

# PLIN021 Sémantická analýza v praxi

OP VK Mezi bohemistikou a informatikou  
[www.projekt-inova.cz](http://www.projekt-inova.cz)

Zuzana Nevěřilová  
[xpopelk@fi.muni.cz](mailto:xpopelk@fi.muni.cz)

Centrum zpracování přirozeného jazyka, B203  
Fakulta informatiky, Masarykova univerzita

4. března 2013

Víceznačnost, granularita

Redundance

Word Sense Disambiguation

Algoritmy WSD

## Je víceznačnost problém v NLP?

rychlý  $\Rightarrow$  fast

auto  $\Rightarrow$  car

---

rychlé auto  $\Rightarrow$  fast car

vysoký  $\Rightarrow$  high

škola  $\Rightarrow$  school

---

vysoká škola  $\Rightarrow$  university?

## Granularita významu (sense): kočka

- 2. malá n. středně velká šelma s hustým kožichem; zool. rod Felis
  - 1. malá kočkovitá šelma, chovaná v domácnostech
  - 3. samice kočkovité šelmy vůbec
- 4. ob. kožišina na límci, kolem krku n. ramen
- 6. věc připomínající někt. vlastnost kočky
  
- 5. kocovina (Haš.)
  
- 7. druh důtek

## Jak stanovit kritéria pro „tu správnou“ granularitu?

podle syntaktických kritérií:

- Lord *zanechal* v závěti všechn svůj majetek místnímu sirotčinci.
- Student *zanechal* studia podáním písemné žádosti.

# Jak stanovit kritéria pro „tu správnou“ granularitu?

podle sémantických kritérií:

- abstraktní × konkrétní
- životný × neživotný
- člověk × zvíře
- emoce
- doména

## Frekvence užívání slova v daném významu

holub a opilý živý tulák. Mizerná pouliční **kočka** je zkoumavě pozorovala. Jako by si spojoval

esba Lukáš Fibrich (21. 8. **kočku** Asi málokdo by hledal největší světové

hledal největší světové muzeum věnované **kočkám** právě na Borneu v Malajsii. Pokud někdy

Malajsii. Pokud někdy náhodou navštívíte město **Kočka** , pak si tuhle atrakci rozhodně nenechte

, bude obrovské sousoší tvořené několika **kočkami** . Vydáte-li se procházkou po nábřeží řeky

směrem, záhy narazíte na další kočičí sochy. **Kočkami** - těmi umělými i těmi živými - je město

. Důvod je jednoduchý - Kuching je město **kočkám** zaslíbené. **NEOBVYKLÉ MUZEUM**

mívají krásné husté a delší vlasy. Jako **kočky** předou, když je budete hladit, česat, mírně

smyslem pro humor obklopují se zvířítka - psi, **kočky** , drobné šelmy, tak se staňte znalci tohoto

dáte za pravdu, že letušky byly opravdu **kočky** . U American Airlines nic takového nečekejte

děkuji Katko za prima plavečky, já budu **kočka** , jako vždy super obchůdek, jen chválím

## Redundance v přirozeném jazyce

Tto vtě určitě porozumte přeštoze nní správná.

Studentky se na semináři nudily. Myslete dopředu na finanční zajištění vlastního pohřbu. Ušetříte tím starosti svým blízkým.



## Redundance v přirozeném jazyce

- na úrovni hlásek (některé hláskové sekvence se v daném jazyce nevyskytují)
- na syntaktické úrovni (shoda přísudku s podmětem)
- na úrovni významů (pleonasmy, tautologie)



# Word Sense Disambiguation

Lexikální desambiguace: nalezení významu slova v daném kontextu.

Pro člověka podvědomé, pro počítače „AI complete“.

WSD = „... the problem of computationally determining which “sense” of a word is activated by the use of the word in a particular context.“ [Agirre and Edmonds, 2006]

klasifikační úloha:

- jednotlivé **významy** tvoří **třídy**
- podle **kontextu** se **rozhodujeme**, do kterých tříd **slovo na vstupu** patří

předpoklady: významy jsou **diskrétní** a je jich **konečný počet**, máme nějaký **inventář** významů

- je d'offrir vous my tout tout
- je te t'offrir se me tout que, je te t'offrir t'offrir tout se
- offrir se me

AI complete = aby bylo možné vyřešit tuto úlohu, je potřeba vyřešit všechny. Výraz „AI complete“ je asociací s „NP complete“. Jde o třídu těžkých problémů teoretické informatiky. Těžký znamená neřešitelný v polynomiálním čase.

Příklad NP-úplného problému je batoh: mám batoh o objemu  $x$ , hodnotící funkci dobře zaplněného batohu  $f$  (definovanou jakkoli, ale samozřejmě objektivně) a věci, které chci dát do batohu. Obecně je objem věcí v součtu větší než  $x$ . Jak dobře zaplnit batoh, tj. jak maximalizovat  $f$ ? V praxi se při balení batohu spokojíme s dobrým výsledkem, přestože možná existuje lepší, ale ujel by nám vlak, než bychom na něj přišli. Pokud se nespokojíme s přibližným výsledkem, musíme zkusit všechny možnosti.

# Word Sense Disambiguation: aplikace

- strojový překlad – machine translation (MT)
- inteligentní vyhledávání – information retrieval (IR)
- inteligentní korektor překlepů
- ...

- strojový přelož - machine translation (MT)
- inteligentní vyhledávání - information retrieval (IR)
- inteligentní textové překlady
- ...

Ukázky toho, když počítačový program předpokládá jiný význam? Všichni známe strojově přeloženou Wikipedii, kde byl např. Ludvík dodávka Beethoven (ad 1). Známe také situaci, kdy hledáme na Google nějaký mnohoznačný termín a většinu stránek ignorujeme (můžeme zkusit hledat „evropský los“, ad 2). Kdybychom tak měli korektor, který by odhalil chyby typu „Palivo z potopené Concordie budou přečerpávat do speciálních palivových lidí.“ (ad 3)...

# Word Sense Disambiguation: přístupy

zpravidla ignorují teoretické, psychologické, logické aj. aspekty významu

- hloubkové (zahrnující znalosti o jazyce i o světě)
- povrchové přístupy (bez dalších znalostí, počítají s okolím)

- Hluboké přístupy (základní znalosti o jazyce i o světě)
- Povrchové přístupy (bez základních znalostí, pouze o okolím)

V současnosti mají vrch povrchové přístupy, protože jsou lepší než nic a relativně levné. Hlubkové přístupy nejsou funkční, protože si žádají zpracovat a formalizovat příliš mnoho dat, což je zdlouhavé a drahé. Hlubkové přístupy jsou funkční v omezených doménách, protože tam je „svět“ malý a experty se vyplatí platit.

# Word Sense Disambiguation: přístupy

historický přístup („expertní“): v jakých kontextech může slovo nabývat jakých významů?

kohoutek

- botanika: rostlina
- chovatelství: lopatková kost
- technické vybavení budov: ruční uzávěř
- ...



## └ Word Sense Disambiguation

## └ Word Sense Disambiguation: přístupy

historie té přísluš [experten]: v jakých kontextech může slovo  
nazývat jakých významů?

kontext

- botanika: ušlech
- s hrozbou: lepatou šest
- tet křídla: vyžvení kálov: rati: zvěř
- ...

Expertní přístup je příliš náročný: kdo dokáže najít a vyjmenovat všechny možné kontexty (které navíc stále přibývají)?

# Algoritmy založené na znalostech

historický start: strojově čitelné slovníky (Machine Readable Dictionaries)

reprezentant: Leskův algoritmus (1986) = slovo  $w$ , jehož okolí sdílí nejvíc slov s definicí (nebo s příklady užití) itého významu, má význam  $s_i$

[Kilgarriff and Rosenzweig, 2000]

## Naivní Leskův algoritmus: kočka (SSJČ)

1. malá kočkovitá šelma, chovaná v domácnostech, na venkově zvl. pro hubení myší; kočka domácí (zool.); šedivá, černá, třibarevná k.; hladká srst kočky; k. mňouká, přede; k. číhá na myš; k. chytá ptáky; angorská k.; být falešný, úlisný jako k.; přen. expr. je to k. falešník; to děvče je k. lichotné, úlisné; [x] jsou na sebe jako pes a k. nenávidí se. . .
2. malá n. středně velká šelma s hustým kožichem; zool. rod Felis: k. plavá; k. divoká; k. domácí
3. samice kočkovité šelmy vůbec; rysí k.; lví k.; expr. každá kočkovitá šelma vůbec (tygr, levhart aj.)
4. ob. kožíšina na límci, kolem krku n. ramen
5. kocovina (Haš.)
6. věc připomínající někt. vlastnost u kočky: bot. velký trs ostřic vystupující z rašeliniště (na blatech); tech. pojízdný vozík jeřábu se zdvihacím ústrojím
7. druh důtek; devíticásá k.

## Naivní Leskův algoritmus: vstup

Aminokyselina DL-methionin okyseluje moč, čímž chrání močové ústrojí psů i *koček* (důležitá vlastnost zvláště u kastrovaných jedinců).

{aminokyselina, DL-methionin, okyselovat, moč, čímž, chránit, močový, ústrojí, pes, i, důležitý, vlastnost, zvláště, u, kastrovaný, jedinec}

{aminokyselina, což, DL-methionin, důležitý, chránit, i, jedinec, kastrovaný, moč, močový, okyselovat, pes, u, ústrojí, vlastnost, zvláště}

## Leskův algoritmus: naivní

{aminokyselina, což, DL-methionin, důležitý, chránit, i, jedinec, kastrovaný, moč, močový, okyselovat, **pes**, **u**, **ústrojí**, **vlastnost**, **zvláště**}

**1:** {a, angorský, být, černý, číhat, děvče, domácí, domácnost, expresivně, falešník, falešný, hladký, hubení, chovaný, chytat, jako, kočkovitý, lichotný, malý, mňoukat, myš, na, nenávidět, **pes**, pro, přeneseně, příst, pták, se, srst, šedivý, šelma, to, třibarevný, úlisný, v, venkov, zoologicky, **zvláště**}

**2:** {divoký, domácí, Felis, hustý, kožich, malý, nebo, plavý, rod, s, středně, šelma, velký, zoologicky}

⋮

**6:** {bláto, botanicky, jeřáb, na, některý, ostřice, pojízdný, připomínající, rašelinště, s, technicky, trs, **u**, **ústrojí**, věc, velký, vozík, **vlastnost**, vystupující, z, zdvihací}

**7:** {devíticásá, druh, důtky}

## Leskův algoritmus: naivní

{aminokyselina, což, DL-methionin, důležitý, chránit, i, jedinec, kastrovaný, moč, močový, okyselovat, pes, u, ústrojí, vlastnost, zvláště}

$$D_1 = \{\text{pes, zvláště}\}$$

$$D_2 = \{\}$$

$$D_3 = \{\}$$

$$D_4 = \{\}$$

$$D_5 = \{\}$$

$$D_6 = \{\text{u, ústrojí, vlastnost}\}$$

$$D_7 = \{\}$$

```
{ s m i n h y m k o s , c o t , D i - m a t i k a , t i l d i n j , s b a z i , i j e t e s ,  
k a m t e s y m e t , m a t e j o j o p o s o v i , s e s , a i n y j , d e s t e n t ,  
e d d e t }
```

```
D1 = { s e s , e d i t e t }  
D2 = {}  
D3 = {}  
D4 = {}  
D5 = {}  
D6 = { s i s t o j , d e s t e n t }  
D7 = {}
```

Naivní L. algoritmus určil, že význam slova kočka v uvedené větě je 6. Je to spíš náhoda podpořená tím, že u významů 1 a 6 v SSJČ také nejvíc textu.

Jakou roli hrají v určení významu slova spojky a předložky? Jak odfiltrovat z definic a příkladů užití slova, která v určení významu slova nehrají roli? viz dál

## Inverzní četnost v dokumentu [Manning et al., 2008]

Term frequency  $tf$  – četnost znaku  $t$  v určitém dokumentu

Počet dokumentů  $N$

Document frequency  $df_t$  – počet dokumentů, ve kterých se vyskytuje  $t$

Inverse document frequency  $idf_t = \log \frac{N}{df_t}$

Příklad: mějme dokumenty: {Máma mele maso}, {Ema maso solí, z masa bude oběd}, {Máma má Emu}, {Ema má mámu i oběd}

$$N = 4$$

$$df_t(\text{maso}) = 2$$

$$idf_t(\text{maso}) = \log \frac{4}{2}$$

$$N = 4$$

$$df_t(\text{Ema}) = 3$$

$$idf_t(\text{Ema}) = \log \frac{4}{3}$$



## └ Algoritmy WSD

## └ Inverzní četnost v dokumentu [Manning et al., 2008]

Term frequency  $tf$  = počet znaků  $t$  v určitém dokumentuPočet dokumentů  $N$ Documnt frequency  $df_t$  = počet dokumentů, ve kterých se vyskytl znak  $t$ Inverzní četnost term frequency  $idf_t = \log \frac{N}{df_t}$ 

Příklad: máme dokumenty: { Máma mála maau }, { Ema maau sola, z maau laia obbí }, { Máma má Ema }, { Ema má máma í ubbí }

 $N = 4$  $df_t(\text{maau}) = 2$  $idf_t(\text{maau}) = \log \frac{4}{2}$  $N = 4$  $df_t(\text{Ema}) = 1$  $idf_t(\text{Ema}) = \log \frac{4}{1}$ 

Proč nepočítáme s frekvencí znaku  $tf$ ? V krátkých dokumentech je  $tf$  nižší než v dlouhých, neznámá to ale, že je tam  $t$  méně důležité.

Co vyjadřuje  $idf_t$ ? Čím běžněji se slovo vyskytuje, tím je  $idf_t$  nižší.

Co když je  $df_t = 0$ ? Používá se zvýšení jmenovatele (přičte se 1, aby ve jmenovateli nebyla 0).

K čemu se používá  $tf$ ? Dohromady s  $idf_t$  se používá při výpočtu míry (váhy) TF-IDF (počítá se jako násobek  $tf \times idf_t$ . Čím vyšší TF-IDF nějaké slovo v souboru dokumentů má, tím víc se předpokládá, že má vyšší relevanci.

## Leskův algoritmus: jednoduchý

pro každý význam  $s_i$  slova  $w$ :

nastav váhu  $v$  na 0:  $v(s_i) := 0$

najdi množinu slov  $O$  v okolí slova  $w$

pro každé slovo  $o_j$  z okolí  $O$

pro každý význam  $s_i$

pokud se  $o_j$  nachází v definici n. př. užití  $D_i$

přičti  $v(o)$  k  $v(s_i)$ :  $v(s_i) := v(s_i) + v(o)$

vyber  $s_i$  s nejvyšším  $v(s_i)$ : *return*  $\max(v(s_i))$

váha slova  $v(o) = idf_o$




## Leskův algoritmus: jednoduchý

1. malá kočkovitá šelma, chovaná v domácnostech, na venkově zvl. pro hubení myší; kočka domácí (zool.); šedivá, černá, třibarevná k.; hladká srst kočky; k. mňouká, přede; k. se plíží, číhá na myš; k. chytá ptáky; angorská k.; být falešný, úlisný jako k.; přen. expr. je to k. falešník; to děvče je k. lichotné, úlisné; [x] jsou na sebe jako pes a k. nenávidí se. . .
2. malá n. středně velká šelma s hustým kožichem; zool. rod Felis: k. plavá; k. divoká; k. domácí
3. samice kočkovité šelmy vůbec; rysí k.; lví k.; expr. každá kočkovitá šelma vůbec (tygr, levhart aj.)
4. ob. kožíšina na límci, kolem krku n. ramen
5. kocovina (Haš.)
6. věc připomínající někt. vlastnost kočky: bot. velký trs ostřic vystupující z rašeliníště (na blatech); tech. pojízdný vozík jeřábu se zdvihacím ústrojím
7. druh důtek; devítiocasá k.

## Leskův algoritmus: jednoduchý

Aminokyselina DL-methionin okyseluje moč, čímž chrání močové ústrojí psů i *koček* (důležitá vlastnost zvláště u kastrováných jedinců).

$i$	$D_i$	$v(s_i)$
1	$D_1 = \{\text{pes}(1, 525), \text{zvláště}(1, 83)\}$	3,355
2	$D_2 = \{\}$	0
3	$D_3 = \{\}$	0
4	$D_4 = \{\}$	0
5	$D_5 = \{\}$	0
6	$D_6 = \{\text{u}(0, 363), \text{vlastnost}(1, 79), \text{ústrojí}(2, 32)\}$	3,173
7	$D_7 = \{\}$	0

-  Agirre, E. and Edmonds, P. (2006).  
*Word sense disambiguation: algorithms and applications.*  
Text, speech, and language technology. Springer.
-  Kilgarriff, A. and Rosenzweig, J. (2000).  
English senseval: Report and results.  
In *Proceedings of the 2nd International Conference on Language Resources and Evaluation*, pages 1239–1244.
-  Manning, C. D., Raghavan, P., and Schtze, H. (2008).  
*Introduction to Information Retrieval.*  
Cambridge University Press, New York, NY, USA.