasks. As the outcome of our experiments, we present TaxoLLaMA, the everything-in-one model, lightweight due to 4-bit quantization and LoRA. It achieves 11 SotA results, 4 top-... [▽ More](#)

Submitted 14 March, 2024; **originally announced** March 2024.

Comments: 18 pages, 8 figures

2. [arXiv:2402.13302](#) [pdf] [cs.CL](#)

Enhancing Modern Supervised Word Sense Disambiguation Models by Semantic Lexical Resources

Authors: Stefano Melacci, Achille Globo, Leonardo Rigutini

Abstract: ...Lexical Resources (SLRs) is mostly restricted to knowledge-based approaches. In this paper, we enhance "modern" supervised WSD models exploiting two popular SLRs: WordNet and WordNet Domains. We propose an effective way to introduce semantic features into the classifiers, and we consider using the SLR structure... [▽ More](#)

Submitted 20 February, 2024; **originally announced** February 2024.

Comments: The 11th International Conference on Language Resources and Evaluation (LREC 2018)

Journal ref: Proceedings of The 11th International Conference on Language Resources and Evaluation (LREC 2018)

3. [arXiv:2402.01720](#) [pdf] [cs.CY](#) [cs.AI](#) [cs.CL](#) [cs.LG](#) [doi](#) [10.48550/arXiv.2402.01720](#)

Deep Learning Based Amharic Chatbot for FAQs in Universities

Authors: Goitom Ybrah Hailu, Shishay Welay

Abstract: ...and effectively addressed challenges such as Amharic Fidel variation, morphological variation, and lexical gaps. Future research could explore the integration of Amharic WordNet to narrow the lexical gap and support more complex questions. [▽ More](#)

Submitted 26 January, 2024; **originally announced** February 2024.

Report number: AksumUniv-CS-2024

Journal ref: Machine Learning (cs.LG), V1, 2024

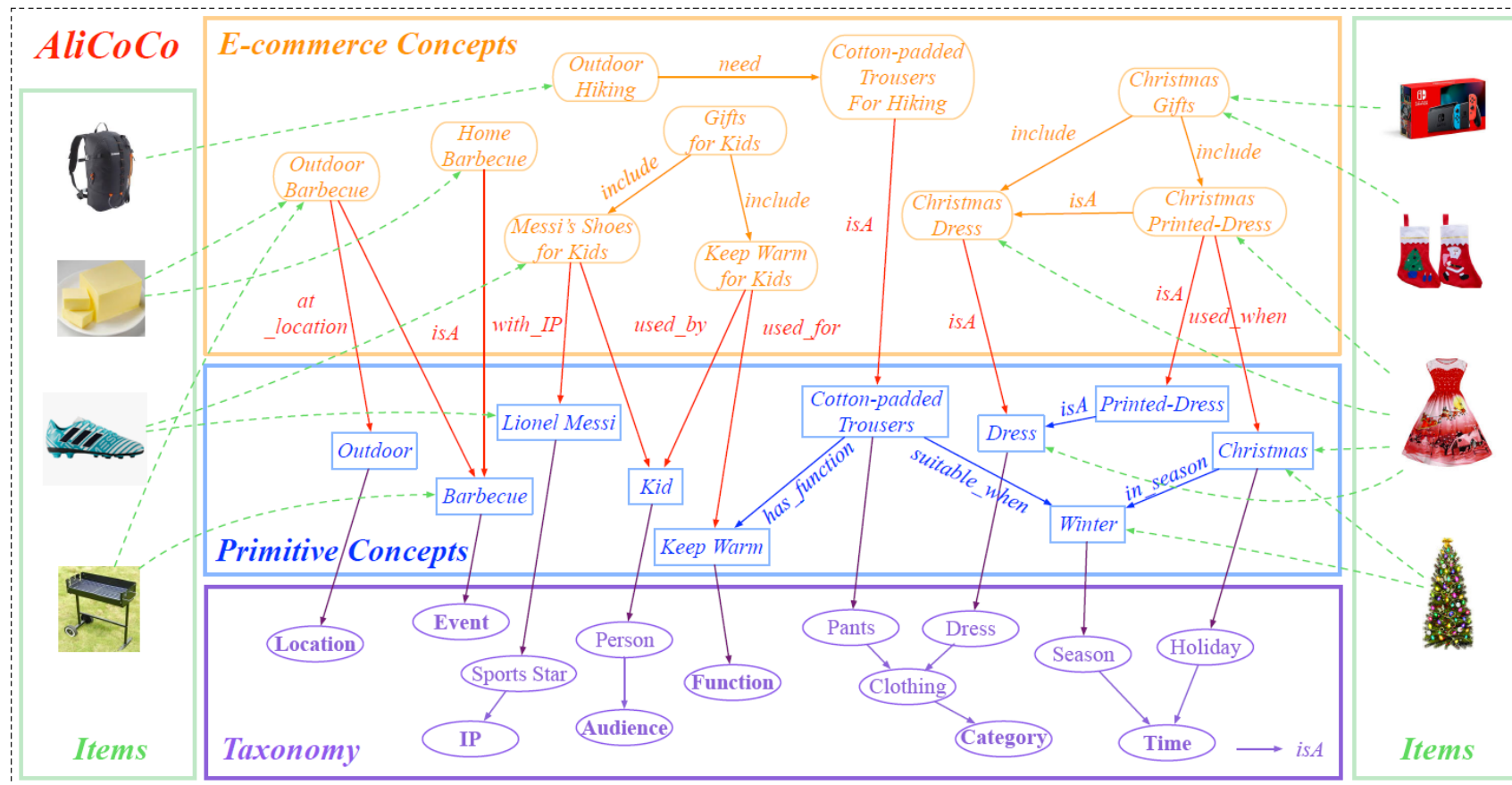
Open Multilingual Wordnet

Summary of Wordnets

Wordnet	Lang	Synsets	Words	Senses	Core	Licence
Albanet	als	4,675	5,988	9,599	31%	CC BY 3.0
Arabic WordNet (AWN v2)	arb	9,916	17,785	37,335	47%	CC BY SA 3.0
BulTreeBank Wordnet (BTB-WN)	bul	4,959	6,720	8,936	99%	CC BY 3.0
Chinese Open Wordnet	cmn	42,312	61,533	79,809	100%	wordnet
Chinese Wordnet (Taiwan)	cmn	4,913	3,206	8,069	28%	wordnet
DanNet	dan	4,476	4,468	5,859	81%	wordnet
Greek Wordnet	ell	18,049	18,227	24,106	57%	Apache 2.0
Princeton WordNet	eng	117,659	148,730	206,978	100%	wordnet
Persian Wordnet	fas	17,759	17,560	30,461	41%	Free to use
FinnWordNet	fin	116,763	129,839	189,227	100%	CC BY 3.0
WOLF (Wordnet Libre du Français)	fra	59,091	55,373	102,671	92%	CeCILL-C
Hebrew Wordnet	heb	5,448	5,325	6,872	27%	wordnet
Croatian Wordnet	hrv	23,120	29,008	47,900	100%	CC BY 3.0
MultiWordNet	ita	35,001	41,855	63,133	83%	CC BY 3.0

Alicoco

- 电商本体



END
