

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

29. října 2012

Algoritmy strojového učení

- „pravidlové“
 - rozhodovací seznamy
 - rozhodovací stromy
- „matematické“
 - pravděpodobnostní: naivní Bayesovský (Duda et Hart, 1973)
 - maximální entropie: Berger 1996
 - podobnostní: k-NN ve vektorovém prostoru (Ng et Lee, 1996)
- „promluvové“
 - one sense per discourse (Gale 1992)
 - one sense per collocation (Yarowsky, 1995)
 - redundance atributů

Rozhodovací seznam

```
if (zvíře má chobot) then output(slon)
if (zvíře má pruhy) then output(zebra)
if (zvíře má ploutve & zvíře není ryba) then
output(žralok)
```

Rozhodovací strom

savec?

žije ve vodě?

žije na souši?

žije v moři?

žije v řece?

býložravec?

masožravec?



„Matematické“ algoritmy zde uvedené jsou každý úplně jiný, spíš jde o reprezentanty různých skupin algoritmů.

Algoritmy strojového učení

- „pravidlové“
 - rozhodovací seznamy
 - rozhodovací stromy
- „matematické“
 - pravděpodobnostní: naivní Bayesovský (Duda et Hart, 1973)
 - maximální entropie: Berger 1996
 - podobnostní: k-NN ve vektorovém prostoru (Ng et Lee, 1996)
- „promluvové“
 - one sense per discourse (Gale 1992)
 - one sense per collocation (Yarowsky, 1995)
 - redundance atributů

Naivní Bayesovský klasifikátor

Naivní Bayesovský alg. předpokládá nezávislost znaků (což nemusí být správně), ale je rychlý.

$$P(C|F_1, \dots, F_n) = \frac{P(C) \cdot P(F_1, \dots, F_n|C)}{P(F_1, \dots, F_n)}$$

| zvíře | velikost | barva | potrava |
|-------|----------|----------|------------|
| slon | velký | šedý | býložravec |
| slon | střední | šedý | býložravec |
| kráva | velká | černá | býložravec |
| kráva | velká | strakatá | býložravec |
| kráva | malá | strakatá | býložravec |
| kráva | velká | bílá | býložravec |
| vlk | velký | černý | masožravec |
| vlk | malý | šedý | masožravec |

Naivní Bayes pro **velkého černého býložravce**

| zvíře | velikost | barva | potrava |
|-------|----------|----------|------------|
| slon | velký | šedý | býložravec |
| slon | střední | šedý | býložravec |
| kráva | velká | černá | býložravec |
| kráva | velká | strakatá | býložravec |
| kráva | malá | strakatá | býložravec |
| kráva | velká | bílá | býložravec |
| vlk | velký | černý | masožravec |
| vlk | malý | šedý | masožravec |

Na základě těchto dat můžeme vypočítat, že zvíře, které vidíme, bude: na 25 % slon, na 50 % kráva a na 25 % vlk, tj. $P(\text{slon}) = \frac{2}{8}$, $P(\text{kráva}) = \frac{4}{8}$ a $P(\text{vlk}) = \frac{2}{8}$.

Podmíněné pravděpodobnosti jsou $P(\text{černá barva}|\text{slon}) = 0$, $P(\text{černá barva}|\text{kráva}) = \frac{1}{4}$, $P(\text{černá barva}|\text{vlk}) = \frac{1}{2}$.

Naivní Bayes pro **velkého černého býložravce**

$$P(C|F_1, \dots, F_n) = \frac{P(C) \cdot P(F_1, \dots, F_n|C)}{P(F_1, \dots, F_n)}$$

$$P(\text{slon}) = \frac{2}{8}, P(\text{kráva}) = \frac{4}{8}, P(\text{vlk}) = \frac{2}{8}, P(\text{černý}|\text{slon}) = 0,$$

$$P(\text{černý}|\text{kráva}) = \frac{1}{4}, P(\text{černý}|\text{vlk}) = \frac{1}{2}, P(\text{velký}|\text{slon}) = \frac{1}{2},$$

$$P(\text{velký}|\text{kráva}) = \frac{3}{4}, P(\text{velký}|\text{vlk}) = \frac{1}{2}, P(\text{býložravec}|\text{slon}) = \frac{2}{2},$$

$$P(\text{býložravec}|\text{kráva}) = \frac{4}{4}, P(\text{býložravec}|\text{vlk}) = 0$$

$$P(\text{slon}|\text{velký, černý, býložravec}) = P(\text{slon})P(\text{velký}|\text{slon}) \cdot$$

$$P(\text{černý}|\text{slon}) \cdot P(\text{býložravý}|\text{slon}) = 0.25 \cdot 0.25 \cdot 0 \cdot 1 = 0$$

$$P(\text{kráva}|\text{velký, černý, býložravec}) = P(\text{kráva})P(\text{velký}|\text{kráva}) \cdot$$

$$P(\text{černý}|\text{kráva}) \cdot P(\text{býložravý}|\text{kráva}) = 0.5 \cdot 0.75 \cdot 0.25 \cdot 1 = \mathbf{0.09375}$$

$$P(\text{kráva}|\text{velký, černý, býložravec}) = P(\text{vlk})P(\text{velký}|\text{vlk}) \cdot P(\text{černý}|\text{vlk}) \cdot P(\text{býložravý}|\text{vlk}) =$$

$$0.25 \cdot 0.5 \cdot 0.5 \cdot 0 = 0$$

Shrnutí: algoritmy strojového učení

- „pravidlové“
 - rozhodovací seznamy
 - rozhodovací stromy
- „matematické“
 - pravděpodobnostní: naivní Bayesovský (Duda et Hart, 1973)
 - maximální entropie: Berger 1996
 - podobnostní: k-NN ve vektorovém prostoru (Ng et Lee, 1996)
- „promluvové“
 - one sense per discourse (Gale 1992)
 - one sense per collocation (Yarowsky, 1995)
 - redundance atributů

Word Sense Disambiguation

úkolem WSD je zjistit, jaký význam (z inventáře významů) má slovo ve vstupním textu

ukázali jsme si dva reprezentanty metod pro WSD: Leskův algoritmus pracující se slovníkovými definicemi a příklady užití a Yarowského algoritmus strojového učení

Word Sense Disambiguation: slabiny

největší slabinou je inventář významů

proto existují jednak snahy vytvořit dobré inventáře, jednak snahy úplně se inventářím vyhnout (HyperLex, [Véronis, 2004])

Word Sense Disambiguation: shrnutí

- všechny algoritmy pro WSD pracují s kolokacemi
- všechny pracují s určitým oknem, ve kterém kolokace sledují

PLIN021 Sémantická analýza v praxi

└ Slabiny WSD

└ Word Sense Disambiguation: shrnutí

- všechny algoritmy pro WSD pracují s lokalizací
- všechny pracují s určitým oknem, ve kterém lokalizace hledají

Ono okno může zásadně ovlivňovat průběhy algoritmů. Není žádná „doporučená velikost“ okna. Hlavním důvodem je to, co možná tušíme: různá slova mají různý dopad na význam promluvy. Sledováním velikosti a kvality tohoto okna (tj. kontextu) se budeme zabývat o něco později, až budeme znát také přístupy z úplně opačného konce.

Word Sense Disambiguation: měření kvality

soutěž SENSEVAL (www.senseval.org)

- vyhodnocení systémů pro WSD
- od roku 1998 (Senseval-1, -2, -3, Semeval-2007, -2010)
- od Semeval-1 jsou úkoly různé (např. přiřazení emoce ke krátkému textu, detekce metonymie ...)
- čeština (zatím) chybí
- data z proběhlých kol jsou k dispozici

soutěž SENSEVAL (www.senseval.org)

- vyhodnocovací systém pro WSD
- od roku 2003 (Senseval-2, -3, Semeval2007, -2010)
- od Semeval-3 jsou k dispozici i tzv. příkazní soubory ke každému testu, takže můžete soutěžit...
- celá ta stránka i tady
- data z prohlížečů i od jiných zdrojů

Ne, že by bylo třeba se Senseval/Semeval účastnit. Je dobré podívat se na ručně anotovaná data (málokdy je máme). Mnoho prací se také na soutěže odvolává.

soutěž SENSEVAL (www.senseval.org)

- vyhodnocovací systém pro WSD
- od roku 2003 (Senseval-2, -3, Semeval2007, -2010)
- od Semeval-2 jsou k dispozici i tzv. přifazení smyslu ke kontextu (sémantická, lexikální informace ...)
- celá řada jazyků (kyj)
- data z prototypů ke každému jazyku

Cokoli ze Senseval/Semeval je inspirací pro BP nebo referát.

Formalismy pro reprezentaci znalostí

WSD je zajímavá a „klasická“ disciplína, řada vědců ale WSD odmítá kvůli uvedeným slabinám.

Navíc, WSD je kvantitativní analýza, neříká nic o významu.

Při studiu významu se chceme posunout dál, skutečně k jádru věci. Chceme pracovat nejen se slovy, ale i se znalostmi: jazykovými i obecnými. . .

└ Formalismy pro reprezentaci znalostí

└ Formalismy pro reprezentaci znalostí

WSD je zejména „deskriptivní“ disciplína, která věnuje sk WSD
odmítavě i velmi složitým otázkám.

Navíc, WSD je kvantitativní analýza, měří se u výstupu.

Přítáží výstupu se chceme promatřit, stát se k jisté věci.
Chceme pracovat s jejím obsahem, ale i se znalostí jazykových
oblastí...

Může to vypadat, že dost bylo matematiky. Na chvíli si od ní odpočineme a osvěžíme termíny z lingvistiky.

Časem se k matematice zase vrátíme...

Sémantické rysy (semantic features)

matka = FEMALE + ADULT + HAS CHILD

batole = HUMAN – ADULT

- sémantické rysy jsou „atomy“ významu
- sémantické rysy jsou distinktivní rysy
- význam je definován pomocí s. rysů a pravdivostních podmínek

[Croft and Cruse, 2004]

Sémantické rysy: synonymie

| výraz | <i>jít</i> | <i>procházet</i> |
|-------|--|--|
| rysy | MOTION ON FOOT SELF-PROPELLED MEDIUM VELOCITY | MOTION ON FOOT SELF-PROPELLED MEDIUM VELOCITY |

Sémantické rysy: antonymie

| výraz | <i>jít</i> | <i>běžet</i> |
|-------|--|--|
| rysy | MOTION ON FOOT SELF-PROPELLED MEDIUM VELOCITY | MOTION ON FOOT SELF-PROPELLED HIGH VELOCITY |

Sémantické rysy: problém s antonymií

| výraz | <i>jít</i> | <i>letět střemhlav</i> |
|-------|--|--|
| rysy | MOTION ON FOOT SELF-PROPELLED MEDIUM VELOCITY | MOTION ON WINGS GRAVITY-PROPELLED HIGH VELOCITY |

└ Sémantické rysy

└ Sémantické rysy: problém s antonymií

| výraz | ry | ry |
|-------|-----------------|-------------------|
| | MOTION | MOTION |
| | ON FOOT | ON WINGS |
| | SELF-PROPELLED | GRAVITY-PROPELLED |
| | MEDIUM VELOCITY | HIGH VELOCITY |

antonymie nebo raději obecně opozitnost?

Sémantické rysy: problém s antonymií

| výraz | <i>jít</i> | <i>loudat se</i> |
|-------|--|---|
| rysy | MOTION ON FOOT SELF-PROPELLED MEDIUM VELOCITY | MOTION ON FOOT SELF-PROPELLED LOW VELOCITY |

2012-10-29

PLIN021 Sémantická analýza v praxi

└ Sémantické rysy

└ Sémantické rysy: problém s antonymi

| výraz | ry | ry |
|-------|-----------------|----------------|
| | MOTION | MOTION |
| | ON FOOT | ON FOOT |
| | SELF-PROPELLED | SELF-PROPELLED |
| | MEDIUM VELOCITY | LOW VELOCITY |

Sémantické rysy (jako všechny ostatní teorie) mohou být užitečné. Při jejich formálním uchopení narazíme na problémy: kolik rysů má každý výraz? Jak poznáme, které jsou podstatné a které ne? Trošku to připomíná problém, jak najít vhodnou otázku ve hře Myslím si zvíře. . .

Výběrová omezení (selectional restrictions)

slouží pro desambiguaci závislosti větných členů [Allen, 1995, kap. 10.1]

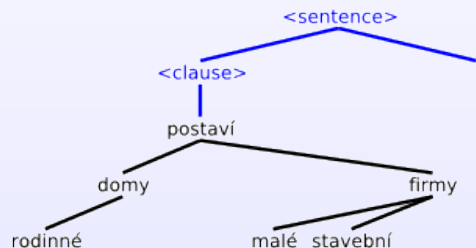
- Koupila jsem si pletenou čepici a šálu.
- Koupila jsem si nealkoholické pivo a křupky.

pletená šála – OK, nealkoholické křupky – NOT OK

Výběrová omezení (selectional restrictions)

slouží pro desambiguaci závislosti větných členů [Allen, 1995, kap. 10.1]

Rodinné domy postaví malé stavební firmy.



AGENT = rodinné
domy
THEME = malé
stavební firmy

AGENT = malé
stavební firmy
THEME = rodinné
domy

(AGENT postavit PERSON | INSTITUTION)
(THEME postavit BUILDING)



Allen, J. (1995).

Natural Language Understanding (2nd ed.).

Benjamin-Cummings Publishing Co., Inc., Redwood City, CA, USA.



Croft, W. and Cruse, D. (2004).

Cognitive linguistics.

Cambridge textbooks in linguistics. Cambridge University Press.



Véronis, J. (2004).

Hyperlex: Lexical cartography for information retrieval.

In *Computer Speech and Language: Special Issue on Word Sense Disambiguation*, page 23.