PLIN021 Sémantická analýza v praxi

OP VK Mezi bohemistikou a informatikou
www.projekt-inova.cz

Zuzana Nevěřilová
xpopelk@fi.muni.cz

Centrum zpracování přirozeného jazyka, B203
Fakulta informatiky, Masarykova univerzita

22. dubna 2013

Ontologie

O. je značně nadužívaný pojem, v informatice znamená „formální a explicitní specifikaci sdílené konceptualizace“ [Gruber, 2009]

- formální
- explicitní
- sdílené pojmy

O. je značně nadužívaný pojem, v informatice znamená „formální a explicitní specifikaci sdílené konceptualizace“ [Gruber, 2009]

- formální
- explicitní
- sdílené pojmy

Ontologie se skládá z uvedených součástí a má uvedené vlastnosti, jinak ale může mít libovolný „tvar“. O. můžeme chápat jako nadpojem pro taxonomie, sémantické sítě atd. Bohužel kvůli nadužívání termínu v mnoha oblastech se setkáme s odmítáním zařadit určité projekty pod pojem ontologie. Vždy, když se hovoří o o., je potřeba ujasnit si, co tím myslíme. Nám v tomto kurzu bude stačit tento jednoduchý pohled a budeme se soustředit hlavně na různé „tvary“ a využití o.

Ontologie

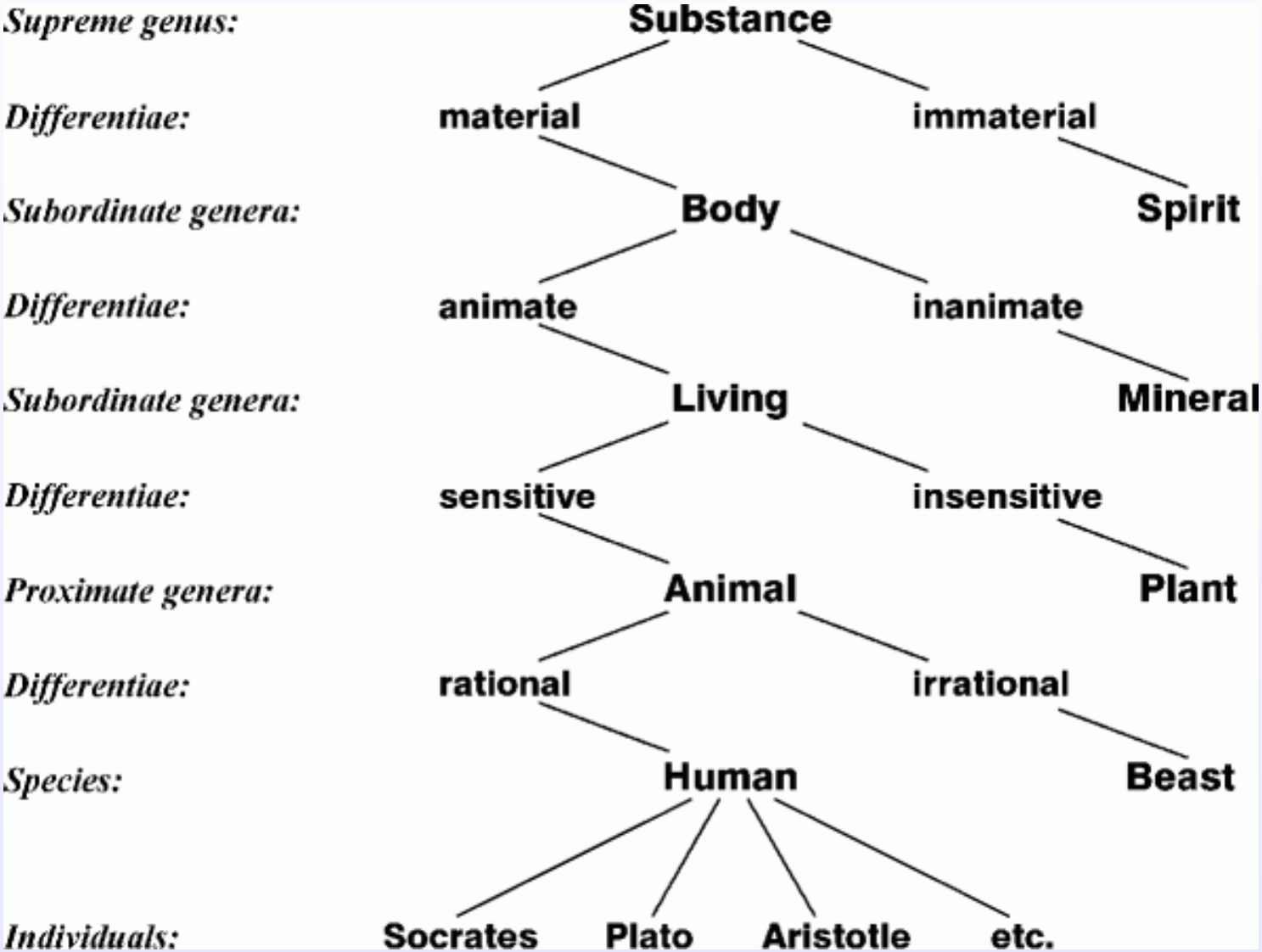
- slovník (glosář, inventář pojmů . . .)
- taxonomie (tezaurus, inventář relací . . .)

Taxonomie (stromy)

- Aristoteles – kategorie (všech) entit, které mohou lidé vnímat
- Porfyrios – uspořádal kategorie
- Carl Linné – klasifikace (všech) organismů

důležité rysy: uzly jsou třídy (organismů, entit . . .), třídy jsou strukturované do stromu (podtřída, nadtřída), uzly na stejné úrovni se vzájemně vylučují (implicitní předpoklad)

Porfyriův strom (John. F. Sowa)



Taxonomie (stromy)

relace is a

relace member of

třída $\times$ instance

Pes je masožravec, nepohrdne však ani ovocem.

Alík má rád švestky.

Americký prezident je zároveň předsedou vlády.

Americký prezident má babičku v Africe.

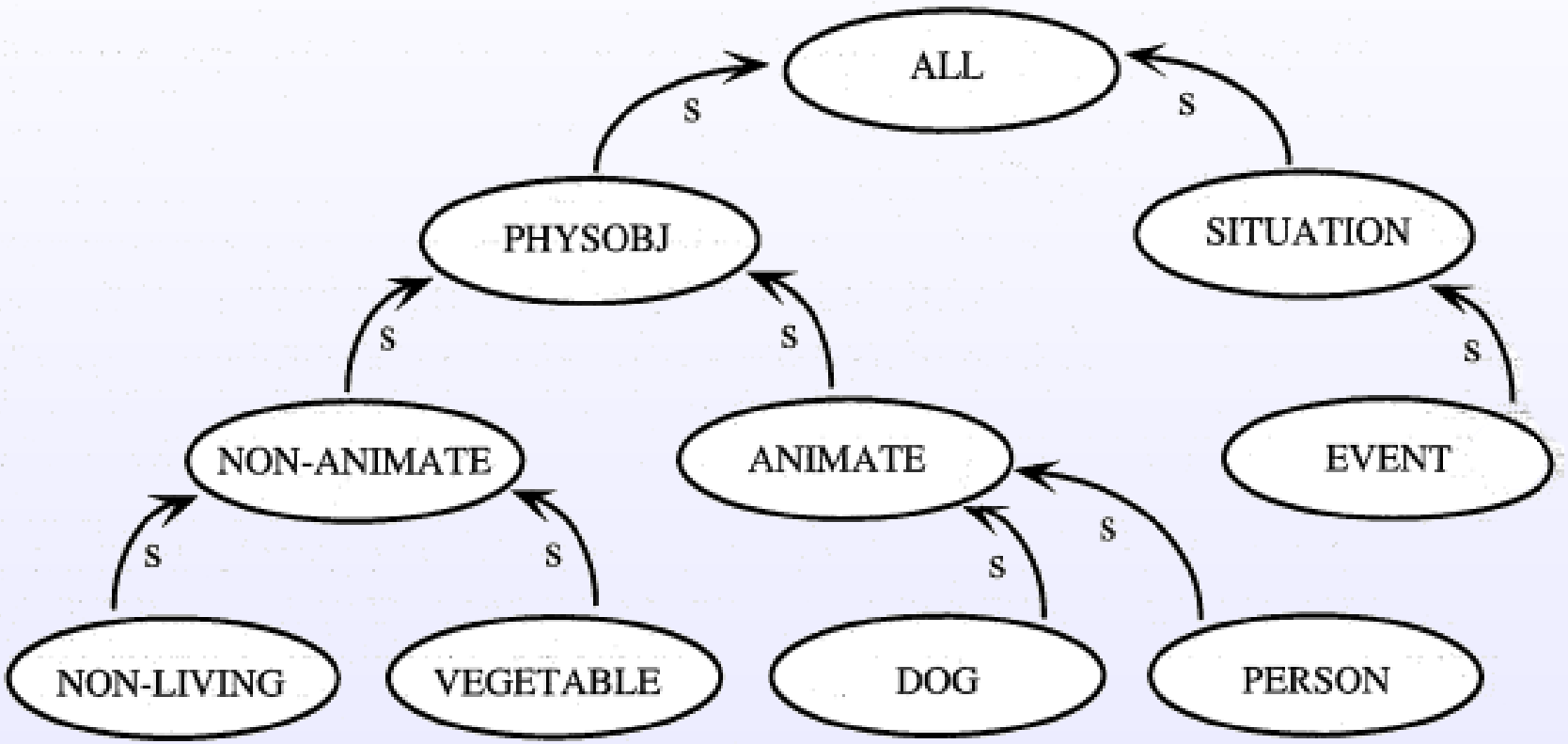
Taxonomie (stromy) a slovníkové definice

klasická definice = genus proximum + differentia specifica

Počítač je v informatice elektronické zařízení, které zpracovává data pomocí předem vytvořeného programu.

Elektronické zařízení je zařízení, jehož funkce závisí na elektrickém proudu nebo na elektromagnetickém poli.

Sémantické sítě I



sémantická síť = reprezentace lexikálních znalostí
[Collins and Quillian, 1969]

Sémantické sítě II

uzly = entity (třídy nebo instance), jednomu konceptu odpovídá jeden uzel

hrany = vztahy mezi uzly (binární relace)

Sémantické sítě

- nadtyp–podtyp, is a, is-a, isa (hypo/hyperonymie)
- instance třídy, member of
- část–celek, has a (holo/meronymie)
- upřesnění akce (troponymie)
- příčina–následek
- ...

- nadtyp–podtyp, is a, is-a, isa (hypo/hyperonymie)
- instance třídy, member of
- část–celek, has a (holo/meronymie)
- upřesnění akce (troponymie)
- příčina–následek
- ...

Jednotlivé podsítě, kde uzly spojují relace jednoho druhu, jsou stromy (tj. taxonomie). Je to docela logické – v každém druhu relace máme nějaké uspořádání „od nejmenšího po největší“.

Např. nadtyp–podtyp je klasická taxonomie. Část–celek taky, protože objekt x se skládá z částí a a b, část a se skládá z částí m a n (které jsou patrně menší než a i menší než x).

- nadtyp–podtyp, is a, is-a, isa (hypo/hyperonymie)
- instance třídy, member of
- část–celek, has a (holo/meronymie)
- upřesnění akce (troponymie)
- příčina–následek
- ...

Důležitou vlastností taxonomií (tj. i těch částí sém. sítě, které tvoří taxonomii) je tranzitivita. Využíváme ji v odvozování (viz dál).

Odbočka k odvozování

Fakt F = tvrzení s pravdivostní hodnotou (např. ptáci létají)
Báze znalostí (knowledge base) KB = (pokud možno konzistentní) soubor faktů (např. ptáci létají, vlaštovka je pták)
Pokud z KB plyne F a přidáme další fakt takový, že KB je stále konzistentní, je KB monotónní reprezentace.[Allen, 1995]

ptáci létají
vlaštovka je pták

---

vlaštovka létá

Odbočka k odvozování

ptáci létají
tučňák je pták

---

tučňák létá

kromě tučňáka ptáci létají
tučňák je pták

NOT(tučňák létá)

kromě tučňáka ptáci létají
pštros je pták

---

pštros létá

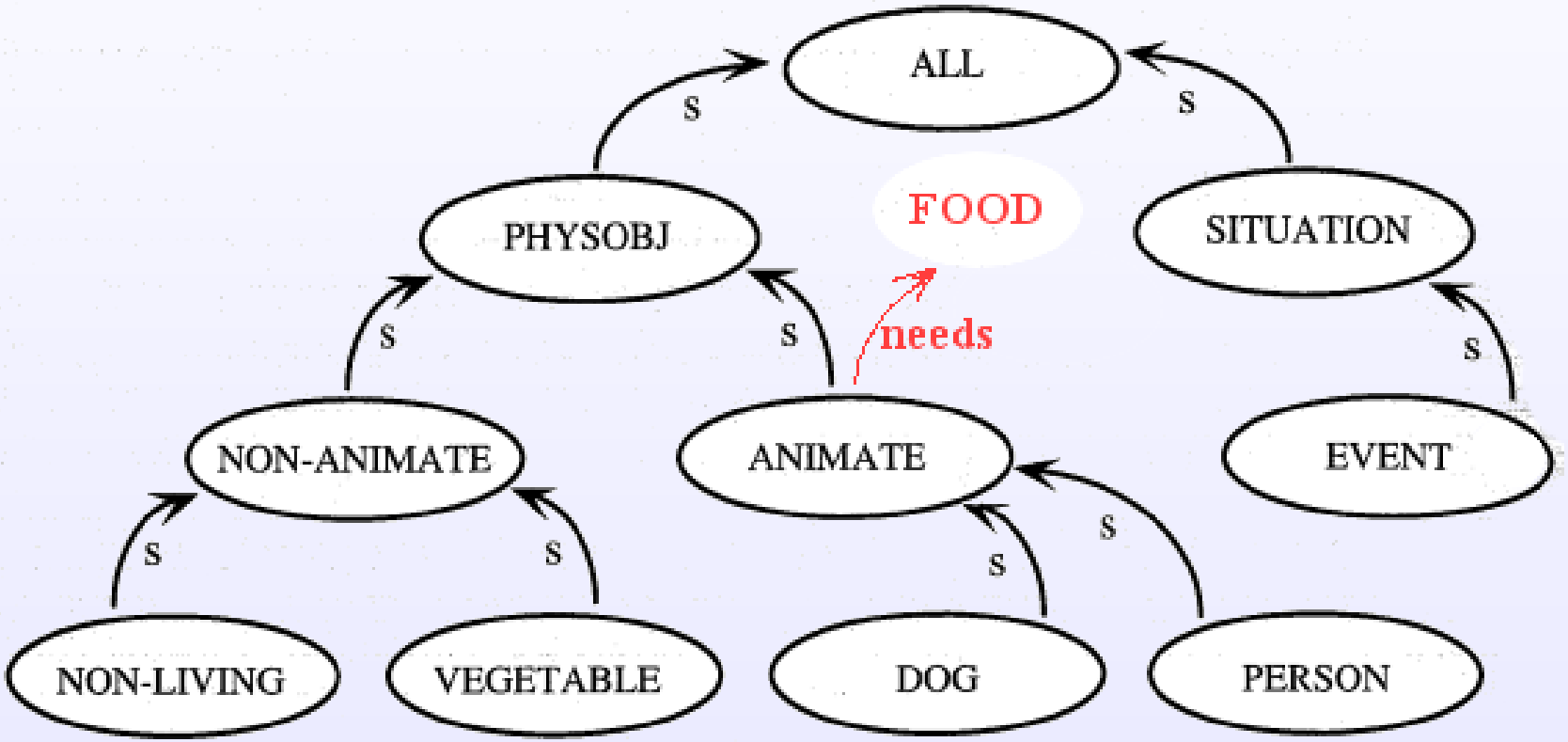
kromě tučňáka, pštrosa, mláďat, mrtvých ptáků, ptáků se zraněnými křídly . . . ptáci létají

Odbočka k odvozování

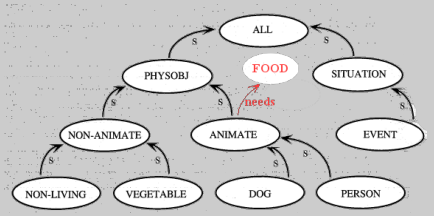
Používáme implicitní pravidlo (*default rule*), tj. ptáci létají, dokud neřekneme jinak. Uvedeme-li implicitní pravidlo, má přednost před obecným faktem.

Ptáci létají, ale tučňák ne.

Sémantické sítě – dědičnost



odvozování je monotónní



odvozování je monotónní

Praktická ukázka dědičnosti a odvozování:

1. silniční vozidlo má (has part) volant
2. dodávka je (isa) silniční vozidlo
3. dodávka má (has part) volant
4. Mercedes Sprinter je (member of) dodávka
5. Mercedes Sprinter má (has part) volant

Sémantické sítě

WordNet a EuroWordNet

český WordNet

Sítě z Wikipedie, dbpedia

Rámce – použití

Rámce můžeme použít pro desambiguaci slov i celých vět [Laparra and Rigau, 2009].

[Bernard Lansky]$_{STUDENT}$ *studied* [the piano]$_{SUBJECT}$ [with Peter Wallfisch]$_{TEACHER}$.

Rámce – použití

Rámce můžeme použít pro doplnění implicitní (nezmiňované) znalosti.

Koupila jsem ojetou felicii. Byly to vyhozené peníze.

koupit:

- má_činitele člověk/instituce/skupina
- má_benefaktora člověk/instituce/skupina
- má_předmět výrobek/nemovitost/zvíře/rostlina/přírodnina
- má_část činitel dá peníze
- má_část benefaktor dá předmět

Rámce se používají hodně. FrameNet nebo VerbaLex jsou zajímavé i svým velkým rozsahem, nejsou to jen nějaké experimenty s uměle vybranými jevy.

Rámce – použití

Rámce můžeme použít pro doplnění implicitní (nezmiňované) znalosti.

Koupila jsem ojetou felicii. Byly to vyhozené peníze.

koupit:

- má_činitele člověk/instituce/skupina
- má_benefaktora člověk/instituce/skupina
- má_předmět výrobek/nemovitost/zvíře/rostlina/přírodnina
- má_část činitel dá peníze
- má_část benefaktor dá předmět

Skripty, scénáře (Abelson)

skript: v restauraci, prvky skriptu mohou být rámce

- host (člověk, není v zaměstnání, má u sebe *peníze*, sedí na *židli*, jí *jídlo*)
- číšník (člověk, je v zaměstnání)
- kuchař (člověk, je v zaměstnání)
- místnost (obsahuje *židle*, stoly, příjemnou teplotu)
- jídlo (uvařil *kuchař*, donesl *číšník hostovi*)
- peníze (zaplatil *host číšníkovi* za *jídlo*)

„Pepovi u večeře zazvonil telefon. Chvíli poslouchal, pak položil telefon a opustil restauraci.“

Skripty, scénáře (Abelson)

„Pepovi u večeře zazvonil telefon. Chvíli poslouchal, pak položil telefon a opustil restauraci.“

Předpokládáme, že mezi „položil telefon“ a „opustil restauraci“ se stalo:

- Číšník donesl účet.
- Pepa zaplatil.
- Pepa se oblékl.

Skripty, scénáře (Abelson)

skript: v restauraci

- host (člověk, není v zaměstnání, má u sebe *peníze*, sedí na *židli*, jí *jídlo*)
- číšník (člověk, je v zaměstnání)
- kuchař (člověk, je v zaměstnání)
- místnost (obsahuje *židle*, stoly, příjemnou teplotu)
- jídlo (uvařil *kuchař*, donesl *číšník hostovi*)
- peníze (zaplatil *host číšníkovi* za *jídlo*)

Usuzování v rámcích může být implicitní (podobné jako v sém. sítích) i speciální pro daný rámec.
Usuzování v rámcích může být nemonotónní.

Příklad: host zaplatil $\Rightarrow$ číšník má u sebe peníze

Příklad: každý host musí zaplatit svoji útratu.

Skripty, scénáře (Abelson)

skript: v restauraci

Skripty popisují typické situace. Stereotypická je i informace o zaplněnosti slotů, např. restaurace musí mít číšníka.

Pořadí ve scénáři je chronologické: host přijde do restaurace, objedná si jídlo, kuchař jídlo uvaří, host sní jídlo, host zaplatí číšníkovi. . .

Můžeme nějak měřit vybočení ze stereotypu?


Allen, J. (1995).
Natural Language Understanding (2nd ed.).
Benjamin-Cummings Publishing Co., Inc., Redwood City, CA, USA.


Collins, A. M. and Quillian, M. R. (1969).
Retrieval time from semantic memory.
Journal of Verbal Learning and Verbal Behavior, 8(2):240–247.


Gruber, T. (2009).
Ontology.
In Liu, L. and Özsu, M. T., editors, *Encyclopedia of Database Systems*, page 1963–1965. Springer Verlag.


Laparra, E. and Rigau, G. (2009).
Integrating wordnet and framenet using a knowledge-based word sense disambiguation algorithm.
In *RANLP*, Borovets, Bulgaria.