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

11. dubna 2012

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 – 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.

zjistit, jak je na tom český FrameNet (PhD práce J. Materny)

Ontologie

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

- formální
- explicitní
- sdílené pojmy

Ontologie

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

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

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.

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)

- 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)

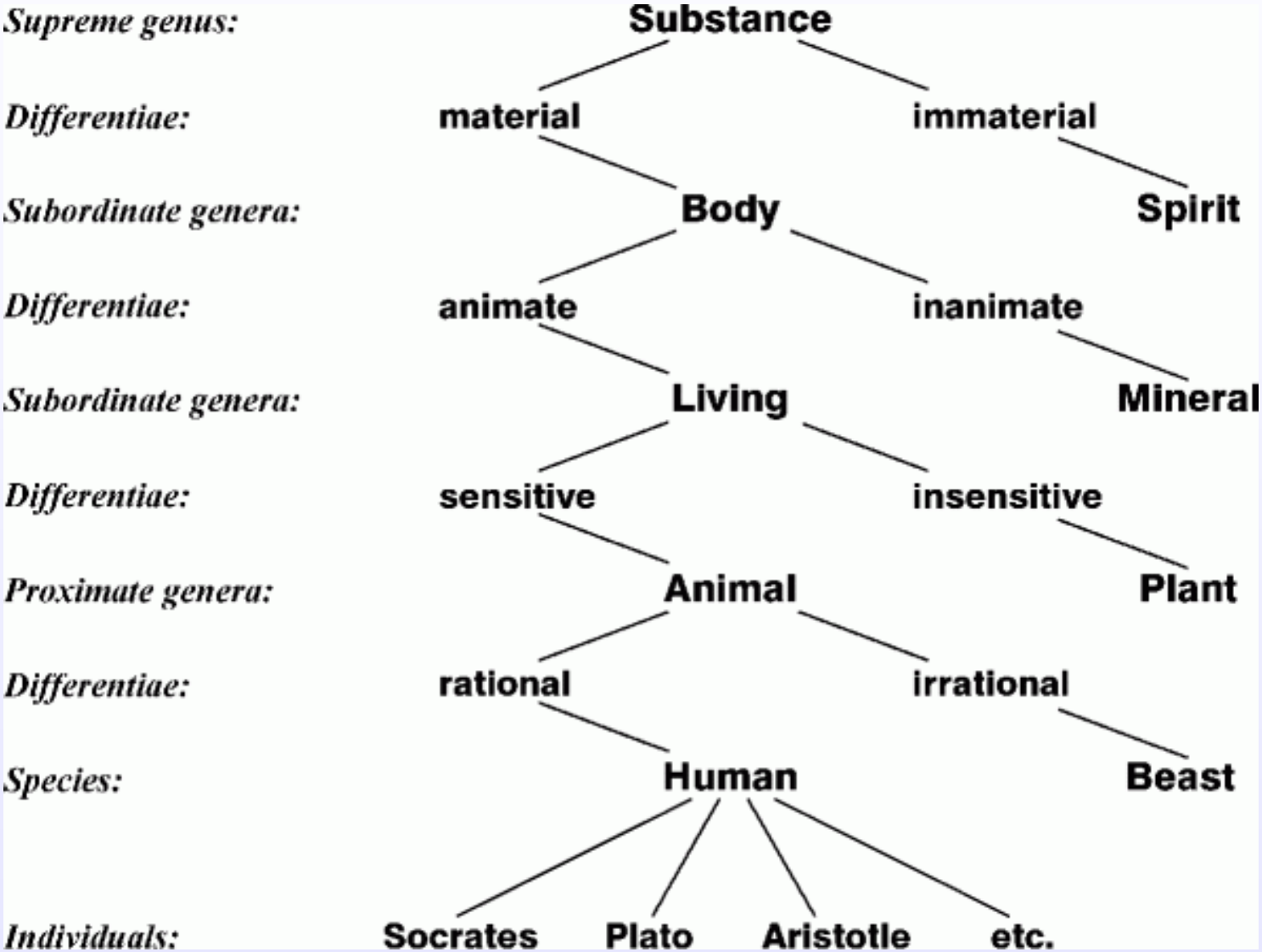
opět – každý z nich má pro své dílo zcela jiné motivy, výsledek je ale docela podobný...

- 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)

Najít dobré ukázky (obrázky k výše uvedeným). U Aristotela je těch kategorií 10: substance, kvantita, kvalita, relace, místo, čas, bytí na pozici, bytí ve stavu, dělání, ovlivnění

Porfyriův strom (John. F. Sowa)



Taxonomie (stromy)

relace is a

relace member of

relace is a

relace member of

Připomenout rozdíl mezi třídou a instancí. Jak je v přirozeném jazyce rozeznáváme? Například podle jména:

Pes je masožravec, nepohrdne však ani ovocem.

Alík má rád švestky.

V mnoha případech je to složitější:

Americký prezident je zároveň předsedou vlády.

Americký prezident má babičku v ohrožení.

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.

klasická definice = genus proximum + differentia specifica

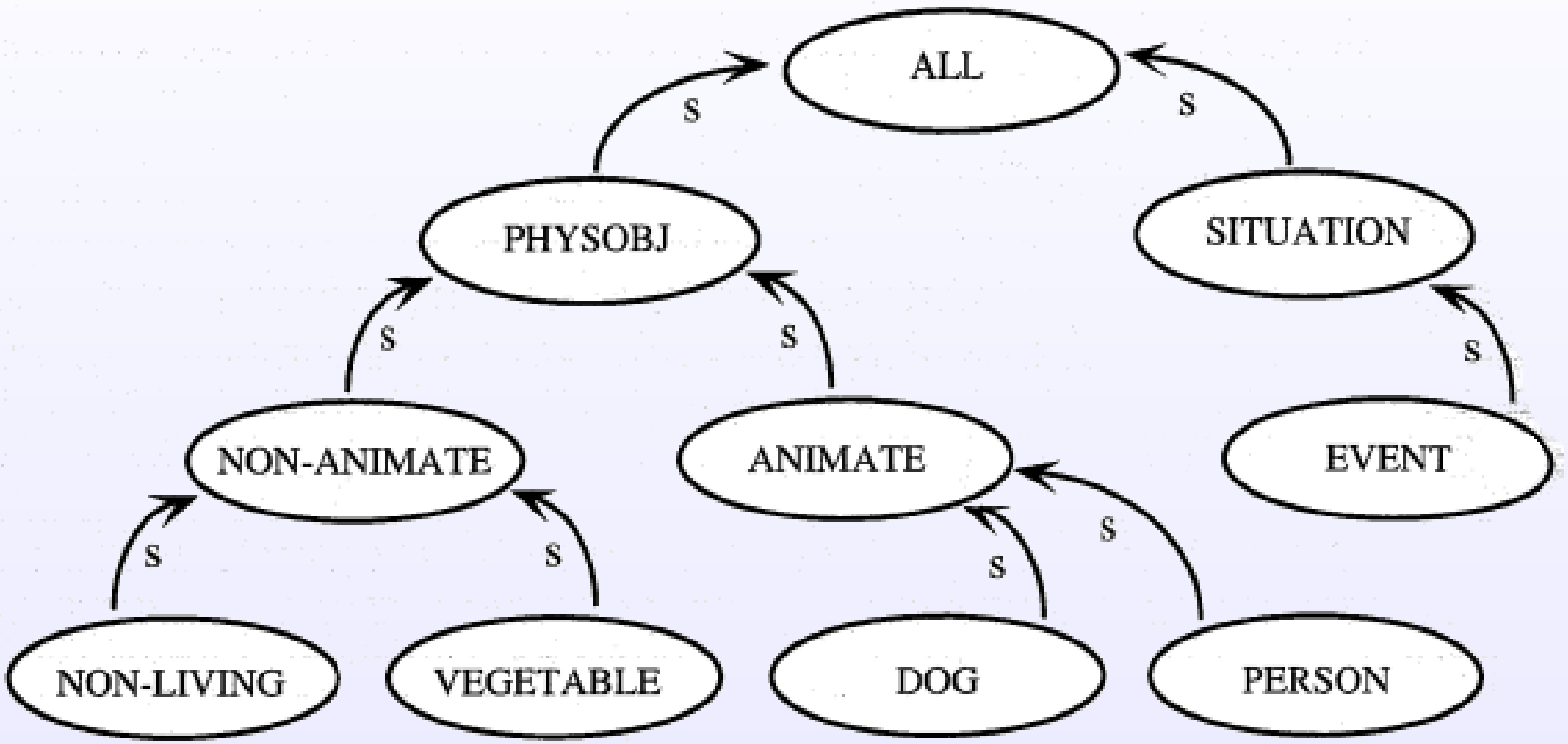
Počítač je v informatice elektronické zařízení, které zpracovává data pomocí předem vytvořeného programu.

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

Nabízí se možnost zpracování encyklopedií jako bází znalostí. Problém je, že málokterá „lidská“ encyklopedie je počítačově zpracovatelná. Teď jsme už aspoň ve fázi počítačově čitelných.

Existuje definice a definice. Některé jsou vědecké, jiné jsou common-sense. V encyklopediích najdeme častěji ty vědecké. Příklad z Encyclopaedia Britannica: papír je tenký, hladký list z dřevitého vlákna vyrobený na drátěném sítu vs. WordSmyth: papír je tenký materiál, který se používá k psaní nebo balení věcí

Sémantické sítě



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

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

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

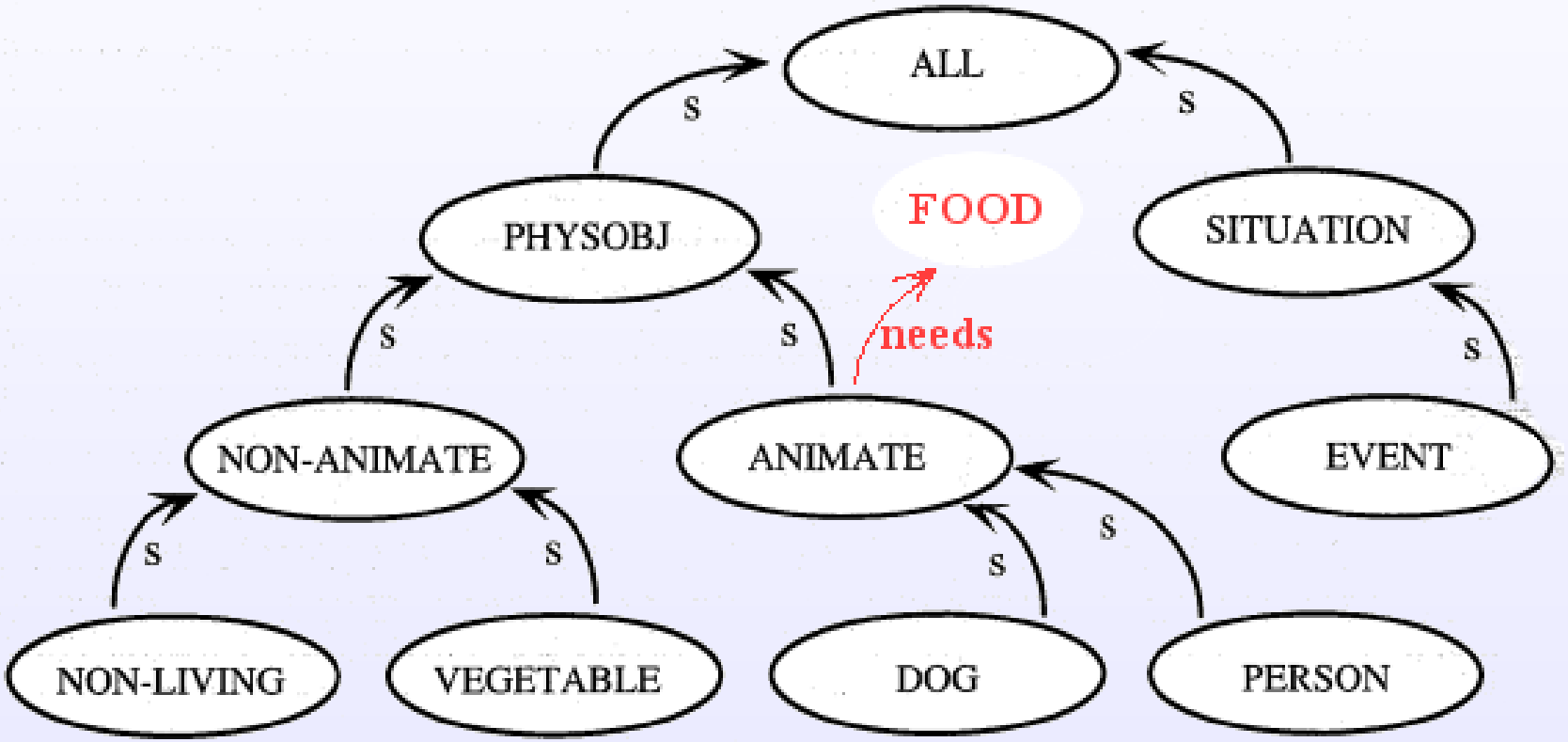
kromě tučňáka, pštrosa, mláďat, mrtvých ptáků, ptáků se zraněnými křídly . . . ptáci létají

Odbočka k odvozování

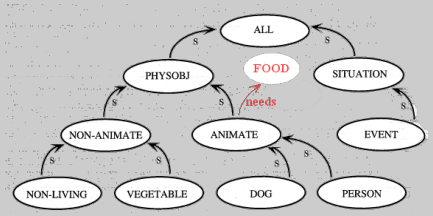
Používáme implicitní pravidlo (*default rule*), tj. ptáci létají, dokud neřekneme jinak. Uvedeme-li implicitní pravidlo, má přednost před obecným faktem.

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

Podíváme se, jestli uvedené „lezení po větvích“ platí pokaždé.

Sémantické sítě

WordNet a EuroWordNet

český WordNet

Sítě z Wikipedie, dbpedia, ArtNet

Nachystat ukázky. Český WordNet je jistě dobré téma pro BP. Podobně česká dbpedia (která dosud neexistuje).

Asociativní sítě

Někdy totožné se sémantickými sítěmi, jindy u asociativních sítí neplatí předpoklad, že jeden koncept odpovídá jednomu uzlu.

Někdy totožné se sémantickými sítěmi, jindy u asociativních sítí neplatí předpoklad, že jeden koncept odpovídá jednomu uzlu.

Uvedený předpoklad typicky neplatí u sítí, které jsou generovány automaticky.

Někdy totožné se sémantickými sítěmi, jindy u asociativních sítí neplatí předpoklad, že jeden koncept odpovídá jednomu uzlu.

Sémantické (asociativní) sítě se používají velmi mnoho. Je jich mnoho druhů, je mnoho způsobů jejich zápisu, jsou v módě (protože sémantický web je v módě a protože web tvoří také síť).

V souvislosti se sémantickým webem můžeme zmínit jazyky sémantického webu: RDF, RDFS, OWL. Podrobněji se jim můžeme věnovat, pokud bude čas (a chuť). Trochu jsme nakousli odvozování. Podle experimentů z kognitivní vědy (dohledat) lidé nemají v hlavě všechno, ale odvozují. Oblast odvozování znalostí ze znalostních bází by vydala na zvláštní seminář, přesto se odvozování nemůžeme úplně vyhnout. Logická reprezentace bývá zpravidla chápána mimo jazyk, přesto cítíme, že je neoddělitelnou součástí promluvy.


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.